ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ

О. В. Редич

ІНФОРМАЦІЙНО-АНАЛІТИЧНІ ТЕХНОЛОГІЇ В ОПОДАТКУВАННІ





Редич Олександр Володимирович

Кандидат економічних наук, доцент, доцент кафедри інтелектуальних управляючих та обчислювальних систем УДФСУ. Керівник курсів за міжнародними освітніми програмами Cisco Network Academy, Oracle Academy. Автор понад 70-ти наукових та науково-методичних робіт.

Сфера наукових інтересів: методи і засоби моделювання економічних систем, ризик-менеджмент, системи підтримки прийняття рішень, технології інтелектуального аналізу даних (Data Mining), штучний інтелект (AI), машинне навчання (ML), обробка великих даних (BigData), програмування інтернет-речей (IoT).

У навчальному посібнику значну увагу приділено СКБД Oracle та тенденціям розвитку цієї платформи, на базі якої функціонує інформаційна система податкового відомства.

9 789663 375205

ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ

ІНФОРМАЦІЙНО-АНАЛІТИЧНІ ТЕХНОЛОГІЇ В ОПОДАТКУВАННІ

Серію «Митна справа в Україні» засновано в 2001 р.
Її продовженням є серія
«Податкова та митна справа в Україні», яку започатковано в 2015 р.

У СЕРІЇ «ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ»
ВИЙШЛИ ДРУКОМ:

2018

«Податкова діяльність держави в умовах становлення інформаційної постіндустріальної економіки»
«Зовнішня торгівля України. Митна статистика (1991–2016 роки)»
«Митна логістика» «Інноватика на фінансових ринках»
«Особливості документування махінацій у сфері державних фінансів, при вчиненні яких використовуються зареєстровані на територіях з пільговим оподаткуванням компанії»
«Міжнародне співробітництво держав у сфері захисту національних меншин»
«Адміністрування податків і платежів»
«Регулювання фінансових ринків в умовах глобалізації»
«Управління митними ризиками: теорія та практика»
«Митна вартість транспортних засобів: теоретико-методологічні засади контролю в Україні та міжнародна практика»
«Фінансовий ринок»
«Підприємництво в умовах формування нової економічної і технологічної нормальності суспільства»
«Актуальні проблеми удосконалення системи протидії митним правопорушенням»
«Оподаткування зовнішньоекономічної діяльності» «Фіскальна політика: теоретичні та практичні аспекти юридичної практики»

2019

«Електронний аудит – сучасна форма електронного контролю»
«Підприємницька діяльність»
«Митна наука. Анотований покажчик результатів наукових досліджень (2009–2018 рр.)»

СЕРІЯ «ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ»

Серію «Митна справа в Україні» засновано в 2001 р.
Її продовженням є серія «Податкова та митна справа в Україні»,
яку започатковано в 2015 р.

Редакційна колегія:
Пашко П. В., д.е.н. (голова)
Бережнюк І. Г., д.е.н.
Деркач Л. В., к.ю.н.
Крисоватий А. І., д.е.н.
Мазаракі А. А., д.е.н.
Ченцов В. В., д.н.держ.упр., д.іст.н.

У СЕРІЇ «МИТНА СПРАВА В УКРАЇНІ» ВИЙШЛИ ДРУКОМ:

2002

«Основи митної справи»

«Митний контроль та митне оформлення»

2003

«Таможенное оформление морских грузов»

«Основы таможенного дела в Украине»

«Митний контроль на автомобільному транспорті»

2004

«Митний контроль на залізничному транспорті»

«Митне оформлення автотранспортних засобів»

«Основи митної справи в Україні»

«Митний кодекс України та нормативно-правові акти, що регулюють його застосування»

«Митний контроль на повітряному транспорті»

«Митне регулювання зовнішньоекономічної діяльності в Україні»

«Коментар до Митного кодексу України»

2005

«Порядок заповнення вантажної митної декларації»

2006

«Основи мистецтвознавчої експертизи та вартісної оцінки культурних цінностей»

«Історія митної справи в Україні»

«Митне регулювання зовнішньоекономічної діяльності в Україні»

2008

«Основи митної справи в Україні»

2009

«Таможенный контроль: на пути к международным стандартам»

«Таможенный контроль в морских пунктах пропуска»

«Митне оформлення енергоносіїв»

«Митна безпека»

«Міжнародні перевезення товарів автомобільним транспортом: теорія і практика»

«Підготовка та підвищення кваліфікації митних брокерів і декларантів: збірник нормативних документів і відомчих роз'яснень»

2010

«Святий Матфей. Митар. Апостол. Євангеліст»

«Історія митної діяльності: Україна в європейському контексті»

2011

«Митні інформаційні технології»

«Особливості митної класифікації товарів»

«Інфраструктура відкритих ключів: технології, архітектура, побудова та впровадження»

«Розподілені бази даних»

2012

«Інформаційно-аналітичне забезпечення діяльності митних органів»

«Митний кодекс України» / Коментар до ст. 6»

«Митна політика та митна безпека: концептуальне визначення та шляхи забезпечення»

«Архітектура комп'ютера»

«Комп'ютерна схемотехніка»

2013

«Актуальні питання теорії та практики митної справи»

«Історія митної діяльності»

«Механізми державного управління митною справою»

«Формування системи митного аудиту»

2014

«Митна енциклопедія, т. 1»

«Митна енциклопедія, т. 2»

«Митна енциклопедія, т. 3»

У СЕРІЇ «ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ» ВИЙШЛИ ДРУКОМ:

2015

«Митне регулювання України: національні та міжнародні аспекти»
«Українська митниця: вчора, сьогодні, завтра»
«Митна політика держави та її реалізація митними органами»
«Механізми державного управління митною справою»
«Історія митної діяльності»
«Митне право України»
«Реформування податкової системи України: сучасні виклики та орієнтири»
«Митна справа»
«Міжнародні економічні відносини: у 2-х частинах»
«Адміністративно-правове сприяння соціально-економічному розвитку в Україні: теоретико-методологічні засади»
«Бухгалтерський облік в умовах застосування інформаційних технологій: історія та практика»
«Розслідування злочинів, пов'язаних із неналежним виконанням професійних обов'язків медичними працівниками: сучасний стан, проблеми та напрями удосконалення»
«Формування управлінської компетентності майбутніх економістів у процесі професійної підготовки: теорія і практика»
«Let Develor English skills»
«Кримінальний процес України»
«Європейська інтеграція»
«Протидія розслідуванню кримінальних правопорушень у сфері оподаткування: теорія та практика»
«Інформатика та обчислювальна техніка: практикум»

2016

«Регулятивний потенціал фінансового ринку в умовах глобальних викликів»
«Правовий статус фіскальних органів України та держав-членів Європейського Союзу»
«Психологічні основи у професійній підготовці фахівців ПМ»
«Математика для економістів. Вища математика. Частина І. Лінійна алгебра. Аналітична геометрія»
«Сучасні інформаційні системи і технології: управління знаннями»
«Економічна кібернетика: введення в спеціальність»
«Інформаційні системи та технології у фінансових установах»
«Фінансове забезпечення інноваційного розвитку»
«Практикум з інформаційних систем в економіці»
«Практикум з WEB-програмування»
«Малий та середній бізнес у взаємовідносинах із податковими органами: соціологічна інтерпретація»
«Моделювання систем захисту інформації»
«Формування бюджетної стратегії України: теорія, методологія, практика»
«Актуальні питання розвитку банківської системи України»
«Система державного фінансового контролю: формування та розвиток в Україні»
«Соціалізація студентської молоді у процесі професійної підготовки»
«Розвиток бухгалтерського обліку в умовах інтеграційних процесів»
«Правове забезпечення фінансової безпеки України»
«Grammar for Students of Economics (для економістів)»
«English Grammar for law Students (для юристів)»
«Алгоритмізація та програмування»
«Основи психології та конфліктології для фахової підготовки економістів»
«Проблеми соціологічного, психологічного та педагогічного забезпечення податкових відносин в Україні»
«Інформаційні системи і технології»
«Інформаційні системи і технології на підприємстві»
«Теорія ймовірностей та математична статистика: практикум»
«Економічні перспективи підприємництва України»
«Актуальні проблеми теорії держави і права»
«Математика для економістів. Функція багатьох змінних. Ряди»

2017

«Теорія держави і права»

«Реформування податкової та митної систем України»

«Історія митної діяльності»

«Механізм державного управління митною справою»

«Митна політика держави та її реалізація»

«Митне право України»

«Стан та перспективи розвитку
бухгалтерського обліку та оподаткування в Україні»

«Державні цільові фонди»

«Аналітично-контрольні аспекти регулювання
фінансово-господарської діяльності суб'єктів економіки»

«Міжнародне фінансове право»

«Концептуалізація інституту уповноважених економічних операторів:
проблеми забезпечення безпеки та спрощення міжнародної торгівлі»

«Тактика розслідування фінзлочинів, вчинених із використанням
зареєстрованих у зоні пільгового оподаткування компаній»

«Інститут митної вартості імпортних товарів в Україні:
актуальні питання та вектори розвитку»

«Державна фіскальна служба України: управління кадровим потенціалом
в умовах інституційних змін»

«Фінансовий ринок України: розбудова інфраструктурної компоненти»

«Податкова політика України: теоретико-правовий аспект»

«Митна статистика України»

«Запобігання митним правопорушенням: охорона і захист прав інтелектуальної власності
на митному кордоні України»

«Інформаційні системи і технології в податковій справі»

«Митна справа України

«Україна – територія можливостей для розвитку підприємництва»

«Актуальні проблеми приватного права»

«Антикризове управління фінансовими корпораціями»

«Оподаткування електронної комерції в Україні»

«Протидія агресивному податковому плануванню:
світовий досвід та виклики для України»

2018

«Податкова діяльність держави в умовах становлення інформаційної
постіндустріальної економіки»

«Зовнішня торгівля України. Митна статистика (1991–2016 роки)»

«Митна логістика»

«Інноватика на фінансових ринках»

«Особливості документування махінацій у сфері державних фінансів,
при вчиненні яких використовуються зареєстровані на територіях
з пільговим оподаткуванням компанії»

«Міжнародне співробітництво держав у сфері захисту національних меншин»

«Адміністрування податків і платежів»

«Регулювання фінансових ринків в умовах глобалізації»

«Управління митними ризиками: теорія та практика»

«Митна вартість транспортних засобів: теоретико-методологічні засади
контролю в Україні та міжнародна практика»

«Фінансовий ринок»

«Підприємництво в умовах формування нової економічної і технологічної
нормальності суспільства»

«Актуальні проблеми удосконалення системи протидії митним правопорушенням»

«Оподаткування зовнішньоекономічної діяльності»

«Фіскальна політика: теоретичні та практичні аспекти юридичної практики»

2019

«Електронний аудит – сучасна форма електронного контролю»

«Підприємницька діяльність»

«Митна наука. Анотований покажчик результатів наукових досліджень (2009–2018 рр.)»

УНІВЕРСИТЕТ ДЕРЖАВНОЇ ФІСКАЛЬНОЇ СЛУЖБИ УКРАЇНИ

СЕРІЯ «ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ»
Заснована 2015 року і є продовженням
започаткованої у 2001 році серії «Митна справа в Україні»

Редич О. В.

ІНФОРМАЦІЙНО-АНАЛІТИЧНІ ТЕХНОЛОГІЇ В ОПОДАТКУВАННІ

Навчальний посібник

Ірпінь
2019

УДК 519.81:336.221(075.8)
ББК 65.261.41(я73)
Р 33

Рекомендовано до друку рішенням Вченої ради
Університету державної фіскальної служби України
(протокол №13 від 20 грудня 2018 р.)

Рецензенти:

Лондар С.Л., д.е.н., професор, директор ДНУ «Інститут освітньої аналітики»;

Pінпа С.П., д.е.н., професор кафедри економічної кібернетики Університету державної фіскальної служби України.

Редич О. В.

Р 33 Інформаційно-аналітичні технології в оподаткуванні: навчальний посібник / О.В. Редич. – Ірпінь : Університет ДФС України, 2018. – 438 с. (Серія «Податкова та митна справа в Україні, т. 127).

ISBN 978-966-337-520-5

У навчальному посібнику розглянуто основні аспекти інформаційного та технічного забезпеченння процесів адміністрування податків, інформаційно-облікові та сервісні системи податкового відомства держави, сучасні технології аналітики, що вже використовуються, і ті, що можуть бути впроваджені в операційній діяльності аналітичних підрозділів.

Значну увагу приділено СКБД Oracle та тенденціям розвитку цієї платформи, на базі якої функціонує інформаційна система податкового відомства. Розглянуто нову опцію Oracle Advanced Analytics, що інтегрує в собі широкі можливості з накопичення, збереження та захисту даних, зокрема Великих Даних (Big Data) та інтелектуального аналізу (Data Mining). Серед методів інтелектуального аналізу проаналізовано нейронні мережі, асоціативні правила, методолоігію дерев рішень, нечіткі множини (Fuzzy) та ін. Кожен розділ завершується лабораторним практикумом.

Може бути корисний студентам, викладачам, науковцям, що вивчають, викладають, досліджують інформаційні технології та їх прикладний аспект в економічній і податковій сферах.

УДК 519.81:336.221(075.8)
ББК 65.261.41(я73)

© Редич О. В., 2019
© Університет державної фіскальної служби України, 2019

ISBN 978-966-337-520-5

ЗМІСТ

ВСТУП 9

РОЗДІЛ 1. ІНФОРМАЦІЙНА ІНФРАСТРУКТУРА ПОДАТКОВОГО ВІДОМСТВА УКРАЇНИ 12

1.1. Інформаційне та технічне забезпеченння процесів адміністрування податків 12
1.1.1. Погляд на податкову систему з точки зору кібернетики 12
1.1.2. Інформаційно-аналітичне забезпечення процесів адміністрування податків 20
1.1.3. Інтегрована інформаційно-телекомунікаційна система податкового відомства України 25
1.1.4. Інформаційно-облікові та сервісні системи податкового відомства України 29
Питання для самоконтролю до теми 1.1 36

1.2. Операційні процеси та інформаційні підсистеми податкового відомства 37
1.2.1. Операційний процес «Облік платників податків» 37
1.2.2. Операційний процес «Обробка податкової звітності та платежів» 39
1.2.3. Операційний процес «Облік платежів» 40
1.2.4. Операційний процес «Податковий аудит» 43
1.2.5. Аналітична система та система адміністрування доступу до даних 45
Питання для самоконтролю до теми 1.2 46

1.3. Методи та засоби моделювання і аналізу податкової інформації 47
1.3.1. Класифікація методів підтримки прийняття рішень 47
1.3.2. Оцінювання результативності податкового відомства методом інтегральної згортки показників 50
1.3.3. Оцінювання податкової політики на основі методології Лаффера 66

1.3.4. Аналітичні показники для оцінювання податкового зобов'язання регіону 74
1.3.5. Моделі цільових функцій регіональної податкової політики 78
1.3.6. Методологічні основи управління ризиками в адмініструванні податків 84
1.3.7. Імовірнісне та кількісне оцінювання податкових ризиків 97
Питання для самоконтролю до теми 1.3 105

1.4. СКБД Oracle – базова платформа інформаційно-аналітичної системи податкового відомства України 107
1.4.1. Клієнт-серверна інфраструктура інформаційної системи податкового відомства 107
1.4.2. Логічні та фізичні об'єкти сервера бази даних Oracle 111
1.4.3. Логічні структури екземпляра бази даних Oracle. Табличні простори і схеми. 114
1.4.4. Об'єкти та логічні структури зберігання і доступу до даних 116
1.4.5. Структури оперативної пам'яті екземпляра БД Oracle 118
1.4.6. Фонові процеси екземпляра сервера Oracle 120
1.4.7. Утиліти для написання запитів до БД Oracle 121
Питання для самоконтролю до теми 1.4 122

Лабораторний практикум 1 124
Лабораторна робота 1.1
Оцінювання результативності підрозділів податкового відомства на основі інтегральної згортки показників 124
Лабораторна робота 1.2
Оцінка ймовірнісних показників податкових ризиків 127
Лабораторна робота 1.3
Організація робочого простору на хмарному ресурсі Oracle Application Express (APEX) 130

РОЗДІЛ 2. ТЕХНОЛОГІЇ АНАЛІЗУ ПОДАТКОВОЇ ІНФОРМАЦІЇ. МОВА SQL ТА СКБД ORACLE 136

2.1. Пакети прикладних програм для аналізу інформації та технології доступу до баз даних 136
2.1.1. Огляд прикладних пакетів статистичного та інтелектуального аналізу 136
2.1.2. Технології доступу до баз даних ODBC, DAO, ADO, OLE DB, RDO 139
2.1.3. Особливості з'єднання з БД Oracle через ODBC 141
2.1.4. Об'єктно-орієнтована технологія доступу до даних OLE DB 143
2.1.5. Технологія ActiveX Data Object (ADO) 145
Питання для самоконтролю до теми 2.1 148

2.2. SQL – мова структурованих запитів до баз даних 149
2.2.1. Загальні відомості та категорії команд SQL 149
2.2.2. Основні типи даних 153
2.2.3. Створення та модифікація таблиць БД Oracle 155
2.2.4. Команда SELECT. Основи запитів на мові SQL 159
Питання для самоконтролю до теми 2.2 173

2.3. Особливості застосування аналітичних функцій Oracle 174
2.3.1. Огляд аналітичних функцій SQL Oracle 174
2.3.2. Формування запитів у конструкції «вікна» 193
2.3.3. Використання метасимволів для підготовки запитів Select 195
2.3.4. Використання статистичних аналітичних функцій 198
2.3.5. Використання функцій перетворення типів даних 204
Питання для самоконтролю до теми 2.3 208

2.4. АІС «Галузь» як джерело податкової інформації для аналізу і прогнозування 209
2.4.1. Основні відомості про АІС «Галузь» 209
2.4.2. Особливості роботи із довідниками схеми «Еталон» 215
2.4.3. Структура таблиць зведених показників АІС «Галузь» 218
2.4.4. Рекомендації щодо формування та оптимізації аналітичних запитів 220
Питання для самоконтролю до теми 2.4 222

Лабораторний практикум 2 223
Лабораторна робота 2.1
Доступ до баз даних Oracle на основі технології ODBC. Відбір даних та побудова зведених таблиць. 223
Лабораторна робота 2.2
Організація OLEDB доступу до БД Oracle із пакета Statistica та проведення дисперсійного аналізу даних 239
Лабораторна робота 2.3
SQL запити до БД АІС «Галузь» (Oracle SQL Developer) 248
Лабораторна робота 2.4
Аналітичні SQL запити до БД АС «Податкова звітність» (Oracle SQL Developer) 260

РОЗДІЛ 3. МЕТОДИ ТА ЗАСОБИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ПОДАТКОВОЇ ІНФОРМАЦІЇ 267

3.1. Огляд сучасних платформ і напрямів розвитку технологій інтелектуального аналізу 267
3.1.1. Загальні відомості про інтелектуальний аналіз (Data Mining) 267
3.1.2. Класифікація методів Data Mining 270
3.1.3. Технології Oracle Advanced Analytics 274
3.1.4. Опція поглибленої аналітики Oracle Data Mining 275
3.1.5. Опція Oracle R Enterprise 279
Питання для самоконтролю до теми 3.1 283

3.2. Регресійно-кореляційні моделі податкових процесів 284
3.2.1. Основні поняття регресійно-кереляційного аналізу 284
3.2.2. Оцінка значимості регресійної моделі 289
3.2.3. Регресійно-кореляційна модель оцінювання тіньової економіки і податкових втрат 293
3.2.4. Податкове навантаження як фактор ризику економічної безпеки держави 298
3.2.5. Прогнозування податкових надходжень на основі авторегресійних моделей динамічних рядів 305
3.2.6. Нелінійна оптимізація у визначенні точок Лаффера 310
Питання для самоконтролю до теми 3.2 313

3.3. Побудова класифікаційних моделей на базі технологій Data Miming 315
3.3.1. Нейронні мережі у задачах формування і виявлення образів суб'єктів оподаткування 315
3.3.2. Вибір структури мережі і методу навчання 319
3.3.3. Огляд структур та типів нейромереж 321
3.3.4. Кластерний аналіз і класифікація регіонів на основі параметрів податкового навантаження 328
3.3.5. Дискримінантний аналіз у прогнозуванні платоспроможності суб'єктів оподаткуввання 331
Питання для самоконтролю до теми 3.3 338

3.4. Застосування асоціативних правил в аналізі діяльності господарюючих об'єктів 339
3.4.1. Основні відомості про асоціативні правила 339
3.4.2. Основні характеристики асоціативних правил 342
3.4.3. Пошук асоціативних правил 346
3.4.4. Стратегії ведення пошуку правил 349
Питання для самоконтролю до теми 3.4 353

3.5. Застосування методології дерев рішень у класифікаційних завданнях 354
3.5.1. Основні поняття методології дерев рішень 354
3.5.2. Алгоритми дерев рішень ID3 і C4.5 356

3.5.3. Застосування класифікаційних методів
Oracle Data Mining 362
Питання для самоконтролю до теми 3.5 368

3.6. Застосування методології нечітких множин в аналізі ризиковості платників податків 369
3.6.1. Основні відомості про нечіткі множини 369
3.6.2. Система нечіткого логічного висновку 372
3.6.3. Використання методології нечітких множин в оцінюванні податкових ризиків 373
Питання для самоконтролю до теми 3.6 377

Лабораторний практикум 3 379
Лабораторна робота 3.1
Кластерний аналіз регіонів за показниками податкових надходжень (Пакет Statistica) 379
Лабораторна робота 3.2
Дискримінантний аналіз фінансових показників підприємств за допомогою пакета Statistica 385
Лабораторна робота 3.3
Оцінювання параметрів кривої Лаффера у середовищі Statistica 393
Лабораторна робота 3.4
Карти Кохонена (середвище Deduktor Studio) 403
Лабораторна робота 3.5
Дослідження методів класифікації в середовищі Oracle Data Miner 409
Лабораторна робота 3.6
Дослідження побудованих моделей Decision Trees і застосування їх для прогнозування поведінки платників податків 418

ПРЕДМЕТНИЙ ПОКАЖЧИК 425
ПОЗНАЧЕННЯ ТА СКОРОЧЕННЯ 429
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 434

ВСТУП

Від ефективності функціонування податкової системи залежить фінансова спроможність держави, ефективність державного регулювання економічними перетвореннями. Прийняття необґрунтованих рішень у цій галузі діяльності може призводити до надзвичайних економічних втрат, погіршення соціально-економічного стану суспільства. Це пояснюється тим, що рішення посадових осіб, пов'язаних з процесами оподаткування, впливають як на макроекономічні процеси у суспільстві, так і на фінансовий стан кожного громадянина.

Підвищення ефективності діяльності підрозділів податкового відомства, формування ефективних управлінських рішень та оцінки наслідків їх реалізації є комплексним завданням, складність вирішення якого пояснюється властивою цій структурі невизначеністю, конфліктною природою процесів досягнення цілей управління та умовами функціонування економічного середовища.

Незважаючи на те, що процеси прийняття рішень у державному управлінні, зокрема оподаткуванні, є більш систематизованими та регламентованим, ніж у інших сферах, відсутність впроваджених засобів аналітики, апробованих економіко-математичних методів податкових процесів породжує велику кількість неефективних процедур адміністрування податками, ускладнює взаємовідносини між платниками податків і податковими органами.

Оскільки відомі апробовані економіко-математичні моделі функціонування податкових систем в основному орієнтовані на стабільну ринкову або адміністративно-командну економіку і базуються на застосуванні різноманітних варіацій регресійного аналізу, необхідно розширити використання методів і засобів аналітики та моделювання податкових процесів у процесах прийняття рішень, адекватних сучасному перехідному стану економіки України.

Схвалена Кабінетом Міністрів 17 січня 2018 року Концепція розвитку цифрової економіки та суспільства України на 2018–2020 орієнтована на прискорення економічного зростання через

ефективність її державних установ, до яких належить і податкове відомство України. Концепція передбачає перехід від сировинного типу економіки, яка споживає природні ресурси, до високотехнологічних виробництв та ефективних процесів за допомогою IT-технологій та комунікацій. Уряд готовий спільно з експертним середовищем та бізнесом напрацьовувати та втілювати рішення з підтримки технологічного прориву, підтримки стартапів, осучаснення освіти державного управління та місцевого самоврядування.

Мета посібника: сформувати у студентів цілісне розуміння про інформаційну систему податкового відомства як сукупності організаційної, методологічної, інформаційної інфраструктур, висвітлити актуальні питання застосування як підприємницьким середовищем, так і державними контролюючими установами інформаційно-аналітичних технологій, засобів і методів штучного інтелекту, інтелектуального аналізу, направлених на підвищення ефективності прийняття рішень у сфері оподаткування.

Завдання посібника: висвітлити особливості інформаційно-аналітичної інфраструктури податкового відомства України (ДФС України 2014–2018 рр.), надання студентам системних знань і практичних навичок з використання засобів аналізу даних, особливості застосування методів та засобів статистичного аналізу інформації, інтелектуального аналізу даних, Data Mining, зокрема технологій на платформі Oracle.

У першому розділі посібника висвітлено питання інформаційно-аналітичного забезпечення, основні операційні процеси відомства, методи та засоби моделювання і аналізу податкової інформації, розглянуто платформу Oracle як основу інформаційно-аналітичної системи податкового відомства України.

У другому – розглянуто питання використання спеціалізованих статистичних пакетів, особливості технологій доступу до БД, аналітичні функції мови структурованих запитів до БД SQL, особливості аналізу даних АІС «Галузь».

У третьому – розкрито питання застосування методів та засобів регресійно-кореляційного аналізу, огляд технологій Data Mining, особливості застосування Oracle Advanced Analytics, Oracle Data

Mining, особливості побудови класифікаційних моделей, кластерного, дискримінантного аналізу, дерев рішень як компонентів Data Mining пакетів Statistica, Deductor.

Кожен розділ завершується лабораторним практикумом, теми – питаннями для самоконтролю. Посібник призначений для отримання теоретичних знань, використання під час підготовки і виконання самостійної роботи, лабораторних практикумів з предметів «Інтелектуальні системи підтримки прийняття рішень», «Інтелектуальний аналіз даних», «Інформаційно-аналітичні технології в податкових і митних органах», «Штучні нейронні мережі в комерції та бізнесі» освітньої програми спеціальності 122 «Комп'ютерні науки», спеціалізації «Інформаційні управляючі системи і технології в економіці».

РОЗДІЛ 1
ІНФОРМАЦІЙНА ІНФРАСТРУКТУРА ПОДАТКОВОГО ВІДОМСТВА УКРАЇНИ

У розділі висвітлюються особливості прийняття рішень в адмініструванні податків, компоненти інформаційної системи, основні операційні процеси, інформаційно-телекомунікаційна інфраструктура, основні платформи інформаційно-аналітичної системи податкового відомства України (ДФС України, станом на 01.12.2018): ІТС «Податковий блок» та її підсистеми. Для розуміння призначення інформаційно-аналітичної системи податкового відомства України у розділі подано аналіз податкової системи як складної кібернетичної системи, систематизовано показники ефективності її функціонування, показані цільові функції і критерії.

1.1. Інформаційне та технічне забезпечення процесами адміністрування податків

1.1.1. Погляд на податкову систему з точки зору кібернетики

У загальному значенні під *системою* будемо розуміти сукупність взаємопов'язаних елементів, які взаємодіють як ціле для досягнення загальної мети. Тому першим елементарним рівнем описування будь-якої системи повинно бути визначення множини елементів, що належать до системи.

Будь-яка реальна система вміщує організацію, проте не всі організації є системою. Організація або група організацій утворюють систему за наявності загальної цілі функціонування для усіх її елементів. Доповнення упорядкованої множини елементів системи сукупністю взаємозв'язків утворює організацію. Розгляд податкової системи відповідно до принципів системного аналізу

передбачає подання її як саморегулюючої економічної системи з наявністю процесів управління, накопичення та аналізу інформації. Необхідно зазначити, що в науковій економічній літературі *податкова система* розглядається, зазвичай, як *система податків* (прямих, непрямих, місцевих, загальнодержавних тощо). У цьому посібнику податкову систему розглядаємо як систему з точки зору кібернетики – науки про інформацію, зв'язки й управління.

Проблема регулювання податкової системи, що виникає у суспільстві, повинна розглядатися у контексті взаємодії рівнів державного управління. Прийняття рішень під час управління такою системою повинно забезпечуватися декомпозицією цілей взаємодіючих рівнів, визначенням множини альтернатив досягнення цілей, що зіставляються за визначеними критеріями ефективності і вибором прийнятних для конкретної ситуації способів управління.

Термін *«адміністрування податків»* у широкому розумінні означає процеси управління та регулювання відносинами у сфері оподаткування, що забезпечують здійснення податкового менеджменту завдяки процедурам контролю за справлянням (стягненням) податків, методологічного забезпечення та здійснення планування, обліку і контролю виконання планових показників податкових надходжень, дотримання суб'єктами податкових відносин податкового законодавства.

Відповідно до вищезазначеного можна виокремити такі стратифікаційні рівні державного управління, які забезпечують функціонування податкової системи: 1) концептуальний рівень: система принципів та соціально-правових основ взаємовідносин держави і платників податків; 2) регулюючий рівень: Президент, Верховна Рада, Кабінет Міністрів, що безпосередньо встановлюють та корегують систему податків і відносини у сфері оподаткування; 3) виконавчий рівень: Державна податкова служба та інші установи, що здійснюють фіскальні функції; 4) цільовий рівень –

Державний бюджет; 5) контролюючий рівень: Рахункова палата, Державне казначейство, Державна контрольно-ревізійна служба, що контролюють надходження та видатки бюджету. Особливе місце у податковій системі як складній соціально-економічній саморегулюючій системі займає *суб'єктно-об'єктний рівень оподаткування* – платники податків та об'єкти (база) оподаткування, що не належать до категорії державного управління і є об'єктами дії як виконавчого рівня, так інших рівнів. Стратифікаційна модель податкової системи України як комплекс взаємодіючих рівнів державного управління наведена на рис. 1.1.



Рис. 1.1. Стратифікаційна модель податкової системи України як комплекс взаємодіючих рівнів державного управління

На концептуальному рівні принципи функціонування податкової системи визначає Конституція України. Статтями Конституції 67, 74, 92, 116, 143 закріплюються права і обов'язки громадян, територіальних громад і господарюючих суб'єктів та держави щодо взаємовідносин у питаннях оподаткування.

Регулюючий (законодавчо-нормативний) рівень податкової системи забезпечують Верховна Рада України, Кабінет Міністрів, Президент України. Відповідно до Конституції України Державний бюджет України і бюджетна система України, система оподаткування, податки і збори встановлюються виключно законами України. Тобто на цьому рівні визначаються базові принципи формування податкових відносин між державою та суб'єктами оподаткування відповідно до цілей та завдань, визначених суспільством.

Виконавчий рівень утворює Державна податкова служба України та інші установи, що здійснюють фіскальні функції, відповідно до Податкового кодексу, внутрішніх нормативів та іншого відомчого законодавства, що регулюють процедури сплати податків і обов'язкових платежів.

Контролюючий рівень утворюють Рахункова палата Верховної Ради, Державна аудиторська служба та Державна казначейська служба України, які контролюють використання бюджетних коштів, координують видатки та акумулюють і перерозподіляють надходження до бюджету. Аналітична інформація контролюючих органів використовується на законодавчо-нормативному рівні для регулювання процесів оподаткування.

Податкова система як складна система є комплексом взаємодіючих стратифікаційних рівнів державного управління, що формують та здійснюють податкову політику, направлену на досягнення визначених цілей відповідно до законодавчо затвердженої системи податків, методології визначення суб'єктів і об'єктів оподаткування та принципів формування податкових відносин у суспільстві.

Модернізація інформаційно-аналітичної системи податкового відомства та її структурних підрозділів повинна враховувати наведену стратифікаційну, кібернетичну модель податкової системи України як комплексу взаємодіючих рівнів державного управління. Це важливо вже зараз, коли державними відомствами накопичені величезні обсяги (сотні терабайт) інформації, які потрібно уміти аналізувати, використовувати, оберігати від несанкціонованого доступу.

Міжнародний досвід показує, що ефективне управління діяльністю податкового відомства ґрунтується на системному підході, досягненнях теорії податкового, проектного менеджменту, бізнес-аналізі, теорії прийняття рішень, інтеграції в інформаційній системі інтелектуальних технологій аналізу даних.

Управління діяльністю органів податкового відомства України (адміністрування податків і зборів) повинно бути направлене на забезпечення виконання покладених на відомство функцій та завдань на основі ефективного використання інтелектуальних і матеріальних ресурсів, якісного розвитку процесів адміністрування податків і податкових відносин. У рамках такої моделі управління діяльністю органів новоствореної Державної податкової служби України можна планувати систему управління якістю (далі – СУЯ). Основним компонентом СУЯ може бути система підтримки прийняття рішень (далі – СППР), функціонування якої, у свою чергу, базується на системі управління ризиками (система ризик-менеджменту (далі – СРМ) та системі оцінювання результатів діяльності підрозділів та територіальних органів (далі – СОРД) (рис. 1.2).

Одним із основних компонентів *Комплексної системи управління діяльністю ДПС України* (далі – КСУ), як це показано на структурній схемі, є *Система оцінювання результатів діяльності органів ДПС України* (далі – СОРД). Завданням СОРД є формування та обчислення сталих у часі, незмінних за основним економічним змістом показників, що характеризують основні результати операційної діяльності органів ДПС України, дозволяють класифікувати їх за рівнем податкових ризиків та якості діяльності у динаміці.



Рис. 1.2. Узагальнена структура системи управління діяльністю органів податкового відомства України

Аналітичну підтримку процесів адміністрування податків можна подати класичною схемою підготовки і прийняття рішень, доповнивши її блоком оцінювання результатів і зворотним зв'язком корегування та уточнення моделі процесу оподаткування. Подана схема є структурною схемою процесу прийняття рішень, що забезпечується класичними процесами: цілепокладання, розробки і прийняття рішень, організації виконання і контролю (рис. 1.3).

На сьогодні термін «система підтримки прийняття рішень» (СППР) або англійською *Decision Support System (DSS)* визначає спеціалізовану інтегровану інформаційну технологію для забезпечення досягнення визначених цілей.

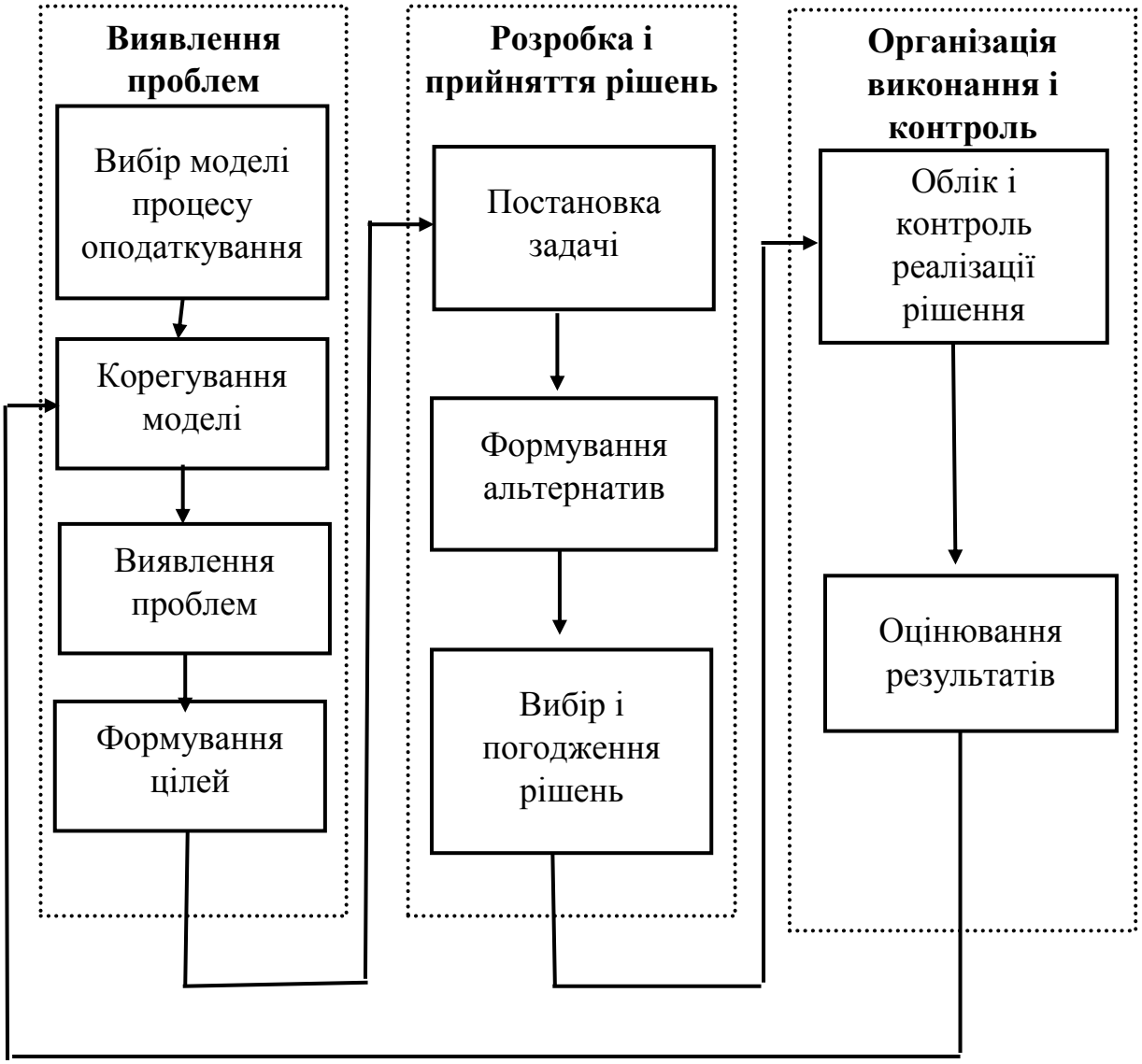


Рис. 1.3. Функціональна схема процесів підтримки прийняття рішень в адмініструванні податків

Система підтримки прийняття рішень – це інтерактивна прикладна інформаційна система, що забезпечує кінцевим користувачам, які приймають рішення, легкий і зручний доступ до даних і моделей з метою прийняття ефективних рішень у слабоструктурованих і неструктурованих ситуаціях у різних галузях людської діяльності.

Для досягнення суб'єктом управління поставлених цілей і завдань, яким у цьому разі виступає відповідний рівень управлін-

ня ДПС України, повинен мати у своєму розпорядженні систему реалізації цих цілей, тобто систему управління. Під час створення системи управління необхідно враховувати складність об'єктів управління, які за класифікацією поділяють на прості і складні. Так, податкова система загалом і органи ДПС зокрема належать до складних систем у зв'язку із наявністю множини взаємопов'язаних різнорідних частин, наявністю випадкових збурень як у самій системі, так і поза її межами (інфляція, зміни законодавства тощо). На рис. 1.4 подана схема управління органом ДПС України як складним об'єктом, що включає такі елементи: модель управління, формування цілі управління (цілепокладання), формування альтернатив рішень, управлінські дії, корегування моделі, базу оподаткування, звітність, корекція моделі.



Рис. 1.4. Схема управління органом податкового відомства

Власне СППР повинна забезпечувати вказані процеси на основі вибраної моделі управління та процесів оподаткування. Під час побудови моделей процесів оподаткування для застосування їх у прийнятті рішень необхідно враховувати *полісистемність* середовища, у якому функціонують суб'єктний і об'єктний рівні та багатокритеріальність цільових функцій взаємодіючих рівнів податкової системи. Наведені вище елементи інтегрованої інформаційної інфраструктури податкового відомства вказують на потребу державних закладів у спеціалізованих аналітичних технологіях та методологічному забезпеченні прийняття обґрунтованих

рішень у формуванні цілей і основних завдань, що повинні бути вирішені у процесах оподаткування. Актуальними завданнями є дослідження сучасних підходів до оцінювання ефективності податкової політики, систематизація і уточнення показників різних аспектів діяльності державних установ, що впливають на її формування і здійснення. Вирішення цих проблем дасть змогу зменшити ризики отримання негативних результатів під час прийняття рішень в оподаткуванні.

1.1.2. Інформаційно-аналітичне забезпечення процесів адміністрування податків

Метою інформаційного забезпечення підрозділів процесів адміністрування податків є отримання деталізованої та агрегованої інформації стосовно усіх аспектів адміністрування податків, яка має слугувати основою для прийняття управлінських рішень. Досягнення цієї мети передбачає вирішення ряду завдань, таких як збирання первинної інформації, її зберігання, перерозподіл між структурними підрозділами органів і їхніми працівниками, підготовка до переробки, власне обробка, надання суб'єкту управління в зручному поданні аналітичних довідок, забезпечення прямих і зворотних зв'язків її циркуляції тощо. При сучасних потоках інформації ефективне вирішення цих завдань неможливе без використання потужних обчислювальних ресурсів, методів та засобів формування ефективних запитів до баз даних.

Державна податкова служба України збирає, зберігає та поширює інформацію, здійснює різноманітні інформаційні процеси та використовує інформаційні ресурси. Безперервна циркуляція інформації між структурними підрозділами ДПС України, відомством та зовнішніми реципієнтами є неодмінними і необхідними атрибутами адміністрування податків. За допомогою інформаційної взаємодії із іншими державними відомствами ДПС України надає послуги підприємницькому середовищу, великому бізнесу, координує взаємовідносини суспільства і держави щодо реалізації податкової політики, спрямовує свій розвиток на забезпечення ефективності податкової системи як системи податків і як кібер-

нетичної системи управління, перетворення відомства на ефективну сервісну службу.

Важливим у цьому аспекті є *інформаційно-аналітичне забезпечення* органів податкового відомства.

Інформаційно-аналітичне забезпечення процесів адміністрування податків – це діяльність органів відомства в установленому законом порядку із застосуванням принципів, методів, способів, правил, схем та алгоритмів, за якими здійснюється пошук даних, їх збирання, обробка, накопичення та зберігання, що спрямована на надання підрозділам необхідних для вирішення ситуації даних в обсязі, достатньому для функціонування системи.

Для *інформаційно-аналітичного напряму* реалізації інформаційної функції податкового відомства України використовують інформацію:

1) що надійшла від платників податків та податкових агентів, зокрема інформацію, що міститься у податкових деклараціях, розрахунках, звітах про використання доходів (прибутків) неприбуткової організації, що зазначені у звітних документах;

2) що міститься у наданих великими платниками податків в електронній формі копіях документів з обліку доходів, витрат та інших показників, пов’язаних із визначенням об’єктів оподаткування (податкових зобов’язань), первинних документах, що ведуться в електронній формі, регістрах бухгалтерського обліку, фінансовій звітності, інших документах, пов’язаних з обчисленням та сплатою податків і зборів;

3) про фінансово-господарські операції платників податків;

4) про застосування реєстраторів розрахункових операцій;

5) від органів виконавчої влади, органів місцевого самоврядування та Національного банку України, зокрема інформацію про об’єкти оподаткування, що надаються та/або реєструються такими органами. Зазначена інформація повинна містити, зокрема, вид, характеристики, індивідуальні ознаки об’єкта оподаткування (у разі наявності), за якими його можна ідентифікувати;

6) про результати здійснення державного контролю за господарською діяльністю платника податків;

7) що міститься у звітних документах (крім персоніфікованої статистичної інформації), які подаються платником податків органам виконавчої влади та/або органам місцевого самоврядування;

8) про встановлені органами місцевого самоврядування ставки місцевих податків, зборів та надані такими органами податкові пільги;

9) про дозволи, ліцензії, патенти, свідоцтва на право провадження окремих видів діяльності, яка повинна містити, зокрема: найменування платника податків, якому видані такі дозволи, ліцензії, патенти; податковий номер або реєстраційний номер облікової картки фізичної особи; вид дозвільного документа; вид діяльності, на провадження якої видано дозвільний документ; дату видачі дозвільного документа; строк дії дозвільного документа, інформацію про припинення (зупинення) дії дозвільного документа із зазначенням підстав такого припинення (зупинення); сплату належних платежів за видачу дозвільного документа; перелік місць провадження діяльності, на яку видано дозвільний документ;

10) про експортні та імпортні операції платників податків;

11) від банків, інших фінансових установ – інформація про наявність та рух коштів на рахунках платника податків;

12) від органів влади інших держав, міжнародних організацій або нерезидентів;

13) за наслідками податкового контролю;

14) іншу інформацію, оприлюднену як таку, що підлягає оприлюдненню відповідно до законодавства та/або добровільно чи за запитом надана контролюючому органу в установленому законом порядку.

Не менш важливим напрямом реалізації інформаційної функції податкового відомства є формування і облік *інформації за результатами контрольно-перевірочної роботи*.

Така інформація надається контролюючим органам періодично або на окремий письмовий запит контролюючого органу у визначені терміни. Інформація контролюючим органам може надаватися

періодично. До неї, наприклад, належить розкриття складових податкового кредиту та податкових зобов'язань у розрізі контрагентів.

Контролюючі органи мають право звернутися до платників податків та інших суб'єктів інформаційних відносин із письмовим запитом про подання інформації (вичерпний перелік та підстави надання якої встановлено законом), необхідної для виконання покладених на контролюючі органи функцій, завдань, та її документального підтвердження.

Платники податків та інші суб'єкти інформаційних відносин зобов'язані подавати інформацію, визначену в запиті контролюючого органу, та її документальне підтвердження.

Зважаючи на зазначене вище, потрібно зауважити, що інформаційна функція податкового відомства реалізовується у двох основних напрямах:

1) інформаційно-обліковий;

2) інформаційно-аналітичний;

Водночас необхідно зазначити, що податкова інформація, зібрана за допомогою інформаційно-облікового, інформаційно-аналітичного напрямів, накопичується і зберігається в базах даних інформаційних систем податкового відомства. Станом на 01.12.2018 функціонували такі інформаційні системи ДФС України (електронні сервіси):

1. «Електронний кабінет платника», до якого включені розділи «Податковий календар», «Стан розрахунків з бюджетом», «Листування з органами ДФС», «Облікові дані», «Пошук інформації в реєстрах ДФС», «Подання заяв/запитів», «Подання декларацій про майновий стан і доходи в електронному вигляді», «Сплата податків (фізичним особам)», «Перегляд даних СЕА/ПДВ», «Інші електронні сервіси».

2. «Інформація з реєстрів», що містить розділи «Дані про взяття на облік платників», «Реєстр страхувальників», «Дані реєстру платників ПДВ», «Реєстр платників єдиного податку», «Інформація про РРО», «Інформація про книги ОРО», «Дані реєстру платників акцизного податку з реалізації пального», «Реєстр осіб, які здійснюють операції з товарами».

3. «Митна статистика», що формується на основі системи «Зведені дані митної статистики зовнішньої торгівлі України».

4. «Дізнайся більше про свого партнера», що включає розділи «Перевірка потенційних бізнес-партнерів щодо сумлінності сплати податків до бюджету».

5. «Загальнодоступний інформаційно-довідковий ресурс» – «Запитання – відповіді з бази знань», «Нормативні та інформаційні документи», «Останні зміни в законодавстві», «Письмові консультації», «Актуальні запитання», «Ставки податків та зборів», «Інформація з питань митної справи», «Інтерактивна мапа», «Необхідно знати», «Контакт-центр ДФС», «Сервіс «Пульс», «Електронний цифровий підпис».

6. «Антикорупційний сервіс «Пульс» – «Можливість цілодобового повідомлення про неправомірні дії або бездіяльність працівників ДФС».

7. «Акредитований центр сертифікації ключів» – «Надання послуг електронного цифрового підпису».

8. «Електронна звітність» – «Спеціалізоване клієнтське програмне забезпечення для формування та подання звітності до «Єдиного вікна подання електронної звітності» – «Підтримка користувачів».

9. «Електронна митниця» – «Інформація щодо багатофункціональної комплексної системи «Електронна митниця».

10. «Дані реєстрів волонтерів АТО» – «Інформація про фізичних осіб-благодійників».

11. «Декларування громадян» – «Податкова декларація про майновий стан і доходи».

Отже, нині інформаційна функція податкового відомства України є важливою складовою діяльності даної інституції, що реалізовується шляхом інформаційно-облікового та інформаційно-аналітичного напрямів. Податкова інформація, зібрана за допомогою інформаційно-облікового та інформаційно-аналітичного напрямів, зберігається у базах даних інформаційних систем податкового відомства і використовується для прогнозування і прийняття рішень.

1.1.3. Інтегрована інформаційно-телекомунікаційна система податкового відомства України

Структурно ДПС України є складним розподіленим багатопрофільним комплексом, який поєднує інспекції різних рівнів, пов'язаних безпосередньо або опосередковано з центральним апаратом. Цей комплекс обслуговує підприємства – юридичні особи та фізичні особи платників податків, тому, з погляду інформаційних процесів, цей комплекс характеризується великою кількістю типів документів різноманітної складності, що циркулюють як в самому комплексі, так і виходять за його межі або приймаються ззовні і створюють для обробки та передачі великі обсяги інформації.

Під інтегрованою інформаційно-телекомунікаційною системою податкового відомства будемо розуміти систему програмно-технічних, методичних і організаційних компонентів, які забезпечують створення, передавання, обробку, аналіз та обмін інформацією, що використовується у щоденній роботі його підрозділів. Комплекс технічних і програмних засобів, що забезпечує обмін корпоративною інформацією, є корпоративною мережею *податкового відомства*. Станом на 01.12.2018 корпоративна мережа ДФС України функціонувала на основі розподілених мереж, об'єднаних за допомогою каналів зв'язку в єдину систему збору та обробки інформації.

Інтегрована інформаційно-телекомунікаційна система податкового відомства України як основа єдиного інформаційного простору відомства включала в себе такі основні компоненти:

– корпоративну телекомунікаційну систему та комунікаційне програмне забезпечення (ПЗ);

– засоби організації колективної роботи співробітників (*Groupware*) – автоматизовані телекомунікаційні та інформаційні підсистеми забезпечення операційних процесів;

– програмне забезпечення (ПЗ) для оперативного аналізу інформації і підтримки прийняття рішень (аналітична система);

– інформаційні ресурси і механізми надання інформації на їх основі:

– інформаційну систему управління електронним документообігом;

– організаційну інфраструктуру, що забезпечує функціонування і розвиток *інформаційного середовища*;

– систему підготовки та перепідготовки фахівців і користувачів *інформаційного середовища*;

– інші підсистеми, що забезпечують допоміжні процеси (бюджетування, складання кошторису, правовий супровід діяльності відомства тощо).

Територіально віддалені мережі податкового відомства об'єднуються в єдину віртуальну приватну мережу (далі – VPN) за допомогою використання тунелів між пограничними маршрутизаторами внутрішньої мережі. Особливість розглянутої технології полягає у застосуванні тунелів MPLS. Перевагою тунелів MPLS VPN є автоматичний спосіб їх прокладання. Вигоди, одержані за рахунок застосування технології MPLS, – це прискорене пересування трафіка по мережі провайдера і керування якістю обслуговування (QoS) для тунелів з інжинірингом трафіка. Вказані принципи побудови MPLS VPN находять втілення у реальній мережі за допомогою декількох специфічних механізмів і чітко визначеного поводження усіх компонентів мережі. Насамперед, мережа MPLS VPN поділяється на дві області: мережа IP-клієнтів і внутрішня (магістральна) мережа MPLS провайдера, яка необхідна для об'єднання мереж клієнтів (рис. 1.5).

Для корпоративної мережі податкового відомства характерні такі типи логічних мереж (рис. 1.5):

– внутрішня мережа Intranet;

– зовнішня мережа Internet;

– мережа зберігання даних;

– мережа керування інфраструктурою;

– мережа тестування та впровадження;

– голосова/телефонна мережа.

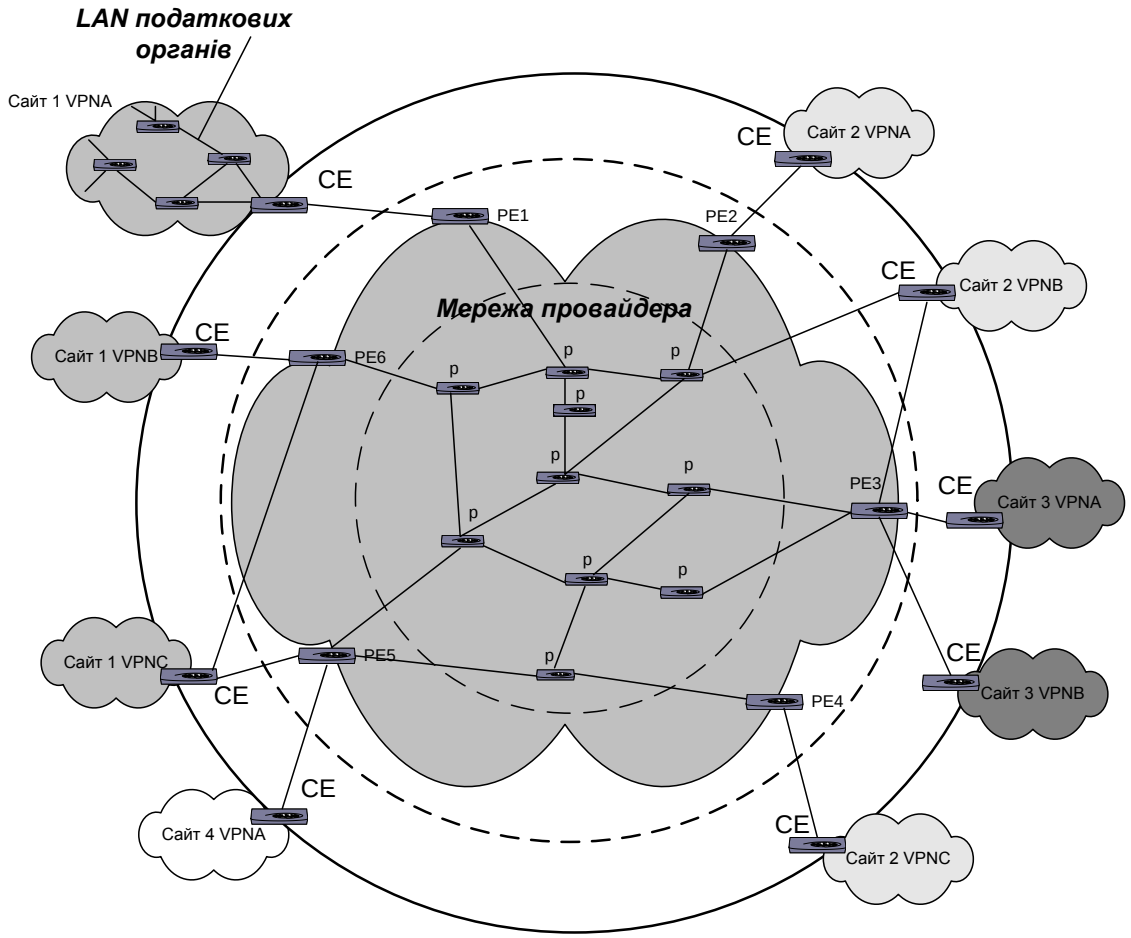


Рис. 1.5. Компоненти MPLS VPN корпоративної мережі податкового відомства України

Внутрішня мережа Intranet. Intranet – це внутрішня (логічна) мережа, заснована на Internet та World Wide Web технології, що використовує локальну (фізичну) мережу для оперативного обміну інформацією.

Необхідною умовою для створення мережі Intranet є підтримка протоколів стеку TCP/IP в локальних мережах. Використання існуючих на сьогодні Web-серверів та програмних застосувань робить Intranet платформо незалежним, внаслідок чого зменшується вартість впровадження традиційних клієнт-серверних програмних застосувань.

Використання WEB-браузера спрощує процедуру інсталяції програмних застосувань та проведення процедури оновлення, оскільки клієнт виконує доступ до цих програмних застосувань через стандартний Web-браузер. Використовуючи стандартну IP-маршрутизацію та схеми тунелювання, а також існуючі комутаційні технології (такі як Layer 3 switching), Intranet мережа може утворюватися

на основі Internet або мережі постачальника комунікаційних послуг. Ці з'єднання розглядають як захищені канали зв'язку та можуть бути утворені на основі IP-тунелів, захищеність даних в яких досягається завдяки використанню IP Security Protocol (IPSec).

Віртуальні приватні мережі. Забезпечення можливостей доступу до корпоративних ресурсів податкового відомства через Internet користувачів, які знаходяться поза периметром його захищеної мережі (зокрема, працівників та організацій-партнерів), вимагає впровадження віртуальних приватних мереж (VPN – Virtual Private Network).

Створення віртуальних приватних мереж вимагає використання комбінації технологій тунелювання, автентифікації та криптування для встановлення сесій приватних мереж через публічну мережу, таку як Internet. Використання віртуальних приватних мереж забезпечує вирішення поточних проблем, пов'язаних з необхідністю використання інфраструктури віддаленого доступу на низькошвидкісних ділянках корпоративної мережі податкового відомства або віддалений доступ через Internet. Організація віртуальних приватних мереж є важливою складовою в організації інтелектуальних мереж, що буде інтегрована як з послугою каталогів, так і з інфраструктурою безпеки. Конфігурація VPN для вирішення завдань ДПС України може бути розділена на дві основні категорії: Intranet VPN (комутація LAN до LAN) та організація віддаленого доступу через VPN.

Intranet VPN (комутація LAN до LAN) – об'єднання офісу ДПС, що безпосередньо підключений до корпоративної мережі ДПС, з підпорядкованим офісом у цілісну мережну інфраструктуру ДПС. На рис. 1.6 показано принцип, згідно з яким повинно відбуватися підключення LAN до LAN.

При такому підключенні пакети даних, що передаються з однієї приватної мережі до іншої, спочатку шифруються та інкапсулюються в пакети із загальною IP адресою (public IP) віддаленого VPN-сервера. Ця маршрутна інформація дозволяє зашифрованим та завантаженим у тунель даним приватної мережі досягати їх кінцевої точки. На прийомній стороні VPN сервером видаляється IP-заголовок загальної мережі (public header) та розшифровується інформація приватної мережі відправника.

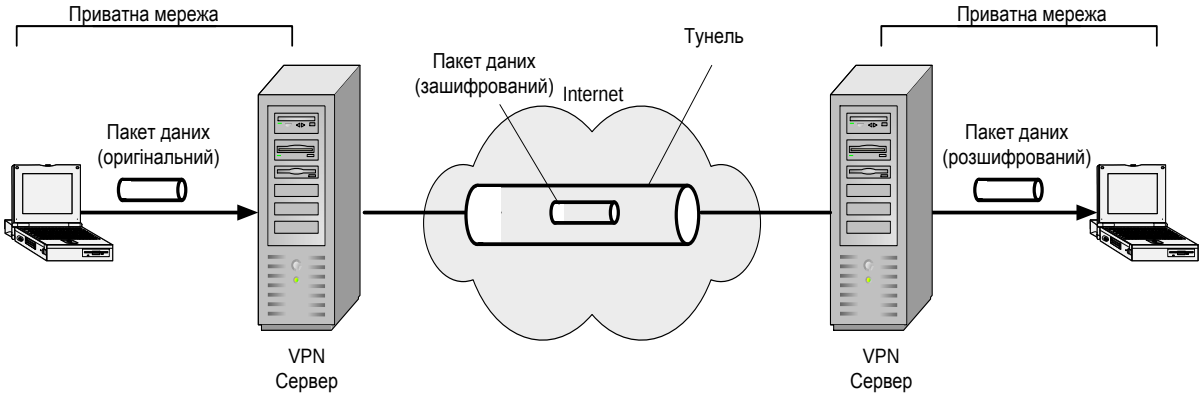


Рис. 1.6. Організація підключення LAN – LAN

1.1.4. Інформаційно-облікові та сервісні системи податкового відомства України

Усю діяльність податкового відомства України підтримують такі автоматизовані, інформаційно-облікові та сервісні телекомунікаційні системи (ІТС, АІС):

– ІТС «Єдина автоматизована інформаційна система» (митний блок);

– ІТС «Податковий блок»;

– ІТС «Єдине вікно подання електронної звітності»;

– ІТС «Управління документами»;

– ІТС «Офіційний веб-портал»;

– ІТС «Державний реєстр фізичних осіб – платників податків»;

– АІС «Скарга»;

– АІС «Суди»;

– АІС «Реєстрація розрахункових операцій (РРО)»;

– Електронний сервіс «Електронний кабінет платника»;

– Спеціалізоване програмне забезпечення для подання електронної звітності;

– Автоматизована інформаційна система попереднього обміну інформацією між ДМК Республіки Білорусь та Державною фіскальною службою України «PRINEX UA».



Рис. 1.7. Поточний та очікуваний стан інформаційно-аналітичного середовища податкового відомства України після модернізації

На рис. 1.7 подано структурні схеми поточного і очікуваного стану інформаційної системи податкового відомства після завершення модернізації. Комплексна система захисту інформації (далі – КСЗІ) дозволить інтегрувати розрізнені на сьогодні підсистеми. Основні функції ІТС «Єдина автоматизована інформаційна система (митний блок)»:

– автоматизація процесів митного оформлення та митного контролю, здійснення контролю доставки вантажів, ведення митних класифікаторів;

– ведення протоколів про порушення митних правил та контрабанди, збереження електронних копій ВМД;

– обмін інформацією із суб'єктами інформаційних відносин.

До складу ІТС **«Єдина автоматизована інформаційна система»** входять: АС «Інспектор»; система корпоративної електронної пошти.

Модуль «**Єдине вікно**» в АСМО «**Інспектор**» включає: систему обміну електронними документами та повідомленнями між

декларантами, митницями і контролюючими органами (санітарно-епідеміологічний, ветеринарно-санітарний, фітосанітарний, екологічний, радіологічний види контролю).

До основних операційних процесів податкового відомства України відносять: 1) облік платників податків; 2) обробку податкової звітності та платежів; 3) облік платежів; 4) податковий аудит. Зазначені основні процеси забезпечує ІТС «Податковий блок», до якої входять відповідні автоматизовані інформаційні підсистеми.

ІТС «Податковий блок» – ключове досягнення проекту «Модернізація державної податкової служби України – 1», впровадження якої було здійснено у 2012 році. ІТС «Податковий блок» розроблено за принципом централізованого накопичення, обробки, систематизації та збереження даних з наданням доступу до них фахівцям податкового відомства всіх рівнів у межах їх функціональних обов'язків. ІТС «Податковий блок» замінила ряд АРМів, які використовувалися у податковому відомстві. Таким чином, зникла потреба багаторазового повторювання у багатьох територіальних підрозділах тих самих даних, витрачаючи значні ресурси на адміністрування цих даних. У базу даних ІТС «Податковий блок» дані заносяться через веб-інтерфейс. Система призначена для автоматизації основних операційних процесів.

До складу ІТС «**Податковий блок**» входять автоматизовані інформаційні підсистеми (АІП):

– «Облік платників податків»;

– «Обробка податкової звітності та платежів»;

– «Облік платежів»;

– «Податковий аудит»;

– «Аналітична система»;

– «Адміністрування системи».

Функції АІП **«Облік платників податків»**:

– облік платників податків (юридичних та фізичних осіб);

– облік об'єктів оподаткування та об'єктів, пов'язаних з оподаткуванням;

– ведення реєстрів платників окремих податків, реєстру суб'єктів ЗЕД тощо.

Основною функцією АІП «**Обробка податкової звітності та платежів**» є прийняття, реєстрація, обробка та збереження даних звітності, поданої платниками податків.

Функції АІП **«Облік платежів»**:

– визначення стану розрахунків з бюджетом кожного платника за кожним податком (ведення ІКПП);

– відшкодування ПДВ, повернення переплат;

– робота з податковим боргом.

Функції АІП **«Податковий аудит»**:

– формування плану-графіка перевірок;

– формування документів на всіх етапах КПР;

– відслідковування оскарження результатів КПР;

– формування та опрацювання запитів зустрічних звірок.

Основною функцією АІП **«Аналітична система»** є формування звітів та зведених інформацій.

Основною функцією АІП **«Адміністрування системи»** є керування доступом до вказаних підсистем.

ІТС **«Єдине вікно подання електронної звітності»** забезпечує автоматизацію процесу приймання та обробки звітності в електронному вигляді, що надсилається суб'єктами господарювання до центральних органів виконавчої влади та фондів загальнообов'язкового державного страхування.

ІТС «Єдине вікно подання електронної звітності» забезпечує виконання таких функцій:

– приймання звітності в електронному вигляді; ведення Єдиного реєстру податкових накладних (далі – ЄРПН) та Єдиного реєстру акцизних накладних (далі – ЄРАН);

– приймання електронних повідомлень від банків, фінансових установ про відкриття/закриття рахунків; приймання та обробку сервісних запитів;

– надсилання інформаційних повідомлень; приймання та обробка договорів про визнання електронних документів в електронному вигляді.

ІТС **«Управління документами»** забезпечує виконання таких функцій:

– реєстрація, облік документів;

– контроль виконання доручень;

– зберігання електронних копій документів (сканкопії).

Функції ІТС «Офіційний веб-портал»:

– оприлюднення інформації про діяльність податкового відомства в мережі Інтернет, доступ до окремих електронних реєстрів та сервісів.

Функції ІТС «Державний реєстр фізичних осіб – платників податків»:

– автоматизація процесу реєстрації фізичних осіб – платників у Державному реєстрі фізичних осіб – платників податків;

– ведення інформаційного фонду Державного реєстру фізичних осіб – платників податків.

Перелік відомостей ІТС «Державний реєстр фізичних осіб – платників податків» затверджено статтею 70 Податкового кодексу України.

Електронний сервіс «**Електронний кабінет платника**» (далі – ЕКП).

ЕКП − це автоматизоване робоче місце платника податків, яке надає йому такі сервіси:

(у публічній, відкритій частині)

– пошук інформації у реєстрах, які веде податкове відомство України;

– ознайомлення з нормативними та інформаційними документами з питань оподаткування, митної справи, єдиного внеску, редакціями відповідей на запитання, з якими звертаються клієнти до контакт-центру;

– отримання інформації про надання послуг електронного цифрового підпису Акредитованим центром сертифікації ключів;

– отримання реквізитів рахунків, на які сплачуються податки, збори, платежі, єдиний внесок;

– складання та подання звернень і запитів для отримання публічної інформації.

(у персональній, закритій частині)

– подання звітності в електронному вигляді, перегляд реєстраційних та облікових даних;

– перегляд даних про стан розрахунків з бюджетом;

– подання заяв, повідомлень, листів, формування запитів для отримання інформації;

– сплата податків за допомогою платіжних систем;

– отримання відомостей про суми виплачених доходів та утриманих податків;

– отримання інформації щодо митного оформлення.

Дані сервіси платник податків може отримати за допомогою будь-якого комп'ютера, під'єднаного до Інтернет.

Спеціалізоване програмне забезпечення для подання електронної звітності забезпечує формування та відправлення до ІТС «Єдине вікно подання електронної звітності» податкових та акцизних накладних, сервісних запитів, договорів про визнання електронних документів в електронному вигляді, а також отримання інформаційних повідомлень, зокрема квитанцій.

Автоматизована інформаційна система попереднього обміну інформацією між ДМК Республіки Білорусь та Державною фіскальною службою України «PRINEX UA» забезпечує автоматизацію інформаційного обміну відомостями щодо переміщення товарів і транспортних засобів через митний кордон Республіки Білорусь і України.

Інформаційні ресурси податкового відомства України, що забезпечують формування періодичних звітів:

– АІС «Галузь» (звітність про надходження, нарахування, податковий борг);

– АС «Фактичні надходження до бюджетів» (звітність за даними Державного казначейства);

– «Регламентована звітність» (звітність з єдиного соціального внеску, відстрочок, ПДВ, результатів контрольно-перевірочної роботи);

– інформація про кількість працюючих та суми нарахованих/виплачених доходів за даними 1-ДФ;

– система підтримки та прийняття рішень;

– формування звітності зіставлення податкових зобов'язань та податкового кредиту для регіонів;

– ризики ПДФО (відомості про ризикових платників юридичних осіб та самозайнятих осіб).

Доступ до інформаційних ресурсів податкового відомства України регламентується нормативами. Зокрема, наказом ДФС від 24.11.2015 № 934 визначено: 1) перелік систем, до яких надається доступ користувачам; 2) формат службової записки про надання прав доступу до інформації в системі. Наказом ДФС від 18.08.2015 № 618 визначено: 1) перелік ролей користувачів; 2) формат запиту на внесення змін до переліку ролей та початкового набору ролей; 3) формат службової записки про надання дозволу на роботу з інформаційними ресурсами. Створення та удосконалення програмного забезпечення регламентується наказом ДПС України від 12.10.2011 № 54, яким визначають: 1) формат заявки на створення (удосконалення) програмного забезпечення; 2) заявку на створення (удосконалення) програмного забезпечення – формалізовані вимоги замовника до програмного забезпечення.

Аналіз функціональності діючих автоматизованих інформаційно-телекомунікаційних систем податковго відомства України, особливо таких, як АІП «Аналітична система» показує актуальність завдань із розширення їхніх аналітичних функцій для забезпечення прийняття ефективних управлінських рішень. Інтеграція в інформаційній системі податкового відомства компонентів оцінювання факторів, що впливають на якість надання платникам податків відповідних послуг, застосування методів прогнозування, пошуку нових знань і прихованих закономірностей у накопичених масивах даних, підвищить ефективність діяльності як структурних підрозділів, так і відомства загалом.

Центральна база даних (далі – ЦБД) є єдиним джерелом інформації для всіх користувачів інформаційної системи, що дозволяє робити їхню роботу більш скоординованою, більш ефективною. При цьому завдяки функціональним можливостям платформи СКБД Oracle кожному користувачу ІТС «Податковий блок» призначено відповідну роль, яка визначає доступ користувача до функціональних можливостей та даних.

Одним з користувачів системи є платник податків. Для платників податків створено електронний сервіс «Електронний кабінет платника податків» (далі − ЕКПП), за допомогою якого платник податків зі свого робочого офісу через Інтернет в режимі онлайн зможе декларувати свої податкові зобов’язання та переглядати

інформацію про стан розрахунків з бюджетом. Впровадження ІТС «Податковий блок» відкрило широкі можливості для запровадження багатьох сервісів для працівників податкового відомства та платників податків.

Питання для самоконтролю до теми 1.1

1. Опишіть податкову систему України як систему управління.
2. Що розуміють під поняттям «система»?
3. Опишіть податкову систему як комплекс взаємодіючих рівнів державного управління.
4. Опишість структуру рівнів органів податкового відомства України.
5. Які два організаційні варіанти передбачається під час створення державних податкових інспекцій?
6. Назвіть основні завдання органів державної фіскальної служби.
7. Які функції покладено на податкове відомство і яким документом вони регламентовані?
8. Назвіть основні складові інформаційно-аналітичного середовища податкового відомства України.
9. Що розуміють під корпоративною інформаційною системою?
10. Які компоненти включає інтегроване інформаційно-аналітичне середовище податкового відомства України?
11. Особливість технології MPLS.
12. Вкажіть компоненти MPLS VPN корпоративної мережі податкового відомства України.
13. Які типи характерні для корпоративної мережі податкового відомства?
14. Чим відрізняється внутрішня мережа Intranet податкового відомства України від зовнішньої Інтернет?
15. У чому суть віртуальних приватних мереж?
16. На які дві категорії може бути розподілена конфігурація VPN для вирішення завдань податкового відомства?
17. Як здійснюється організація з'днання та підключення територіальних органів податкового відомства за схемою LAN–LAN?

1.2. Операційні процеси та інформаційні підсистеми податкового відомства

1.2.1. Операційний процес «Облік платників податків»

Підсистема «Облік платників податків» як складова частина ІТС «Податковий блок» забезпечує узагальнення усієї інформації про реєстрацію та облік платників податків у єдиній базі даних (рис. 1.8).

Операційний процес підсистеми «Облік платників податків» складається із таких процесів: ідентифікація фізичних осіб; взяття на облік платників податків; облік об'єктів господарювання; реєстрація платників окремих податків.



Рис. 1.8. Операційний процес «Облік платників податків»

Єдина база даних з централізованим управлінням із забезпеченням доступу до неї усіх підрозділів податкового відомства через корпоративну мережу та платників податків через електронний

сервіс «**Електронний кабінет платника податків**» забезпечує платникові податків:

– можливість звертатися за отриманням податкових послуг до органів державної податкової служби у режимі on-line: повідомляти про зміни своїх облікових даних, подавати заяву про реєстрацію платником ПДВ, платником єдиного податку;

– спрощення постановки на облік у зв'язку з відміною необхідності відвідувати податковий орган з метою отримання документа про взяття на облік (для платників, відомості щодо яких містяться в ЄДР);

– спрощення постановки на облік за неосновним місцем обліку в податковому органі за місцезнаходженням об'єкта оподаткування;

– спрощення процедур переходу на облік до іншого податкового органу при зміні місцезнаходження платника податків.

Підрозділам ДПС України:

– спрощення процедури обробки відомостей, що надходять від держреєстраторів у зв'язку з реєстрацією, припиненням, внесенням змін у реєстраційні дані платників податків, створенням (ліквідацією) відокремлених підрозділів тощо;

– доступ до всебічної та актуальної інформації про платників податків, що дозволить краще здійснювати відбір платників податків для проведення податкового аудиту, здійснювати аналіз економічних та законодавчих тенденцій і змін;

– спрощення процедури визначення переліку податків та платежів, належних до сплати платникові податків;

– спрощення процедур з припинення платників податків;

– спрощення процедур зі зміни контролюючого податкового органу у зв'язку зі зміною місцезнаходження платника податків;

– скасування необхідності здійснення обміну такими документами між податковими органами (у разі законодавчо закріпленої норми):

– повідомлення про взяття на облік / зняття з обліку платника податків у зв'язку зі зміною місцезнаходження (місця проживання) за ф. № 11-ОПП;

– повідомлення про взяття на облік, створення, унесення змін до відомостей, закриття відокремленого підрозділу платника податків за ф. № 17-ОПП;

– скасування необхідності щорічного перевидання свідоцтва про реєстрацію платником єдиного податку.

1.2.2. Операційний процес «Обробка податкової звітності та платежів»

Платники податків мають можливість подавати податкову звітність у будь-який зручний для них спосіб: у паперовому або в електронному вигляді, особисто, поштою, на магнітних носіях, електронною поштою. Також для платників податків додано нову можливість – подавати звітність у режимі реального часу. Електронний сервіс «Електронний кабінет платника податків» є web-застосуванням клієнтської частини системи інформаційної «Податковий блок» для платника податків.

З 2014 року в електронному кабінеті платника податків запроваджено формування та заповнення платіжних документів щодо сплати податків і зборів (обов'язкових платежів), забезпечення їх сплати до бюджету з використанням банківських платіжних систем, офіційне листування з органами ДФС з питань оподаткування та он-лайн консультування платників податків. Операційний процес підсистеми «Приймання податкової звітності» складається із підпроцесів: DP-01 – прийняття, введення та контроль податкової звітності; DP-02 – контроль введення платежів; DP-03 – обробка податкової звітності та платежів (рис. 1.9).

Переваги централізованої системи приймання та обробки податкової звітності:

– автоматична актуалізація програмного забезпечення на етапі запуску;

– забезпечення роботи з пакетом податкової звітності − набором файлів електронних документів податкової звітності, між якими встановлено зв'язок щодо їх належності до однієї податкової декларації;

– надання на кожний прийнятий пакет документів однієї квитанції про прийняття податкової декларації з додатками;

– забезпечення не лише подання звітності в режимі он-лайн з одночасним отриманням квитанції про її приймання, а й можливості перегляду поданої звітності та відображення інформації: як подана звітність вплинула на стан розрахунків платника податків з бюджетом.



Рис. 1.9. Операційний процес «Обробка податкової звітності та платежів» та його підпроцеси

1.2.3. Операційний процес «Облік платежів»

Особовий рахунок платника податків – це екранна форма, яка надає можливість користувачу бачити інформацію про стан розрахунків платника податків з бюджетом, наявну в системі обліку платежів (рис. 1.10).

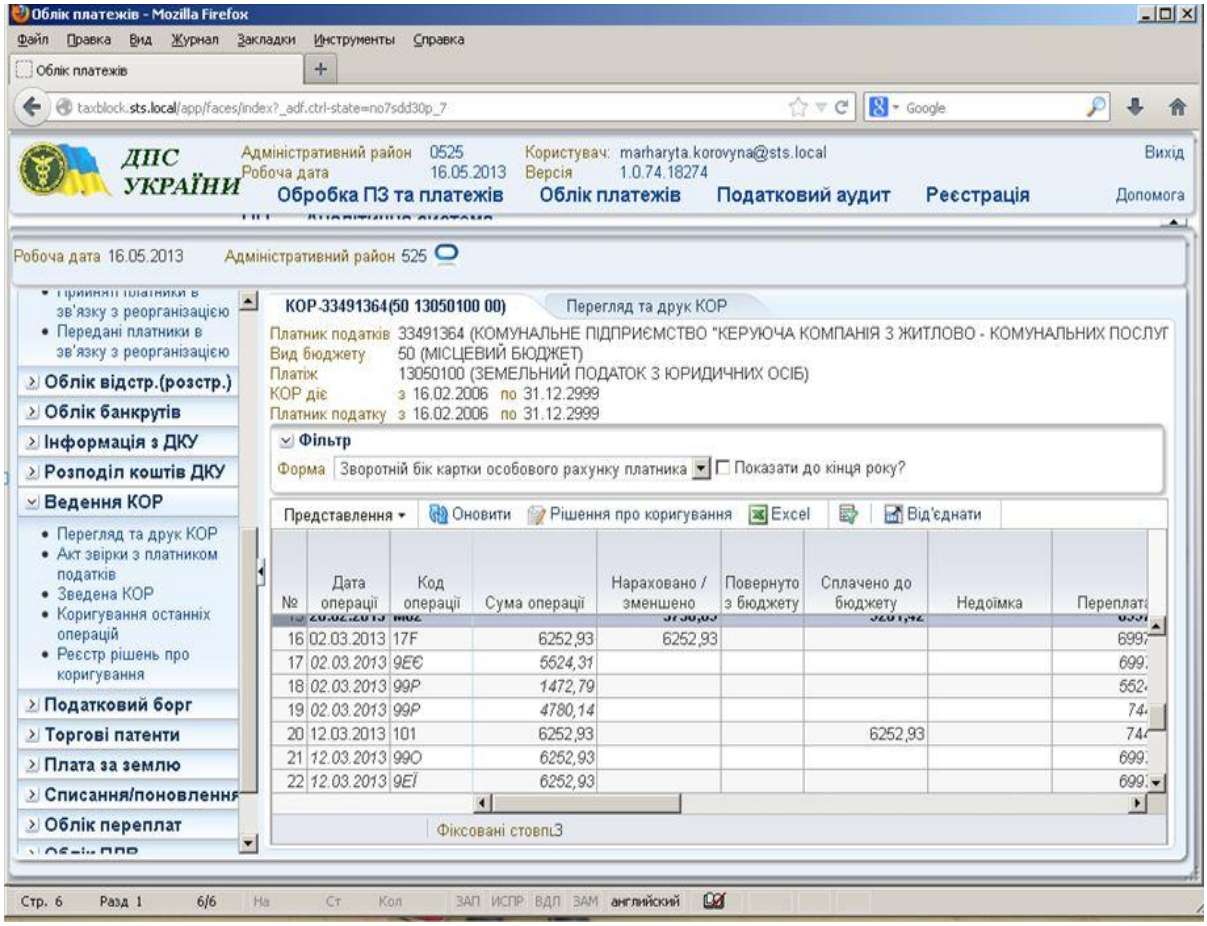


Рис. 1.10. Інтерфейс особової картки платника

Під час перегляду картки особового рахунку в підсистемі «Облік платежів» користувач зможе:

– задати період для перегляду;

– «сховати» непотрібні для перегляду стовпці;

– побачити грошові зобов'язання, за якими не настав граничний строк сплати (до кінця поточного року);

– побачити реєстраційні дані платника податків;

– переглянути документи, на підставі яких проведено операції в КОР (наприклад, подану платником податків податкову звітність тощо);

– отримати інформацію щодо періоду призначення платежу;

– переглянути інформацію лише про переплату (історія переплат) або податковий борг (історія недоїмки);

– побачити інформацію щодо розрахунку пені та штрафу за порушення правил сплати тощо.

Для зручності та оперативності доступу до інформації попередніх років закриття особових рахунків не здійснюється. При

ліквідації, зміні форми власності, назви платника податків, його переходу на сплату консолідованого податку тощо в особових рахунках проставляється дата закінчення терміну призначення платежу.

Операційний процес «Облік платежів» складається з підпроцесів (рис. 1.11):

– ТА-01.Здійснення розрахунків в особових картках платників податків;

– ТА-02. Облік податкових зобов’язань;

– ТА-03. Облік погашення податкових зобов’язань та повернутих коштів;

– ТА-04. Розрахунок та застосування штрафних санкцій;

– ТА-05. Облік та адміністрування переплати;

– ТА-06. Нарахування та облік пені;

– ТА-07. Облік податкового боргу;

– ТА-08. Ведення особових рахунків платників податків та зведених рахунків з платежів.

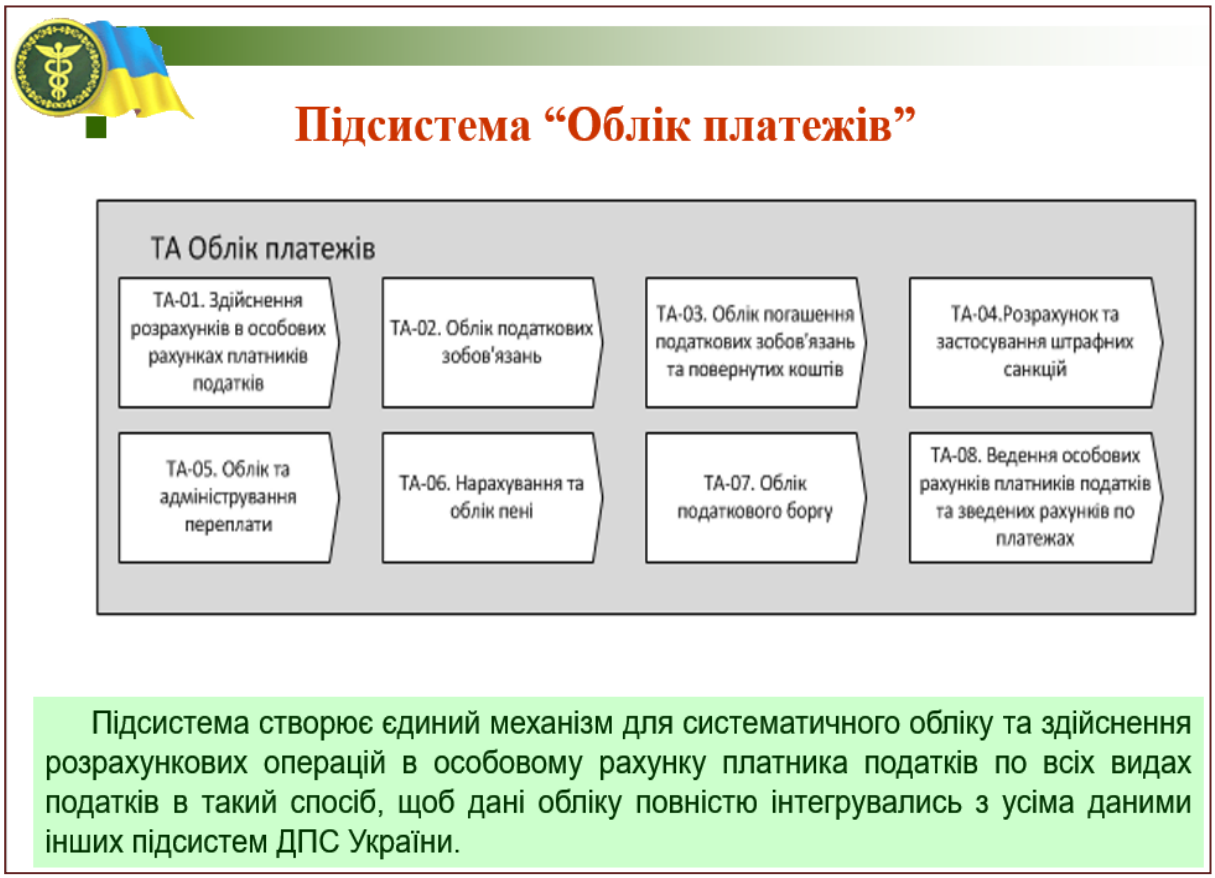


Рис. 1.11. Операційний процес «Облік платежів» та його підпроцеси

У випадку, коли згідно з чинним законодавством платник податків має сплачувати грошові зобов'язання за податком, збором у різних районах/областях України (наприклад, плата за землю), система формує зведену картку за таким податком, збором. Зведена картка надає можливість користувачу підсистеми переглянути інформацію про стан розрахунків платника за таким податком взагалі.

Зведена картка формується також для платників податків, які мають філії та відокремлені підрозділи, та надає можливість переглянути інформацію про стан розрахунків з бюджетом платника податків, включаючи всі його філії та підрозділи, не залежно від того, в якому органі ДФС вони перебувають на обліку.

Працюючи зі зведеною карткою платника податків, користувач може переглянути кожну з його окремих карток особових рахунків (далі – КОР) та КОР його відокремлених підрозділів.

1.2.4. Операційний процес «Податковий аудит»

Сьогодні актуальним є питання зменшення втручання контролюючих органів у діяльність платників податків. Тому в останні роки держава приділяє багато уваги регулюванню процесів проведення податкового аудиту, конкретизації процедур його проведення, прав та обов'язків як контролюючих органів, так і платників податків.

«Податковий блок» як інструмент, за допомогою якого податковий інспектор має можливість здійснювати в автоматичному режимі опрацювання великого масиву інформації, полегшує працю інспектору та скорочує час, витрачений на проведення перевірки.

Операційний процес «Податковий аудит» складається із підпроцесів (рис. 1.12):

– AD-04. Призначення перевірки;
– AD-05. Підготовка до податкового аудиту;
– AD-06. Документальна невиїзна перевірка;

– AD-07. Проведення податкового аудиту;

– AD-08. Оформлення матеріалів податкового аудиту;

– AD-10. Організація аналізу роботи підрозділів податкового аудиту.

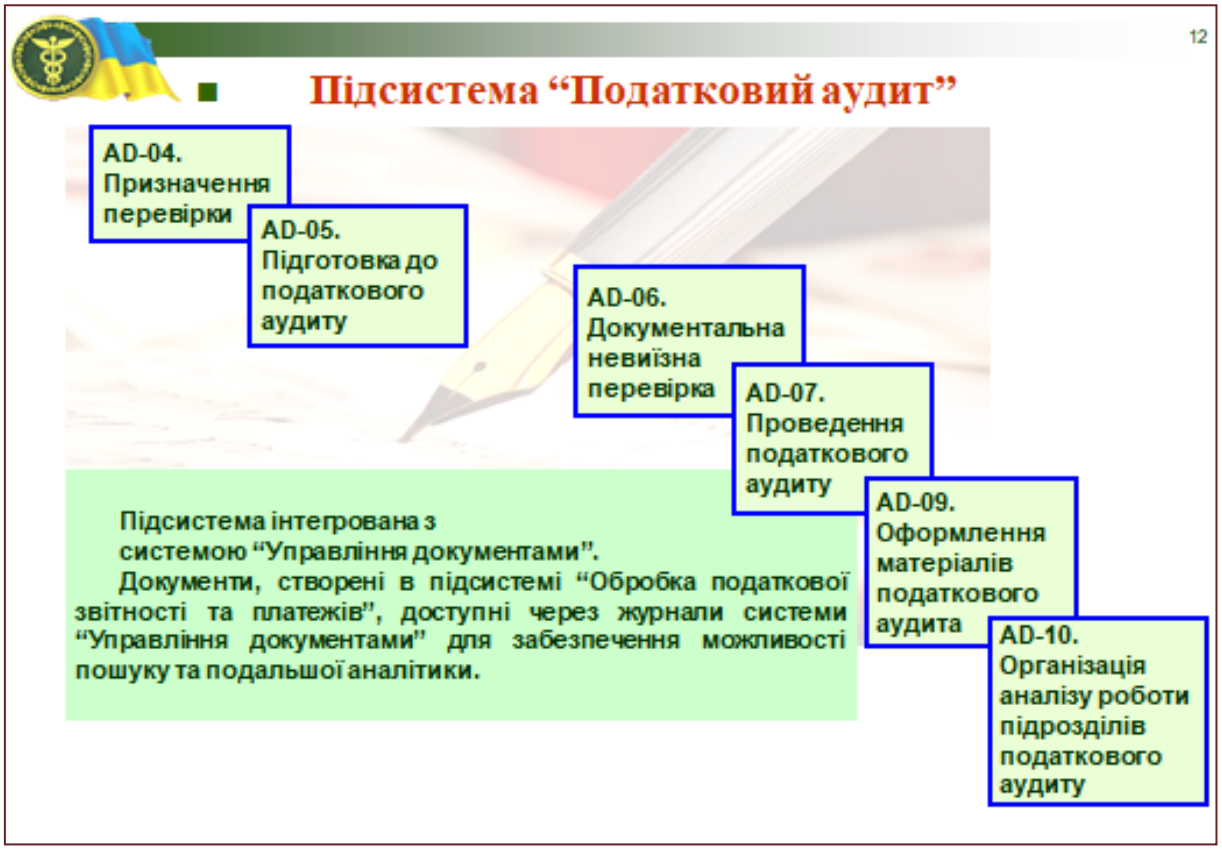


Рис. 1.12. Операційний процес «Податковий аудит» та його підпроцеси

Основними перевагами оновленої підсистеми «Податковий аудит», порівняно з попередньою системою АІП «Аудит», є:

– автоматизація процесів щодо формування у системі практично усіх документів на усіх етапах КПР, зокрема долучення інших документів;

– можливість формування та опрацювання у системі запитів зустрічних звірок;

– створення паспорта перевірки;

– формування звітів «Аудит 1», «Аудит 2п»;

– здійснення моніторингу КПР.

«Паспорт перевірки» забезпечує документування результатів перевірки на кожному етапі проведення перевірки, оперативний

контроль та використання матеріалів перевірок для аналізу діяльності підприємств у напрямі реалізації контрольно-перевірочних заходів.

«Паспорт перевірки» є базою накопичення та систематизації інформації щодо окремого платника, яку отримано з різних джерел, що циркулює в інформаційних системах (опис платника податків, історії його податкової та платіжної поведінки, фінансово-господарської діяльності та актуальних проблем, пов'язаних з оподаткуванням, тощо).

1.2.5. Аналітична система та система адміністрування доступу до даних

Підсистема «Аналітична система»

«Аналітична система» дає можливість отримати доступ до АІС «Скарга», АІС «Суди», АІС «Табель інспекції». Підсистема має режими: «Суб'єкти фіктивного підприємництва», «Співставлення додатків 5 до Декларації з ПДВ», є деякі інформації та аналітичні довідки. Щодо наявності у цій підсистемі суто аналітичних можливостей, зокрема прогнозування, класифікації, регресії та інших методів, то можна констатувати їх відсутність. Розробка й апробація перерахованих методів та програмних засобів в інформаційну систему податкового відомства України є актуальним завданням для підрозділів, що відповідають за формування аналітичних довідок керівництву податкового та інших відомств.

Підсистема «Адміністрування системи»

Підсистема «Адміністрування системи» призначена для управління доступом користувачів до певних функцій системи «Податковий блок» та даних. Функції даної підсистеми використовують весь арсенал СКБД Oracle розмежування доступу до баз даних податкового відомства України.

Питання для самоконтролю до теми 1.2

1. Назвіть чотири основних операційних процеси податкового відомства України.

2. Які підсистеми входять до ІТС «Податковий блок»?

3. Основні функції АІП «Обробка податкової звітності та платежів».

4. Основні функції АІП «Аналітична система».

5. Основні функції ІТП «Єдине вікно подання електронної звітності».

6. Назвіть основні інформаційні ресурси податкового відомства України.

7. Якими документами регламентується доступ до інформаційних ресурсів податкового відомства України?

8. Особливості операційного процесу «Облік платників податків».

9. Призначення «Електронного кабінету платника податків».

10. Особливості операційного процесу «Обробка податкової звітності та платежів».

11. Особливості операційного процесу «Облік платежів».

12. Особливості операційного процесу «Податковий аудит».

1.3. Методи та засоби моделювання і аналізу податкової інформації

1.3.1. Класифікація методів підтримки прийняття рішень

Щодо теоретичних аспектів процедур прийняття рішень, то відомо, що вони можуть відбуватися у таких трьох умовах або ситуаціях: 1) *визначеності*, тобто наявності повної інформації про стан об'єкта, щодо якого приймається рішення; 2) *невизначеності* – повної відсутності інформації про об'єкт, щодо якого приймається рішення; 3) *ризику* – наявності інформації про можливий стан об'єкта (імовірності можливого стану).

Перша умова характерна для ситуацій, коли прийняття рішень здійснюється на основі регламентованих нормативами та інструкціями критеріїв та алгоритмів. Наприклад, у контексті діяльності органів податкового відомства України особливістю прийняття рішень стосовно порушень податкового законодавства є формалізація їх Податковим кодексом, наказами та інструкціями. Такі алгоритми, як правило, автоматизуються, а рішення формується інформаційною системою. Наприклад, нарахування штрафів, пені, формування повідомлень, висновків щодо відшкодування тощо.

Другу умову – невизначеність пов'язують з відсутністю інформації про стан або імовірності можливих значень результату (стану середовища, платника, бази оподаткування тощо). Для обґрунтування рішень в умовах невизначеності, як правило, використовують *теоретико-ігрові методи, технології імітаційного моделювання, прогнозування наслідків управлінських рішень.* Це пояснюється тим, що у більшості випадків для прийняття управлінських рішень використовують неповну і неточну інформацію, яка і утворює ситуацію невизначеності різного ступеня.

Третя умова – прийняття рішення в *умовах ризику* означає, що відома імовірність стану середовища або платника, стосовно якого приймається рішення. Тому, коли говоримо про *управління ризиками*, то йдеться, зокрема, про: *якісний аналіз ризику; формування системи кількісних показників ступеня ризику; кількісний*

аналіз ризику; моделювання та прогнозування ризику; управління ризиком, тобто здійснення дій щодо зниження його ступеня.

Система підтримки прийняття рішень (далі – СППР), функціонування якої базується на *управлінні ризиками,* дає можливість ризик-менеджерам більш ефективно використовувати ресурси, розподіляти відповідальність, покращувати результати діяльності підприємства та забезпечувати прийнятний рівень ризику.

На рис. 1.13 показано один із найпоширеніших підходів до класифікації методів підтримки прийняття рішень, що використовують у СППР. Ці методи можуть використовувати у процесах управління і прийняття рішень у діяльності органів податкового відомства.



Рис. 1.13. Класифікація методів підтримки прийняття рішень

Відповідно до такого підходу класифікації всі методи обґрунтування управлінських рішень поділяють на *кількісні* та *якісні.*

Кількісні методи (або методи дослідження операцій) застосовують, коли фактори, що впливають на вибір рішення, можна кількісно визначити та оцінити.

Якісні методи використовують тоді, коли фактори, що визначають прийняття рішення, не можна кількісно охарактеризувати або вони взагалі не піддаються кількісному вимірюванню. До якісних методів здебільшого належать *експертні методи*.

Кількісні методи, залежно від характеру інформації, яку має особа, яка приймає рішення, поділяють на:

– методи, що застосовують у ситуаціях, що відповідають *першій умові* прийняття рішень – *однозначної визначеності* інформації про ситуацію прийняття рішення (*аналітичні методи та частково методи математичного програмування*);

– методи, що застосовують у ситуаціях, що відповідають *другій умові* прийняття рішень – *невизначеності* інформації про ситуацію прийняття рішення (*теоретико-ігрові методи,* які, залежно від того, що спричиняє невизначеність ситуації: об'єктивні обставини або свідомі дії противника, поділяють на *методи теорії статистичних рішень (критерії Вальда, Гурвіца, Севіджа та ін.)* та *методи теорії ігор*). Усе більшого визнання отримують також методи *нечіткої логіки*, що формують окрему групу;

– методи, що застосовують в умовах, що відповідають *третій умові* прийняття рішень – *імовірнісної визначеності* інформації про ситуацію прийняття рішення (*статистичні методи та частково методи математичного програмування*).

Аналітичні методи характеризуються тим, що встановлюють аналітичні (функціональні) залежності між умовами вирішення задачі (факторами) та її результатами (прийнятим рішенням). До аналітичних належить широка група методів економічного аналізу, що можуть застосовуватися для аналізу діяльності господарюючих суб'єктів, оцінювання бази оподаткування, соціально-економічного розвитку регіону.

Статистичні методи включають методи теорії ймовірностей та математичної статистики. Аналітичні підрозділи ДФС України вже використовують для прогнозування і класифікації такі методи: кореляційно-регресійний; дисперсійний; факторний; кластерний аналізи; методи статистичного контролю якості і надійності та інші. Проте накопичені в інформаційних сховища дані вже вимагають більш ефективних технологій, мов програмування, методів і засобів просіювання великих масивів даних, що належать до так званих методів машинного навчання і обробки великих даних (BigData, Machine Learning, R. Python тощо).

Методи математичного програмування. Математичне програмування – розділ математики, який включає теорію та методи рішення умовних екстремальних задач з кількома змінними. В задачах математичного програмування необхідно вибрати значення змінних (тобто параметрів управління), щоб забезпечити максимум (мінімум) цільової функції за певних обмежень. *Методи математичного програмування* можна застосовувати у процесах оптимізації бюджетування, планування оптимальних витрат на утримання територіальних органів податкового відомства, формування номенклатури послуг, мінімізації використання витратних матеріалів, регулюванні запасів для підтримки інформаційної інфраструктури, плануванні контрольно-перевірочної роботи тощо.

1.3.2. Оцінювання результативності податкового відомства методом інтегральної згортки показників

Щодо практичних аспектів розрахунків та оцінювання результатів діяльності податкового відомства, то для забезпечення порівняльності підрозділів, що функціонують у регіонах із відмінними рівнями соціально-економічного розвитку та розрізняються за податковим потенціалом, кількістю зареєстрованих підприємств і підприємців, можна запропонувати *методологію інтегральної згортки показників.*

Методологія інтегральної згортки показників застосовується тоді, коли потрібно або доцільно на основі значної кількості показників сформувати оцінку явища у вигляді одного показника. Цей показник, на відміну від простого рейтингування, дозволяє виявляти показники та фактори, що впливають на покращання/погіршення загального інтегрального показника, або критерію, виявляти причинно-наслідкові зв'язки.

Інтегральні (комплексні) показники (критерії) використовують для порівняльного аналізу результатів діяльності організацій, фінансового стану підприємств, їхньої надійності як клієнтів банків, можливості надання кредитів, підписання угод тощо. Існує адитивна (1.1) і мультиплікативна (1.2) форми інтегрального показника:

$$I = \max/\min \sum_{i=1}^{n} \pm w_i a_i \qquad (1.1)$$

$$I = \max / \min \prod_{i=1}^{n} a_i^{\pm w_i} \qquad (1.2)$$

$$\sum_{i=1}^{n} w_i = 1, \qquad 0 < w_i < 1 \qquad (1.3)$$

де I – інтегральний показник;

w_i – вагові коефіцієнти;

a_i – нормовані, однонаправлені показники;

n – кількість показників.

Якщо показники мають різні одиниці і непорівняльні діапазони вимірювань, показники необхідно нормувати. Як правило, показники нормують, переводять у діапазон [0,1] або [0–100 %] відомими формулами.

Для нормування, тобто приведення значень критеріїв у діапазон 1–100 або 0–1, можна використати класичний minimax-ний підхід:

1. Для показників, збільшення яких вказує на покращання ситуації (наприклад, при розгляді відносного показника питомої

ваги надходжень за результатами перевірок), для нормування критеріїв використовують формулу (1.4):

$$q_i = q_i(x_i) = \begin{cases} 0, & \text{при } x_i \le \min_i \\ \left(\frac{x_i - \min_i}{\max_i - \min_i}\right) \times 100, & \text{при } \min_i < x_i \le \max_i \\ 100, & \text{при } x_i > \max_i \end{cases} \qquad (1.4)$$

Значення критерію має пряму залежність від значення відносного показника: чим вище значення критерію, тим краща оцінка діяльності підрозділу.

2. Для показників, зменшення яких вказує на покращання ситуації (наприклад, при розгляді відносного показника затратності діяльності підрозділу), для нормування критеріїв використовують формулу (1.5):

$$q_i = q_i(x_i) = \begin{cases} 100, & \text{при } x_i \le \min_i \\ \left(\frac{\max_i - x_i}{\max_i - \min_i}\right) \times 100, & \text{при } \min_i < x_i \le \max_i \\ 0, & \text{при } x_i > \max_i \end{cases} \qquad (1.5)$$

Значення показника (критерію) має зворотну залежність від значення відносного показника: чим вище значення критерію, тим гірша оцінка діяльності підрозділу.

У формулах (1.4) та (1.5):

q_i – критерій;

x_i – значення відносного (якісного) показника;

$\max_i$ – максимальне значення відносного показника;

$\min_i$ – мінімальне значення відносного показника.

У 2011 році апробація методики розрахунків інтегральних показників діяльності податкового відомства здійснювалася Департаментом розвитку та модернізації ДПС України. В основу «Системи оцінювання результативних показників діяльності органів ДПС України» було покладено програмно-цільовий підхід, що базувався на Методичних рекомендаціях щодо визначення

результативних показників бюджетної програми, затверджених наказом Міністерства фінансів України «Про результативні показники бюджетної програми» від 27.10.09 № 1252. (Наказом Міністерства фінансів України «Про затвердження Примірного переліку результативних показників бюджетних програм» від 14.01.11 № 15 поновлено перелік та зміст показників.

Метою впровадження Системи оцінювання результативних показників діяльності податкового відомства є забезпечення керівництва усіх рівнів методами та засобами підтримки прийняття рішень, направлених на підвищення мотивації щодо зростання якості та ефективності адміністрування податків кадрового складу.

Нижче подані основні поняття та визначення цієї методики для розрахунків результативних показників діяльності органів податкового відомства.

Абсолютні показники – фактичні результати здійснення операційних процесів або іншої діяльності (надходження податків та платежів до бюджету, нарахування за результатами перевірок, сплата за результатами перевірок, якість обліку платників, витрати на утримання будівель).

Якісні (відносні) показники – показники, що характеризують якість здійснення операційних процесів та іншої діяльності органів податкового відомства і розраховуються як відношення двох абсолютних показників. Відносні показники є питомою вагою, відсотком, значенням у розрахунку на одного працівника, платника тощо. (У сформованому робочою групою реєстрі відносні показники іменують також критеріями, оскільки на їхній основі вже можна формувати рішення).

Критерії – якісні показники, що розраховують на основі відносних показників і які використовують для розрахунків інтегральних критеріїв прийняття рішень.

Цільовий показник діяльності підрозділу податкового відомства – максимальне або мінімальне значення відносного показника, що встановлюється експертною робочою групою на підставі розрахунків (статистики) декількох звітних періодів і досягнення якого повинні прагнути територіальні підрозділи.

Ваговий коефіцієнт – показник, що характеризує впливовість (важливість) критерію на загальний результат. Вагові коефіцієнти критеріїв формуються експертним шляхом (експертною робочою групою) із дотриманням умови: сума вагових коефіцієнтів повинна дорівнювати «1» (або 100 % при встановленні значень вагових коефіцієнтів у відсотках). У розрахунках використані вагові коефіцієнти, встановлені експертно у зв'язку із неповнотою первинних даних для обчислення всього масиву показників і критеріїв.

Інтегральний критерій результативності діяльності – показник, що характеризує результативність збору податків і платежів та економічну ефективність діяльності територіальних органів податкового відомства як результат виконання встановлених завдань.

Інтегральний критерій якості операційних процесів (процесів адміністрування податків) – показник, що обчислюється на основі критеріїв якості окремих операційних процесів і характеризує загальну якість операційної діяльності теориторіальних органів податкового відомства.

У наведеному нижче прикладі подані розрахунки інтегрального критерію результативності діяльності органів ДПС, інтегрального критерію якості операційних процесів територіальних органів, інтегрального критерію соціально-економічного розвитку регіонів, інтегрального критерію рівня порушень законодавства працівниками податкового відомства.

Для розрахунків інтегрального критерію результативності діяльності територіальних підрозділів ДПС України на основі запропонованої методології у 2010 році були використані такі показники:

1. Встановлені завдання зі збору податків та платежів до загального фонду Державного бюджету у розрізі основних податків (податок на прибуток підприємств, ПДВ, акцизний збір, податок на доходи фізичних осіб).

2. Збір податків і платежів до Загального фонду Державного бюджету всього та у розрізі основних податків.

3. Збір податків і платежів до Зведеного бюджету.

4. Фактична чисельність працівників органів ДПС України.

5. Фактичні видатки на утримання органів ДПС України.

На основі вищезазначених абсолютних показників розраховувалися такі відносні показники:

(А) показники, що характеризують результати збору податків і платежів:

1. Відносний показник виконання встановлених завдань зі збору платежів усього та у розрізі податків до загального фонду ДБ, %.

2. Відносний показник питомої ваги від'ємного відхилення від завдання у сумі невиконання по регіону (для ДПІ) та питома вага від'ємного відхилення від завдання у сумі невиконання по Україні (для регіонів).

3. Відносний показник динаміки збору податків і платежів порівняно з аналогічним періодом минулого року (в реальних цінах), %.

(Б) показники, що характеризують економічну ефективність підрозділів податкового відомства:

1. Відносний показник *затратності,* що характеризує питому вагу затрат у сумі збору податків до Зведеного бюджету з підрозділу.

2. Відносний показник *продуктивності*, що характеризує величину надходжень (збору) у розрахунку на одного працівника податкового підрозділу.

3. Відносний показник *динаміки затратності* як різниця затратності поточного періоду і попереднього періоду, що характеризує її зменшення чи збільшення.

4. Відносний показник *динаміки продуктивності* як відношення продуктивності поточного періоду до продуктивності попереднього періоду, що характеризує її збільшення чи зменшення.

На основі відносних показників розраховувалися відповідні критерії та здійснювалося їхнє нормування (рис.1.14). Для розрахунків інтегрального критерію результативності діяльності органів ДПС (2010–2011 роки) були використані вагові коефіцієнти, встановлені робочою групою експериментальним шляхом. Розрахунки здійснювалися при припущенні рівноважливості критеріїв (1.1, 1.2), (2.2, 2.3), інтегральних критеріїв (1, 2).



Рис. 1.14. Загальна схема розрахунків інтегрального критерію результативності діяльності органів податкового відомства

Для розрахунку інтегрального критерію збору податків і платежів, економічної ефективності встановлено вагові коефіцієнти (табл. 1.1, 1.2).

Такі значення вагових коефіцієнтів встановлені керуючись вимогою рівності «1» суми всіх вагових коефіцієнтів, які задіяні у розрахунку (в нашому прикладі: 0,4+0,4+0,2=1), та припущенням рівноважливості Критерію виконання завдання зі збору податків і Критерію динаміки надходжень, а також недостатньою інформативністю Критерію питомої ваги відхилення від встановленого завдання.

Таблиця 1.1

Показники та вагові коефіцієнти для розрахунку інтегральних критеріїв результативності

	Відносні показники для розрахунку нормованих критеріїв	Нормовані критерії (Формула нормалізації)	Ваговий коефіцієнт
1	**Інтегральний критерій збору податків і платежів** (сума добутків критеріїв 1.1–1.3 на вагові коефіцієнти)		**0,5**
1.1	Відносний показник виконання встановлених завдань до загального фонду	Критерій виконання завдання з надходжень до загального фонду **(формула (1.4)**	0,4
1.2	Відносний показник динаміки збору податків і платежів до загального фонду	Критерій динаміки надходжень до Зведеного бюджету **(формула (1.4)**	0,4
1.3	Відносний показник питомої ваги від'ємного відхилення від завдання	Критерій питомої ваги відхилення від завдання з надходження до Загального фонду у відхиленні по області **(формула (1.5)**	0,2
2	**Інтегральний критерій економічної ефективності діяльності органів ДПС** (сума добутків критеріїв 2.1–2.3 на вагові коефіцієнти)		**0,5**
2.1	Відносний показник затратності	Критерій затратності збору податків і платежів **(формула (1.5)**	0,2
2.2	Відносний показник динаміки затратності	Критерій динаміки затратності збору податків і платежів **(формула (1.5)**	0,4
2.3	Відносний показник динаміки продуктивності	Критерій динаміки продуктивності працівників органів ДПС **(формула (1.4)**	0,4
3	**Інтегральний критерій результативності діяльності органів ДПС** (сума добутків інтегральних критеріїв 1–2 на вагові коефіцієнти)		

Таблиця 1.2

Показники та вагові коефіцієнти, що використані в розрахунку інтегрального критерію збору податків і платежів

№ з/п	Назва показника	Позначення вагового коеф.	Значення
1.	Критерій виконання завдання зі збору податків до Загального фонду (Квз)	$w_{квз}$	0,4
2.	Критерій динаміки надходжень до Зведеного бюджету (Кдзб)	$w_{Кдзб}$	0,4
3.	Критерій питомої ваги відхилення від завдання з надходження до Загального фонду (Кнз)	$w_{Кнз}$	0,2

Розрахунок показників економічної ефективності: показника затратності (економності), продуктивності, динаміки затратності (економності), динаміки продуктивності. Для розрахунку були встановлені такі вагові коефіцієнти (табл. 1.3):

Таблиця 1.3

№ з/п	Назва показника	Позначення вагового коеф.	Значення
1.	Критерій затратності збору податків і платежів **(Кзт)**	$w_{зт}$	0,2
2.	Критерій динаміки затратності збору податків і платежів органу ДПС **(Кдзт)**	$w_{дзт}$	0,4
3.	Критерій динаміки продуктивності працівників органу ДПС **(Кдпр)**	$w_{дпр}$	0,4

Такі значення вагових коефіцієнтів встановлені керуючись вимогою рівності «1» суми всіх вагових коефіцієнтів, які задіяні у розрахунку (в нашому прикладі: 0,2+0,4+0,4=1) та припущеннях рівноважливості критерію динаміки затратності збору податків і платежів органу ДПС (Кдзт) і Критерію динаміки продуктивності працівників органу ДПС (Кдпр)**,** а також недостатньою інформативністю критерію затратності збору податків і платежів (Кзт).

Розрахунок інтегрального критерію результативності діяльності підрозділів податкового відомства. Вагові коефіцієнти для розрахунку інтегрального критерію результативності діяльності органів ДПС (табл. 1.4).

Таблиця 1.4

№ з/п	Назва показника	Позначення вагового коеф.	Значення
1.	Інтегральний критерій збору податків і платежів **(Кзп)**	$\mathbf{W_{зп}}$	0,5
2.	Інтегральний критерій економічної ефективності **(Кеф)**	$\mathbf{W_{еф}}$	0,5

Значення вагових коефіцієнтів встановлені керуючись вимогою рівності «1» суми всіх вагових коефіцієнтів, які задіяні у розрахунку (в нашому прикладі: 0,5+0,5=1) та припущеннях рівноважливості інтегрального критерію збору податків і платежів (Кзп) та інтегрального критерію економічної ефективності (Кеф). Для прикладу у табл 1.5 наведено результати використання методології інтегральної згортки загалом по ДПС України (станом на 01.10.2010).

Таблиця 1.5

	Показники результативності діяльності органів ДПС	Значення показників	Пояснення
1.	**Показники, що характеризують результативність збору податків і платежів**		
1.1.	Відсоток виконання встановлених завдань зі збору платежів всього до Загального фонду ДБ, %	82,97 %	
1.2.	Динаміка зростання (зменшення) надходжень до ЗБ порівняно з аналогічним періодом минулого року (в реальних цінах), %	115,2 %	Станом на 01.10.2010, порівняно з аналогічним періодом минулого року, до ЗБ надійшло на 15,2 % податків більше

Продовження табл. 1.5

2.	**Показники, що характеризували економічну ефективність органів ДПС**		
2.1.	Показник затратності, що характеризує питому вагу затрат у сумі збору податків до Зведеного бюджету по ДПС України, %	1,98%	100 грн податкових надходження вартують державі 2,27 грн
2.2.	Показник продуктивності, що характеризує величину надходжень (збору) у розрахунку на одного працівника ДПС (тис. грн)	2286,72	
2.3.	Показник динаміки затратності як різниця відсотків затратності 2009 і 2010 років, що характеризує її зменшення/збільшення,%	-0,14%	Затратність 2010 року зменшилася на 0,14 % (2009 – 2,12 %, 2010 – 1,98)
2.4.	Показник динаміки продуктивності як відношення продуктивності 2010 року до продуктивності 2009 року, що характеризує її збільшення/зменшення, %	16,3%	Продуктивність збільшилася на 16,3 %. У 2009 – 1970,84 тис. грн на працівника; 2010 – 2286,72 тис. грн на працівника
3.	**Інтегральний критерій результативності діяльності ДПС за даними показниками, %**	65,04 %	

Аналогічні показники були розраховані для кожного органу ДПС (станом на 2010 всього 462 інспекції та 28 обласних ДПА України). На основі статистичного аналізу розподілу значень інтегрального критерію результативності діяльності підрозділів визначено правило класифікації органів ДПС за рівнем результативності.

Таблиця 1.6

Статистичні показники інтегрального критерію результативності діяльності органів ДПС

Назва показника	Значення
Кількість органів ДПС, включаючи регіональні ДПА	490
Середнє значення інтегрального критерію	65,00
Середньо квадратичне відхилення, СКВ	7,15
Мінімальне значення	33,93
Максимальне значення	87,61

Нижче подано гістограму розподілу розрахованих значень інтегрального критерію результативності діяльності органів ДПС. Дані табл. 1.6 та графічний аналіз рис. 1.15 дає право стверджувати, що розподіл відповідає «нормальному закону» (близько 70 % об'єктів, що розглядаються, знаходяться у межах: середнє значення показника ± його середньоквадратичне відхилення):

$$\sim 58 < \text{Ікр } (65{,}0 \pm 7{,}15) < \sim 72$$

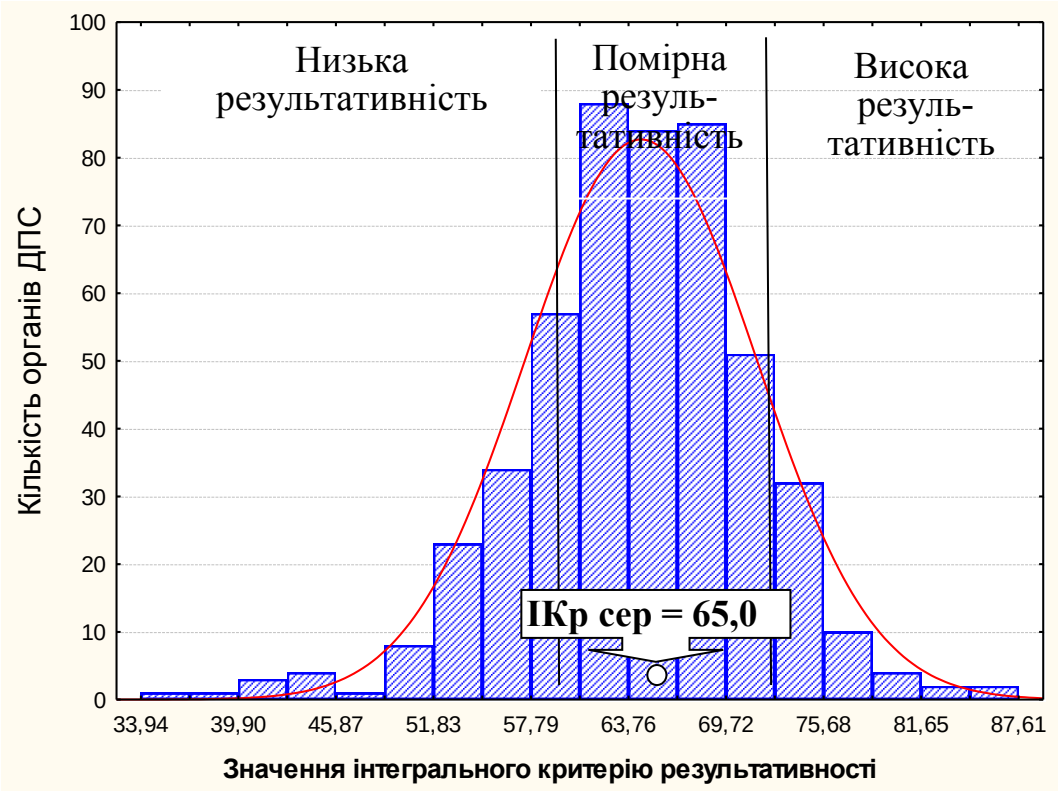


Рис. 1.15. Гістограма розподілу розрахованих значень інтегрального критерію результативності діяльності органів ДПС (2010 р.)

На основі оцінених параметрів статистичного розподілу (табл.1.6) можна формувати правила (критерії) класифікації органів податкового відомства за рівнем результативності. У цьому прикладі для 2010 року критерії класифікації формувалися на основі середнього значення Інтегрального критерію результативності діяльності органів ДПС 65,0 та середньо квадратичного відхилення (СКВ) рівним ~7,0. Відповідно до цих критеріїв органи ДПС за рівнем результативності розподілені на три основні групи (табл. 1.7).

Таблиця 1.7

Критерії розподілу органів ДПС за рівнем результативності

№ групи	Група органів ДПС за рівнем результативності діяльності	Значення інтегрального критерію
1.	Група високої результативності	Ікр > 72
2.	Група помірної результативності	58 < Ікр < 72
3.	Група низької результативності	Ікр < 58

Результати класифікації органів ДПС за значеннями інтегрального критерію результативності діяльності подані у табл. 1.8. Із таблиці 1.8 видно, що до 3-ої групи (низької результативності) входило 74 органи ДПС, тобто 16 % від загальної кількості органів ДПС на той час. Проте загальна сума невиконання встановленого завдання цими інспекціями становить 9 225628 тис. грн, або 53,14 % від усієї суми невиконання встановлених завдань зі збору до Загального фонду Державного бюджету. Це вказує на те, що саме групі органів ДПС низької результативності потрібно приділяти оперативну увагу.

Використання вищенаведеної методики і доведення до кожного органу податкового відомства інформації про його рівень результативності на основі апробованих, перевірених показників та критеріїв буде сприяти формуванню мотивації щодо покращення показників операційної діяльності. Особливо це може бути дієвим під час формування реальних стимулів для покращання показників діяльності, що може наблизити податкове відомство до класу саморегулюючих систем із ознаками самоорганізації та неперервного удосконалення.

Таблиця 1.8

Класифікація органів ДПС за значеннями інтегрального критерію результативності діяльності

№ гр.	Значення інтегрального критерію результативності	Кількість органів ДПС	Питома вага органів ДПС у загальній кількості,%	Сума відхилення від вст. завдань, (тис. грн)	Питома вага відхилення у загальній сумі по Україні, %
1.	ІКр> 72 (Група високої результативності)	57	12,3	–	–
2.	58 <Ікр< 72 (Група помірної результативності)	331	71,5	8 132 563	46,86
3.	ІКр<58 (Група низької результативності)	74	16	9 225 628	53,14
	Усього	462	100	17 358 191	100

Власне вищенаведені результати характеризують реальний стан, і на основі отриманих результатів можна формувати проміжні та стратегічні цілі, що стосуються результативності діяльності підрозділів податкового відомства. У табл. 1.9 показано приклад сформованих стратегічних цілей податкового відомства держави на період до 2020 року, які можна уточнювати і корегувати на основі середніх прогнозних значень ІКр та контрольних параметрів СКВ («сігма»).

На рис. 1.16 показані прогнозні графіки розподілу значень ІКр для кожного контрольного року і межі, визначені значенням «сігма» (СКВ,(σ)), у яких повинні знаходитися ≥ 70 % органів ДПС України. Оскільки метод базується на встановленні двох цільових показників – середнього значення та СКВ (σ), його можна назвати методом «сігма-управління».

Таблиця 1.9

Приклад сформованих стратегічних цілей податкового відомства на основі методу «сігма – управління»

Показник \ роки	Фактичні значення	Цільові показники	
	2010	2015	2020
Середнє значення інтегрального критерію результативності (ІКр) органів ДПС України	65	80	90
Середньо квадратичне відхилення (СКВ, σ)	7	5	3

Щодо сучасних методів та технологій прийняття рішень, особливу увагу на сьогодні доцільно приділяти технологіям інтелектуального аналізу великих масивів даних (BigData) на основі методів машинного навчання (Machine Learning).

Технології інтелектуального аналізу даних об'єднують нові ідеї із галузі комп'ютерних наук, штучного інтелекту, статистики та теорії баз даних, і позначаються як *KDD* або *DM (Knowledge Discovery in Database, Data Mining* – добування *нових* знань із баз даних (датамайнінг).

На рис.1.16: N – кількість органів податкового відомства; Ікр – значення інтегрального критерію результативності.

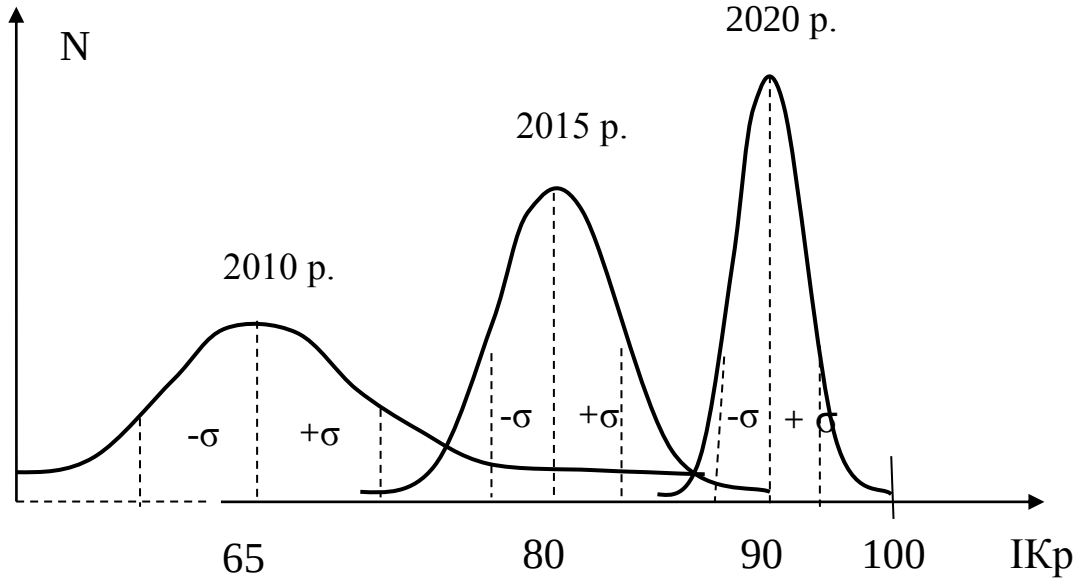


Рис. 1.16. Приклад «сігма–управління» за встановленими цілями з підвищення результативності податкового відомства держави

На сьогодні концепції управління, що базуються на впровадженні технологій оцінювання результатів діяльності та мотивації персоналу, інтегровані у парадигмі розвитку організації, орієнтовані на «управління знаннями», яку можна взяти за основу для побудови інтелектуальної інфраструктури ДПС України. Вона передбачає створення експертного середовища для системного і безперервного обміну, накопичення і використання професійних знань. Зокрема, однією із його складових може бути підсистема формування, моніторингу та контролю цільових показників.

Апробацію економіко-математичних моделей податкових процесів можна здійснювати за допомогою широкого спектру математичних та статистичних пакетів, що реалізують методи *датамайнінгу*. Зокрема, мови програмування Python, R, пакети *Statistica (StatSoft), SPSS,* додатки розширенрої аналітики СКБД Oracle 10g-18 мають широкий спектр методів виявлення і розпізнавання образів, асоціативних правил, дерев рішень, класифікації, кластерного і дискримінантного аналізу, нечітких множин, регресійно-кореляційного аналізу, прогнозування часових рядів (ARIMA та інші). Ці засоби можуть використовувати для підтримки прийняття рішень в управлінні різними аспектами діяльності органів ДПС України.

Отже, враховуючи неповноту інформації та неточність даних інформаційних ресурсів податкового відомства щодо діяльності платників податків, для побудови ефективної системи управління діяльністю органів ДПС України необхідне об’єднання зусиль експертів у галузі оподаткування, сучасних технологій аналізу інформації, управління знаннями. Комплексна система управління податковим відомством повинна включати такі елементи: нормативно-методологічне забезпечення операційної діяльності, ефективну організаційну та інформаційну інфраструктуру; системи: підтримки прийняття управлінських рішень, управління ризиками, контролю результативності діяльності, управління якістю.

1.3.3. Оцінювання податкової політики на основі методології Лаффера

Проблемі визначення критеріїв ефективності й оптимальності податкової системи у науковій літературі приділяють багато уваги С. М. Мовшович, Е. В. Балацький, В. В. Капітоненко, В. Є. Юринець та С. Л. Лондар, С. В. Онишко, Д. Г. Чернік, А. В. Скрипник. Оптимізаційна модель податкової політики, завдяки якій можна визначитися з податковим тиском на економіку, була розроблена, зокрема, А. Лаффером. Суть його теорії полягає у тому, що:

1) вхідним параметром моделі є новостворена суб'єктами господарювання додана вартість;

2) нелінійний зворотний зв'язок виходу оператора податкової системи із входом при перевищенні сукупної податкової ставки певного рівня, в результаті чого бюджетні надходження зменшуються;

3) на виході моделі формується сума мобілізованих бюджетних коштів.

Цій моделі приділили увагу науковці Є. В. Балацький, С. Л. Лондар, С. М. Мовшович.

Розширена модель податкової системи у вигляді математичного оператора, на вхід якого поступає валовий внутрішній (регіональний) продукт (ВВП, ВРП), що створюються суспільством за певний проміжок часу, а виходом оператора є бюджетні надходження, показані на рис. 1.17.

Дослідження у цій галузі проводяться за двома напрямами. Перший (теоретичний) – передбачає моделювання виробничих і фіскальних процесів та теоретичне виведення із побудованих моделей параболічної залежності Лаффера. У таких дослідженнях увага зосереджена на доказах наявності «ефекту перегину» на фіскальній кривій і умов його виникнення. Другий (прикладний) напрям пов'язаний з практичним розрахунком так званих точок Лаффера (точок перегину на фіскальній кривій) для окремих країн, регіонів, що за своєю суттю належить до теорії макроекономічного оцінювання.

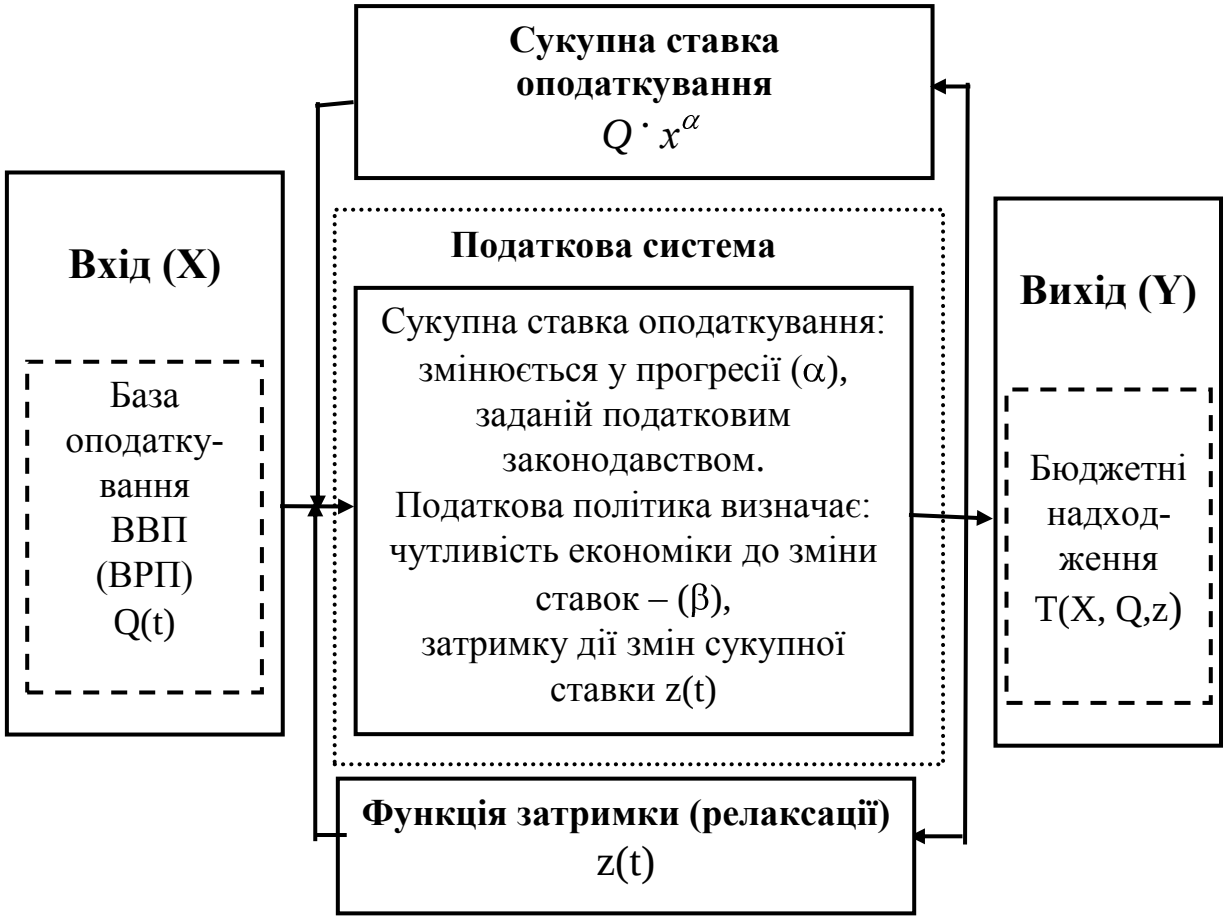


Рис. 1.17. Модель податкової системи як математичного оператора

У моделі, наведеній на рис. 1.17, податкова система як математичний оператор діє на ВВП/ВРП за допомогою сукупної податкової ставки у відповідній прогресії, що визначається законодавством. Вчений С. Л. Лондар пропонує такий вигляд математичного оператора податкової системи (1.6).

$$T(x)= Q\, x^{\alpha}(1-x)^{\beta}, \qquad (1.6)$$

де Q – валовий внутрішній продукт (валовий регіональний продукт), (ВВП, ВРП);

x – сукупна податкова ставка (у частках одиниці);

α – коефіцієнт податкової прогресії;

β – коефіцієнт чутливості економіки до зміни податкової ставки (коефіцієнт згасання економічної активності);

x^{α} – прогресія сукупної ставки оподаткування;

Q – (ВВП, ВРП), у свою чергу, залежать від кількості діючих суб'єктів господарювання $n = (1-x)^{\beta}$.

Для врахування перехідного процесу під час введення змін у систему оподаткування пропонують ввести в оператор функцію затримки z(t):

$z(t)= e^{\pm t/\tau}$ – функція затримки (релаксації) дії зміни податкової ставки на економіку, де τ – час, за який величина податкових надходжень зменшиться/збільшиться в e $(e=2,71828)$ разів від загальної величини релаксації. На статистичних даних показано, що середня тривалість періоду, потягом якого відбувається макроекономічна реакція $\tau = 2,5 \pm 0,8$ місяця. У загальному випадку залежність податкових надходжень від моменту дії зміни ставки x можна записати так:

$$F(t)= A\cdot e^{+t/\tau}+T(x_2), \qquad (1.7)$$

де $A=T(x_1)-T(x_2)\pm\Delta\gamma$ – значення миттєвого збільшення/зменшення надходжень залежно від рівня фіскальності податкової політики;

$T(x_1)$, $T(x_2)$ – надходження до і після перехідного періоду відповідно;

x_1, x_2 – діюча і нова сукупні ставки оподаткування;

$\Delta\gamma$ – зміна величини ВВП/ВРП під час введення нової сукупної ставки оподаткування.

Лінійна залежність кількості $n(x) = N(1-x)$ діючих підприємств від податкової ставки x (крива 1) і відповідна їй крива надходжень у бюджет $F(x)$ (крива 4) та залежності $n(x)= (1-x)^{\beta}$ при $\beta<1$ і $\beta>1$ (криві 2 та 3 відповідно) показані на рис. 1.18 (N – кількість зареєстрованих підприємств).

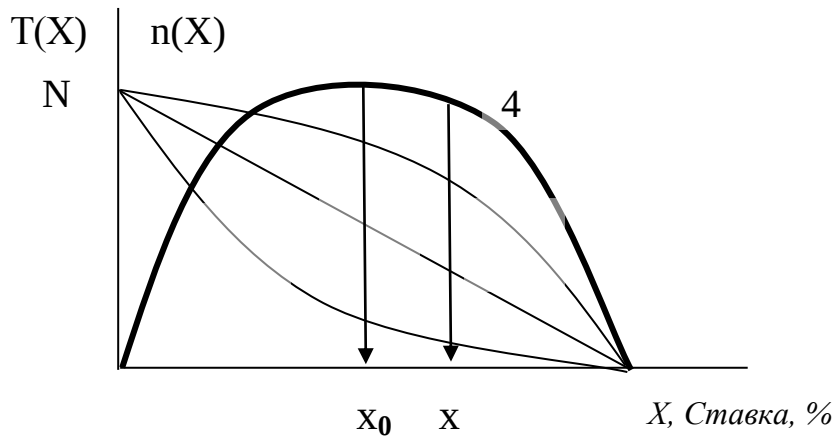


Рис. 1.18. Графіки залежності кількості підприємств *n(x)* і надходжень у бюджет *T(x)* від податкової ставки *x*

При позитивній динаміці національної економіки і, відповідно, при збільшенні сукупної ставки *x* зменшення діючих підприємств незначне (крива 2, β<1).

При економічному спаданні, навіть при незначному збільшенні сукупної податкової ставки x веде до різкого скорочення діючих підприємств (крива 3, β>1). Дослідженнями доведено, що оптимальне значення сукупної ставки x_0 визначається винятково коефіцієнтами α і β:

$$x_0 = \alpha /(\alpha+\beta). \qquad (1.8)$$

Як правило, критерій оптимальності податкової системи на основі цього підходу формулюється так: податкова система є оптимальною, якщо сукупна податкова ставка *x,* якою характеризуються взаємовідносини податкової системи із економічним середовищем, рівна податковій ставці x_0.

У протилежному випадку ці взаємовідносини характеризуються слабким ($x < x_0$) або надмірним ($x > x_0$) податковим тиском на економічне середовище, тобто необхідно розробити комплекс правових, економічних і фінансових заходів, що дозволять мінімізувати різницю $|x-x_0|$.

Проведені вітчизняними економістами дослідження вказують на складність визначення критеріїв оптимальності чи ефективності податкової системи, які на практичному рівні можна було б використовувати у формуванні регіональної податкової політики. Проблемою є трактування поняття «сукупна податкова ставка», її визначення та обчислення показників, за якими можна порівнювати ефективність здійснення податкової політики регіональними державними установами. Порівняльне оцінювання ефективності здійснення регіональної податкової політики доцільно виконувати на основі річних значень податкового зобов'язання і сукупних ставок оподаткування валового регіонального продукту, створеної у регіоні у розрахунку на душу населення та одного реального платника податків. Це дозволить не враховувати відмінності регіонів за кількістю населення, розмірами ВДВ, кількістю платників і визначати такі, де податкова політика була менш ефективною, ніж у інших.

В основі досліджень Є. В. Балацького лежить дослідження залежності обсягу регіонального виробництва Q_i (для узагальнення визначається величиною валового регіонального продукту ($ВРП_i$)) від рівня податкового тиску у регіоні $x_i = T_i/Q_i$, де T_i – сума регіональних податкових надходжень до бюджету, Q_i – фактично валовий регіональний продукт ($ВРП_i$) . Залежність ВРП від сукупної ставки оподаткування x можна апроксимувати нелінійною функцією $Q(x)$, параметри якої необхідно оцінити. При цьому, на відміну від попереднього підходу, розрізняють точки Лаффера першого та другого родів.

Точка Лаффера *першого роду* – це точка x^*, при якій виробнича крива $Q = Q(x)$ досягає локального максимуму, тобто коли виконуються умови: $dQ(x^*)/dx = 0$; $d^2Q(x^*)/dx^2 < 0$. Точка Лаффера *другого роду* – це точка x^{**}, при якій фіскальна крива $T = T(x)$ досягає локального максимуму, тобто коли виконуються умови $dT(x^{**})/dx = 0$; $d^2Q(x^{**})/dx^2 < 0$.

Економічно точка Лаффера першого роду x^* означає межу податкового тиску, при якому економічна система не переходить у режим рецесії. Точка Лаффера другого роду x^{**} показує величину податкового тиску, за межами якої збільшення податкових надходжень стає не можливим. Оцінювання двох точок Лаффера і їх зіставлення з фактичним податковим тиском дозволяє оцінити ефективність податкової системи держави і напрями її оптимізації (рис. 1.19).

Дослідження Є. В. Балацького показують, що нелінійну апроксимацію вказаних залежностей досить апроксимувати поліноміальною регресією другого або третього степеня і, відповідно, на основі *трипараметричного* або *двопараметричного* методу. У цьому посібнику розглядається двопараметричний метод.

В основі двопараметричного методу лежить апроксимація процесу економічного зростання укороченою квадратичною функцією, що включає тільки два параметри (1.9):

$$Q = \beta x + \gamma x^2, \quad T = \beta x^2 + \gamma x^3. \qquad (1.9)$$

Додаткові обмеження, що накладаються на функціональні властивості виробничої системи, задаються рівнянням (1.10):

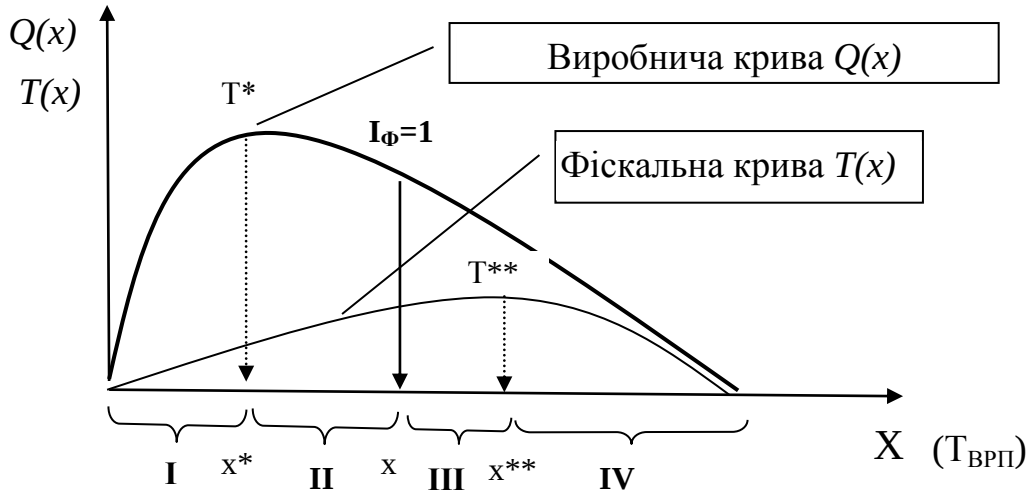
$$\Delta Q=(\beta+2\gamma x)\,\Delta x. \qquad (1.10)$$

Система рівнянь (1.9, 1.10) достатня для визначення параметрів β, γ.

Точки Лаффера першого та другого роду виводяться із (1.9) і розраховуються за формулами (1.11):

$$x^* = -\beta/2\gamma,\ x^{**} = -2\beta/3\gamma. \qquad (1.11)$$

Необхідно виділити особливість цього підходу до застосування концепції Лаффера. Сучасна теорія податків оперує в основному тільки з точками Лаффера другого роду, залишаючи без уваги точки першого роду. Можна розглядати три основних положення значення фактичної сукупної ставки оподаткування відносно до x* і x**: $x^* < x < x^{**}$; $x < x^* < x^{**}$; $x > x^* > x^{**}$ (рис. 1.19).



I – Зона економічного ризику;
II – Зона стимулювання бізнесу;
III – Зона ризику згортання бізнесу;
IV – Зона економічного ризику.

Рис. 1.19. Зони економічного ризику держави і бізнесу на виробничій та фіскальній кривих

При $x^* < x < x^{**}$ з погляду інтересів держави податковий тиск має допустиму величину. Бажаним випадком для господарюючих суб'єктів є ситуація $x < x^*$ і навпаки, для усіх суб'єктів економіки несприйнятним є варіант $x > x^{**}$, тобто варіант надзвичайного податкового тиску.

На основі двопараметричної моделі (1.9), яку можна вважати цільовою функцією, можна дійти висновку про те, що фіскальна політика буде стимулюючою для виробництва у випадку, коли ставка оподаткування валового регіонального продукту ($Т_{ВРПі}$) буде наближатися до точки Лаффера першого роду $Т_{ВРПі} \rightarrow x^*$.

При $Т_{ВРПі} < x^*$ економіка *і*-го регіону входить у зону економічного ризику держави (регіону, групи регіонів) у зв'язку із значним зменшення податкових надходжень через зменшення ставок оподаткування, а при $Т_{ВРПі} > x^{**}$ збільшуються ризики для господарюючих об'єктів через надмірний податковий тиск. Тобто для розвитку економіки бажано, щоб точка $Твдв_і$ позиціонувалася у межах $x^* < Т_{ВРПі} < x^{**}$. Для *j – ї* групи регіонів рівень фіскального тиску у *і*-му регіоні можна оцінювати на основі розрахунку *регіонального індексу фіскальності.* Рівень регіонального податкового тиску будемо оцінювати як відношення фактичної регіональної сукупної ставки оподаткування ВРП ($Т_{ВРПі}$) до значення $(x_j^* + x_j^{**})/2$ – середини інтервалу між точками Лаффера, визначеними для j –ї групи регіонів (точка *x*, що розділяє зони II, III, (1.12).

$$I_{фі} = 2Т_{ВРПі} / (x_j^* + x_j^{**}). \qquad (1.12)$$

Із формул (1.9) можна визначити критичні значення індексу фіскальності для критичних точок *j-ї* групи регіонів:

$I_{фП} = 2Т_{ВРПj} / (x_j^* + x_j^{**}) = 1$, *помірне значення індексу фіскальності для j –ї групи регіонів;*

$I_{фj}^* = 2x_j^* / (x_j^* + x_j^{**}) = 0{,}86/\gamma$, *для точки Лаффера першого роду* x^*;

$I_{фj}^{**} = 2x_j^{**} / (x_j^* + x_j^{**}) = 1{,}14/\gamma$, *для точки Лаффера другого роду* x^{**};

Враховуючи попередні судження, можна виділити такі зони ризику, що визначаються фактичним значеннями регіонального ідексу фіскальності $I_{фі}$ (рис. 1.19):

I: при $I_{фі} < (I_{фj}^{*} = 0,86/\gamma)$ – зона економічного ризику регіону, пов'язаного із надмірно малими значеннями ставок оподаткування і, відповідно, низьким податковими зобов'язаннями, що не дасть змогу виконувати свої функції державним установам;

II: $(I_{фj}^{*} = 0,86/\gamma) < I_{фі} < 1$ – зона помірних податкових зобов'язань, зона стимулювання бізнесу і ефективного розвитку економіки у регіоні;

III: $1 < I_{фі} < (I_{фj}^{**} = 1,14/\gamma)$ – зона високих податкових зобов'язань і економічного ризику бізнесу, пов'язаного із скороченням виробництва через високі ставки оподаткування;

IV: $I_{фі} > (I_{фj}^{**} = 1,14/\gamma)$ – зона надмірних податкових зобов'язань та економічного ризику, пов'язаного із рецесією виробництва та переходом бізнесу регіону у тінь.

Власне допустимими (рекомендованими) значеннями індексу фіскальності податкової політики у регіоні можна вважати діапазон $0,86/\gamma < I_{фі} < 0,14/\gamma$.

З економічного погляду точка x_j^{*} показує межу податкового тиску, при якому виробництво не переходить у режим рецесії. Точка x_j^{**} відображає величину податкового тиску, за межами якої збільшення податкових надходжень стає неможливим. При досягненні скупної ставки оподаткування розміру x_j^{**} і при її перевищенні надмірні податки приводять до того, що чистого прибутку у господарюючих суб'ктів фактично не залишається. Починається спад економічної активності, значна частина доходу переходить у тіньовий сектор.

1.3.4. Аналітичні показники для оцінювання податкового зобов'язання регіону

Податкове навантаження на економіку держави, як правило, оцінюється відношенням фактичних податкових надходжень до бюджету і ВВП. Під час аналізу ефективності оподаткування можна розрізняти показники *фактичного* податкового навантаження на економіку, *номінальног*о податкового навантаження і *потенційного* податкового навантаження на економіку, які повинні враховувати, відповідно, фактичні і потенційні податкові надходження, зокрема пільги, недоїмку, тіньову економіку.

Нижче наведена система показників, які можуть розраховуватися на основі даних про фактичні надходження за забов'язаннями звітного періоду, недоїмку (борг) та надані пільги.

Потенційні податкові надходження регіону ($ППН_Р$) – показник, що у грошовому виразі визначає можливу дохідність регіону як суму фактичних податкових надходжень за зобов'язаннями звітного періоду, недоїмки та наданих законодавством пільг у звітному періоді:

$$ППН_Р = ФПН_Р + Н_Р + ПП_Р, \qquad (1.13)$$

де $ППН_Р$ – потенційні податкові надходження регіону за звітний період;

$ФПН_Р$ – фактичні податкові надходження регіону за зобов'язаннями звітного періоду;

$Н_Р$ – податкова недоїмка у регіоні, створена у звітному періоді;

$ПП_Р$ – пільги, отримані платниками регіону у звітному періоді.

Податкове зобов'язання регіону ($ПЗ_Р$) – показник, що характеризує податкову дохідність регіону як суму фактичних податкових надходжень за зобов'язаннями звітного періоду та недоїмки, створеної у відповідному звітному періоді:

$$ПЗ_Р = ФПН_Р + Н_Р. \qquad (1.14)$$

Наближену величину податкового зобов'язання регіону за зобов'язаннями звітного періоду можна розрахувати як суму

фактичних податкових надходжень і приросту недоїмки за звітний період:

$$ПЗ^t \approx ФПН_t^t + \Delta H_t^t, \qquad (1.15)$$

де верхній індекс у змінних – період виникнення зобов'язань;

нижній індекс – період сплати податків або зміни недоїмки;

$ПЗ^t$ – податкові зобов'язання за зобов'язаннями звітного періоду;

$ФПН_t^t$ – фактичні податкові надходження за зобов'язаннями звітного періоду,

ΔH_t^t – приріст недоїмки за зобов'язаннями звітного періоду.

Фактичні податкові надходження за зобов'язаннями звітного місяця можна визначити як різницю між сумарними фактичними надходженнями у звітному місяці і податковими надходженнями зі сплати зобов'язань попередніх періодів:

$$ФПН_t^t = ФПН_t - \sum_{\tau<t} ФПН_t^\tau, \qquad (1.16)$$

де $ФПН_t^t$ – фактичні податкові надходження за зобов'язаннями звітного місяця;

$ФПН_t$ – сумарні фактичні надходженнями у звітному місяці;

$ФПН_t^\tau$ – податкові надходження зі сплати зобов'язань попереднього періоду τ.

Показник $\sum_{\tau<t} ФПН_t^\tau$ у формулі (1.16) визначає податки, сплачені не в термін, тобто зобов'язання, які з'явилися у попередні періоди і не були погашені. Відповідно, у ті ж періоди формувалася недоїмка, тому погашення цих зобов'язань у поточному періоді є скороченням недоїмки за попередніми зобов'язаннями (для спрощення, як пропонує у своїх працях Є. В. Балацький), будемо вважати, що ні пеня, ні штрафи у недоїмку не входять (1.17):

$$ФПН_t^\tau = -\Delta H_t^\tau \, (\tau \prec \mathrm{t}). \qquad (1.17)$$

Повний приріст недоїмки у поточному місяці дорівнює сумі приросту недоїмки за зобов'язаннями поточного місяця і зміни (скорочення) недоїмки за рахунок сплати за зобов'язаннями попередніх періодів:

$$\sum_{0<\tau<t} \Delta ПН_t^\tau + \Delta H_t^t = \Delta H_t \qquad (1.18)$$

Підставляючи (1.12) у формулу визначення податкового зобов'язання (1.17), отримаємо (1.18):

$$\text{ПЗ}^t = ФПН_t - \sum_{\tau<t} \Delta ФПН_t^\tau + \Delta H_t^\tau = ФПН_t + \\ + \sum_{\tau<t} \Delta ПН_t^\tau + \Delta H_t^\tau = ФПН_t + \Delta H_t . \quad (1.19)$$

Таким чином, для наближеного оцінювання базових агрегованих показників регіонального податкового зобов'язання можна використовувати формулу (1.19).

Враховуючи необхідність порівняльного аналізу регіональної політики та ефективності функціонування податкової системи, доцільним на сьогодні є використання даних валового регіонального продукту (ВРП), виробленого у регіоні, що розраховується Державним комітетом статистики, і який є базою оподаткування, найповніше в агрегованому вигляді характеризує рівень регіональної економічної активності.

Частиною ВРП є заробітна плата, з якої нараховується прибутковий податок і на яку нараховуються збори на обов'язкове пенсійне забезпечення. Ці нарахування включаються виробниками до собівартості продукції, перекладаються у цінах товарів на споживача і оплачуються ним із власних доходів. Іншою частиною є прибуток, що оподатковується податком на прибуток підприємств, та ПДВ, що стягується із загальної величини створеної доданої вартості.

Для оцінювання ефективності здійснення регіональної податкової політики доцільно ввести такі індикатори (індекси):

1) Індекс фіскальної ефективності ($I_{фе}$) – показник рівня фактичної сплати податків, що характеризує, зокрема, ефективність фіскальних зусиль і розраховується як відношення фактичних податкових надходжень регіону за зобов'язаннями звітного періоду ($ФПН_Р$) до розрахованих податкових зобов'язань регіону ($ПЗ_Р$) у звітному періоді:

$$I_{фе} = ФПН_Р / ПЗ_Р, \quad (1.20)$$

де $I_{фе}$ – індекс фіскальної ефективності;

$ФПН_Р$ – фактичні податкові надходження у регіоні за зобов'язаннями звітного періоду;

$ПЗ_Р$ – податкові зобов'язання регіону.

Таблиця 1.10

Ефективна ставка оподаткування ВРП у 1998–2013 роках*

Показник	Позначення	Роки				
		1998	1999	2000	2009	2013
Податкове зобов'язання (млн грн)	ПЗ	30 962	28 671	30272	344 456	314 160
Валовий регіональний продукт (млн грн)	ВРП	82 834	103 847	142 669	913 345	1 522657
Ефективна ставка оподаткування ВРП, %	$Т_{ВРП}$	37,4	27,6	21,2	36,3	21,5

***Розраховано на основі оперативних даних ДПС/ДФС України**

Коефіцієнт виконання планових зобов'язань ($К_{ПЗ}$) характеризує якість прогнозу надходжень і розраховується як відношення фактичних податкових надходжень регіону/держави у звітному періоді до планових податкових надходжень (до зведеного та державного бюджетів).

$$К_{ВП} = ФПН_{ЗП} / ПН_{П}, \qquad (1.21)$$

де $К_{ВП}$ – коефіцієнт виконання планових зобов'язань;

$ФПН_{ЗП}$ – фактичні податкові надходження у звітному періоді;

$ПН_{П}$ – планові податкові надходження.

Індекс потенційних податкових надходжень регіону ($І_{ППН}$) – характеризує рівень податкового зобов'язання регіону і ($ПЗ_{Р}$) щодо потенційних податкових надходжень регіону ($ППН_{Р}$):

$$І_{ППН} = ПЗ_{Р} / ППН_{Р}, \qquad (1.22)$$

де $І_{ППН}$ – індекс фактичних податкових надходжень;

$ПЗ_{Р}$ – податкове зобов'язання регіону;

$ППН_{Р}$ – потенційні податкові надходження регіону.

Коефіцієнт резерву збільшення податкових надходжень ($К_{РПН}$) є зворотним до індексу потенційних податкових надходжень і вказує, на яку відносну величину можуть збільшитися податкові надходження у регіоні при повній сплаті податків, якщо недоїмка (борг) і пільги будуть рівними нулю:

$$К_{РПН} = ППН_Р / ПЗ_Р. \qquad (1.23)$$

Для оцінювання ефективності здійснення податкової політики доцільно також аналізувати такі показники, як витрати на адміністрування податків, динаміку зростання кількості діючих підприємств, динаміку зростання ВРП на душу населення, показники виконання Зведеного бюджету. Такі показники розглядалися у параграфі 1.3.2.

1.3.5. Моделі цільових функцій регіональної податкової політики

Побудуємо цільову функцію регіональної податкової політики на основі таких показників:

B(t) – затрати на забезпечення функціонування державних установ, що здійснюють податкову політику (адміністрування податків);

T(t) – фактичні податкові надходження за зобов'язаннями звітного періоду;

N (t) – податкова недоїмка звітного періоду;

$S_д(t)$ – кількість фактично діючих суб'єктів господарської діяльності (СГД), які сплачують податки;

$S_з(t)$ – кількість зареєстрованих СПД;

*T**, *T***– точки Лаффера першого та другого родів для групи регіонів, до якої входить регіон (відповідно x*, x** див. 1.3.3);

$T_{ВДВij}$ – ставка оподаткування ВРП у регіоні;

За умови рівнозначності показників) цільова функція регіональної податкової політики може бути подана так:

$$\max F(t) = \left[(1 - \frac{B(t)}{T(t)})^{0,25} \cdot (1 - \frac{N(t)}{T(t)})^{0,25} \cdot (\frac{S_Д(t)}{S_З(t)})^{0,25} \cdot (\frac{T_j^*}{T_{ВДВ}})^{0,25} \right] \quad (1.24)$$

Обмеженнями моделі є :

$0 < F(t) < 1$, $B(t) < T(t)$, $N(t) < T(t)$, $S_д(t) < S_З(t)$, $T_{ВДВ} \geq T^*$

Таким чином, критерій ефективності регіональної податкової політики можна записати так:

$$P_{t1,t2} = \int_{t1}^{t2} \left[(1 - \frac{B(t)}{T(t)}) \cdot (1 - \frac{N(t)}{T(t)}) \cdot \frac{S_Д(t)}{S_З(t)} \cdot (\frac{T^* \cdot Q_{РП}(t)}{T_{РП}(t)}) \right]^{0,25} dt \quad (1.25)$$

Обмеження:

$\frac{dQ_{РП}}{dt} > 0$ – умова зростання ВРП у розрахунку на одного діючого (реального) платника;

$\frac{dT_{РП}}{dt} > 0$ – умова зростання податкового зобов'язання у розрахунку на одного реального платника;

$\frac{dT_{РП}^2}{dQ_{РП}^2} > 0$ – умова еластичності $ПЗ_{дн}$ до $ВРП_{дн}$;

$\frac{dS_Д}{dt} > 0,$ – умова зростання кількості реальних платників податків.

Зростання критерію (1.25) є позитивним під час виконання таких умов: збільшення ВРП і податкового зобов'язання у розрахунку на одного реального платника податків у регіоні мають позитивну динаміку; еластичність податкового зобов'язання до валової доданої вартості має прогресивний характер; питома вага кількості діючих суб'єктів підприємництва, які сплачують податки у кількості зареєстрованих, збільшується.

Під час зростання критерію ефективності $P_{t1,t2}$, повинна виконуватися умова дотримання значення індексу фіскальності $I_ф\,(t)$ в зоні II (рис. 1.19), або в межах:

$$2T^* / (T^*+T^{**}) \le I_ф\,(t) \le 1 \qquad (1.26)$$

Використовуючи дану модель, необхідно враховувати нерівномірність розвитку регіонів України, відмінності умов функціонування галузей. Оцінювання точок Лаффера дасть змогу визначити ефективність регіональної податкової політики у минулі періоди і, за наявності прогнозних значень ВРП, податкового зобов'язання та динаміки зростання кількості реальних платників податків, визначити податковий тиск і зони ризику економіки регіону у прогнозованих періодах.

Розглянемо приклад побудови інтегрального критерію оцінювання ефективності функціонування податкової системи на основі таких показників:

1. Коефіцієнт ефективності адміністрування податків, що характеризує рівень справляння податків і який можна записати так:

$$К_{еа} = 1 – Н/ФПН, \qquad (1.27)$$

де $К_{еа}$ – коефіцієнт ефективності адміністрування податків,

$Н$ – наявна недоїмка на кінець звітного періоду,

$ФПН$ – фактичні податкові надходження у звітному періоді.

Позитивними змінами у функціонуванні податкової служби повинно бути наближення цього показника до одиниці.

2. Коефіцієнт результативності перевірочної роботи, що характеризує рівень надходжень за результатами перевірок:

$$К_{рп} = С_{пл} / Н_{ар}, \qquad (1.28)$$

де $К_{рп}$ – коефіцієнт результативності перевірок;

$С_{пл}$ – сплачено до бюджету за результатами перевірок;

$Н_{ар}$ – нараховано за результатами перевірок.

Позитивною тенденцією буде також наближення цього показника до одиниці.

3. Коефіцієнт *виконання встановлених прогнозних завдань з податкових надходжень до бюджету*, що характеризує рівень виконання встановлених прогнозних показників, оцінює якість прогнозування:

$$К_{вп} = ФПН / ПН_{дБ}, \qquad (1.29)$$

де $К_{вп}$ – коефіцієнт виконання завдань з податкових надходжень до бюджету;

ФПН – фактичні податкові надходження;

$ПН_{дб}$ – податкові надходження, враховані в розрахунках держбюджету.

4. Коефіцієнт ефективності функціонування податкового відомства:

$$К_{едо} = 1 – В_{ф} / ФПН, \qquad (1.30)$$

де $К_{едо}$ – коефіцієнт ефективності функціонування податкового відомства;

$В_{ф}$ – фактичні витрати на податкове відомство;

ФПН – фактичні податкові надходження у звітному періоді.

5. Індекс динаміки зростання діючих підприємств:

$$D_{рп} = S_{д} / S_{з},$$

де $D_{рп}$ – індекс динаміки зростання діючих підприємств;

$S_{д}$ – діючі підприємства,

$S_{з}$ – зареєстровані підприємства.

6. Індекс динаміки зміни ВРП на душу населення:

$$D_{вдв} = 1 – ВРП_{днт1} / ВРП_{днт2}, \qquad (1.31)$$

де $D_{вдв}$ – динаміка зміни ВРП на душу населення,

$ВРП_{нт1}$ і $ВРП_{днт2}$ – ВРП на душу населення попереднього та звітного періодів відповідно.

Частина визначених показників характеризує ефективність здійснення податкової політики (індекс потенційних податкових надходжень регіону ($I_{ППН}$), коефіцієнт ефективності функціонування податкового відомства ($К_{едо}$), індекс динаміки зростання діючих підприємств ($D_{дп}$), індекс динаміки ВРП на душу населення ($D_{ВРП}$)), інша – характеризує ефективність функціонування

підрозділів податкового відомства (індекс фіскальної ефективності ($I_{фе}$), коефіцієнт ефективності адміністрування податків ($К_{еа}$), коефіцієнт результативності перевірочної роботи ($К_{пр}$), коефіцієнт виконання бюджету з податкових надходжень ($К_{вп}$).

Подамо цільову функцію ефективності здійснення податкової політики на основі запропонованих показників в адитивному форматі:

$$\max F_{ППІ}(t) = 0,25 \cdot I_{ППІН}(t) + K_{едо}(t) + D_{оп}(t) + D_{ВРП}(t), \quad (1.32)$$

$0 < F_{ПП}(t) < 1.$

Цільову функцію діяльності податкового відомства можна записати так:

$$\max F_{ПС}(t) = 0,25 \cdot \left| I_{фе}(t) + K_{еа}(t) + K_{пр}(t) + K_{еп}(t) \right|, \quad (1.33)$$

$0 < F_{ПС}(t) < 1.$

Інтегральна цільова функція податкової системи на основі двох попередніх функцій за умови рівнозначності складових записується так:

$$\max F(t) = 0,5 \cdot \left[F_{ПП}(t) + F_{ПС}(t) \right] \quad (1.34)$$

$0 < F(t) < 1.$

Відповідно, комбінований критерій ефективності податкової системи у звітному періоді можна записати так:

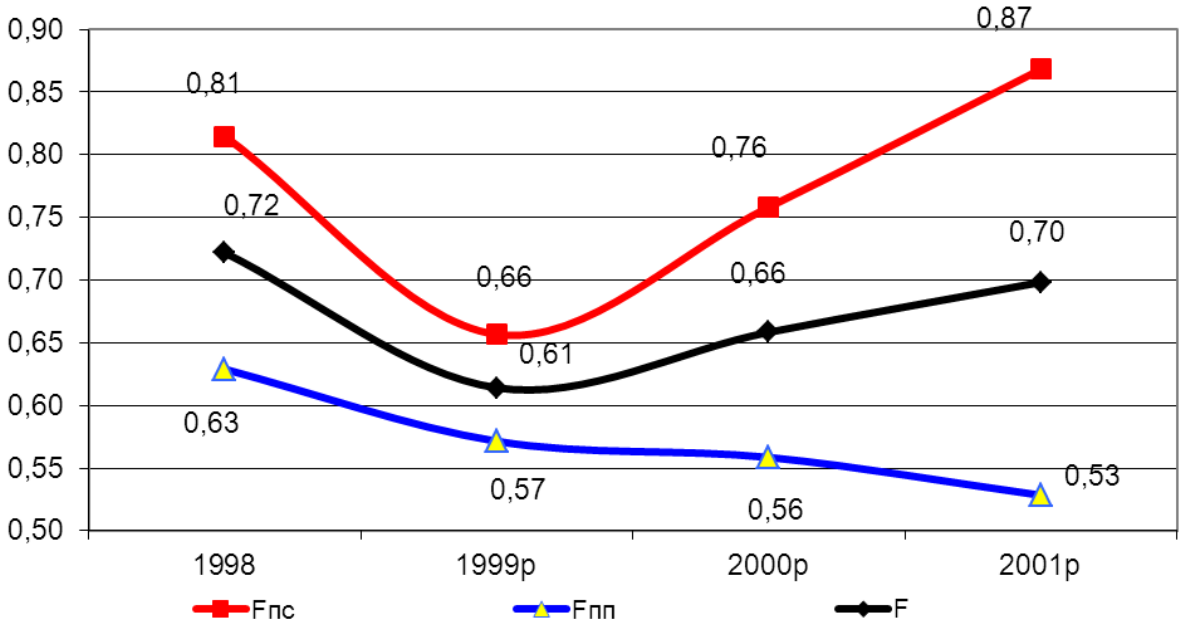
$$\max F_{t1,t2} = 0,5 \cdot \int_{t1}^{t2} \left[F_{ПП}(t) + F_{ПС}(t) \right] dt, \quad (1.35)$$

де t_1 *і* t_2 – початок і кінець звітного періоду.

На рис. 1.20 для прикладу у вигляді окремих графіків наведені цільові функції податкової політики, податкової служби, податкової системи у період економічної кризи 1998–2001 років.

Графіки дають можливість робити висновки щодо ефективності функціонування податкової системи, оцінювати динаміку її складових. Аналіз показує, що податкова політика поступово погіршувалася від – 0,63 у 1998 році до 0,53 у 2001 році. Погіршення

значення критерію ефективності здійснення податкової політики відбувалося за рахунок збільшення розмірів наданих пільг, що відобразилося на індексі потенційних податкових надходжень ($I_{ппн}$), коефіцієнті ефективності функціонування податкового відомства за рахунок значного збільшення витрат на адміністрування податків ($K_{ефо}$), недостатньої динаміки зростання ВРП у розрахунку на душу населення ($D_{вдв}$).



$F_{пс}$ – значення критерію ефективності функціонування податкової служби;
$F_{пп}$ – значення критерію ефективності здійснення податкової політики;
F – значення критерію ефективності функціонування податкової системи.

Рис. 1.20. Приклад аналізу податкової політики, функціонування податкової служби та податкової системи України на основі критеріїв ефективності (1998–2001 рр.)

Проаналізувавши вищезазначене, можна дійти висновку про те, що ефективність здійснення державою податкової політики, відповідно до вибраних складових інтегрованого показника ефективності, становила у 2001 році 53 % від потенційно можливої, водночас ефективність діяльності податкового відомства – 87 %, а ефективність функціонування податкової системи України загалом – 70 %.

1.3.6. Методологічні основи управління ризиками в адмініструванні податків

Податкове відомство держави протягом останніх років впроваджує *в адмініструванні податків ризикоорієнтований підхід* та побудову на його основі *ризикоорієнтованої системи адміністрування податків* (далі за текстом РОСАП). У 2017 ДФС України оновлено ризикоорієнтовану систему податкового контролю та ІТ-відбір платників податків для планових податкових перевірок. Тобто основна увага в оцінці ризиків приділяється саме платникам податків. Проте закони та реалії ринкової економіки обумовлюють впровадження в управлінні методів, орієнтованих на аналіз та оцінку не лише зовнішніх, а й внутрішніх факторів, що впливають на ефективність діяльності підприємств та відомств. Йдеться *про управління, що базується на принципах ризикоорієнтованого менеджменту*.

Міжнародний досвід виявлення ризиків та їх нейтралізації вказує на важливість розуміння операційного контексту як сукупного поєднання усіх операційних процесів, зовнішніх і внутрішніх факторів та обставин, у яких здійснюють діяльність органи ДПС України.

Впроваджуючи ризикоорієнтований підхід, важливим є врахування парадигми корпоративного управління, що впроваджується у податкових службах та у банківській сфері провідних європейських держав, а саме *операційного ризик-менеджменту (далі – ОРМ)*. Необхідність впровадження ОРМ усвідомлюється у міру розвитку організації, коли її діяльність ускладнюється різноплановими процесами, існує широка мережа регіональних підрозділів, багаторівневе операційне середовище. Внаслідок цього актуальним стає оптимізація бізнес-процесів і нерозривно пов’язана із цим задача створення системи виявлення, оцінювання та управління операційними ризиками, що і визначає суть ОРМ.

Необхідно зауважити, що наука про економічний ризик (ризикологія) покликана вивчати на засадах системного аналізу та синергетичного підходу принципи та методи оцінювання, моделювання та управління ризиком. Тобто під час впровадження

ризикоорієнтованого підходу йдеться, зокрема, про: *формування структури та класифікацію ризиків; визначення джерел, факторів та причинно-наслідкового змісту ризиків; формування системи показників ризиків, якісний і кількісний їхній аналіз та вимірювання; моделювання та прогнозування ризиків; управління ризиками та зниження їх ступеня*.

Актуальним завданням є формування методологічних основ управління ризиками в галузі адміністрування податків, систематизація понятійного апарату податкового ризик-менеджменту, ідентифікація *типів, класів, груп та видів* ризиків, що властиві діяльності податкового відомства.

Аналіз, систематизація та узагальнення підходів, представлених у науковій літературі, дають підстави виокремити основні структурні елементи категорії «ризик». А саме:

1. *Ризик* – потенційна, чисельно вимірна можливість втрат. Поняттям ризику характеризується невизначеність, пов'язана з можливістю виникнення у процесі діяльності організації або реалізації нею проекту несприятливих ситуацій і наслідків.

2. *Ризик* – ймовірність виникнення фінансових або матеріальних втрат, збитків, недонадходжень податків, планованих доходів, прибутку.

3. *Ризик* – це невизначеність фінансових результатів у майбутньому.

4. *Ризик* – це вартісне вираження імовірнісної події, що призводить до фінансових, матеріальних і нематеріальних або інших втрат.

5. *Ризик* – ймовірність втрати цінностей (фінансових, матеріальних товарних ресурсів) у результаті діяльності, якщо умови проведення діяльності будуть змінюватися у напрямі, відмінному від передбаченого планами і розрахунками.

У цих визначеннях помітний тісний зв'язок понять «ризик», «ймовірність» та «невизначеність». Щодо поняття *ймовірності,* розрізняють об'єктивну і суб'єктивну (експертну) ймовірність. Наприклад, ймовірність виникнення деякого рівня податкових втрат T через порушення податкового законодавства деяким платником А податків i-го виду діяльності j-ї групи за величиною

задекларованого доходу може бути подана у загальному вигляді формулою (1.36):

$$pT_{Aij} = f_{ij}\,(T), \qquad (1.36)$$

де: $f\,(T_{ij}) = n_{ij}\,(T)/N_{ij}$;

$f_{ij}(T)$ – частота виникнення податкових втрат T через порушення законодавства платниками i-го виду діяльності j-ї групи за величиною задекларованого доходу;

$n_{ij}(T)$ – число випадків порушення податкового законодавства;

N_{ij} – загальне число випадків податкових порушень платниками і-го виду діяльності j-ї групи за величиною задекларованого доходу.

Залежність результатів оцінювання ймовірностей податкових втрат від вхідної інформації, наприклад від недостовірних даних у звітності платників податків, призводить до появи *невизначеності*. *Невизначеність* припускає наявність факторів, за яких результати дій не є детермінованими, а ступінь можливого впливу цих факторів на результати невідомий; наприклад, це неповнота або неточність інформації.

Щодо *ризику у сфері оподаткування*, необхідно враховувати, що він зумовлений невизначеністю та конфліктністю інтересів, які виникають у процесі розподілу створеного продукту між платниками податків та державою. Платники – юридичні й фізичні особи, які безпосередньо створюють матеріальні та нематеріальні блага, – шукають напрями мінімізації податків, щоб віддати державі якомога менше. Держава розраховує на постійні й стабільні надходження в обсягах, принаймні, не менших ніж заплановано в бюджеті на поточний рік і постійно шукає шляхи мобілізації додаткових надходжень. Отже, оскільки суб’єкти, що беруть участь у процесі оподаткування, мають різні інтереси, цілі й сподівання щодо сплати податків, то поняття «податковий ризик» має неоднозначний зміст. До того ж існує суб’єктивне ставлення учасників процесів оподаткування до ризику (є особи схильні до ризиків, нейтральні та несхильні до ризиків), що повинно враховуватися при означенні цих понять. Отже, *податковий ризик* необхідно розглядати з різних позицій: з боку платників податків та держави.

З позиції платника податку *податковий ризик* – це імовірність фінансових та інших втрат через впровадження надмірно фіскального податкового законодавства, можливих додаткових нарахувань платежів, штрафів, пені тощо у результаті перевірок контролюючими органами, порушень законодавства посадовими особами як платника податків, так і контролюючих органів.

Із суб'єктивної точки зору *податковий ризик* з позиції платника податків – це економічна категорія, що відображає особливості сприйняття ним (платником податків) небезпеки можливих втрат (збитків), пов'язаних із адмініструванням та сплатою податків.

Платники податків, схильні до ризику, як правило, йдуть на порушення законодавства виводячи в «тінь» частину податкової бази за будь-яких умов. Частина платників податків вимушена ризикувати, щоб забезпечити діяльність підприємства, не будучи схильними до ризику. Ще одна категорія платників податків, маючи преференції і «захист» у вищих ешелонах влади, також, не будучи схильними до ризику, залишають у «тіні» значну частину неоподаткованої доданої вартості.

Усі розглянені категорії платників податків під час проведення контролюючими органами перевірок можуть отримати донарахування, штрафи тощо, але кожна категорія має різний ступінь ризику і його грошовий вимір. Можна зауважити, що дослідження суб'єктивної складової податкових ризиків є на сьогодні актуальним завданням у галузі оподаткування та ризикології.

Виходячи з наведених означень понять «податковий ризик», можна дати означення суб'єкта і об'єкта ризику у сфері оподаткування.

Суб'єктами податкового ризику виступають держава (державні контролюючі органи) та юридичні й фізичні особи, що є платниками податків. Однак вказані суб'єкти керують різними об'єктами ризику.

Об'єктами податкового ризику для держави є:

– діяльність платників податків, пов'язана з отриманням доходів та нарахуванням і сплатою податків;

– діяльність органів ДПС України (контролюючих органів),

операційна діяльність (процеси адміністрування податків), надання послуг платникам податків та державним установам, інші покладені на них функції.

Податковий ризик з позиції держави – це імовірність втрат податкових надходжень через порушення законодавства платниками податків та посадовими особами контролюючих органів, впровадження недосконалого податкового законодавства, неефективне адміністрування податків та інші зовнішні і внутрішні чинники.

Із суб'єктивної точки зору податковий ризик держави – це економічна категорія, яка відображає особливості сприйняття відповідальними суб'єктами управління процесами оподаткування невизначеності кінцевого результату щодо забезпечення надходжень коштів до державного бюджету унаслідок можливого впливу низки негативних факторів.

Щодо операційної діяльності органів ДПС України, всі процеси, що здійснюються у підрозділах цього відомства, можна розподілити на *основні* та *забезпечуючі*. До *основних операційних процесів*, за результатами яких оцінюють основну діяльності органів податкового відомства України, можна віднести такі процеси: 1) реєстрація платників податків; 2) обробка податкової звітності та платежів; 3) облік платежів; 4) податковий аудит; 5) апеляції платників податків; 6) погашення податкового боргу; 7) вирішення справ у судах за участі органів податкового відомства; 8) організація боротьби з відмиванням доходів, одержаних злочинним шляхом; 9) досудове слідство (кримінальні розслідування); 10) зв'язки з громадськістю; 11) обслуговування платників податків.

Забезпечуючі операційні процеси (процеси організаційно-управлінського блоку) – здійснюють основні операційні процеси. До них можна віднести такі: 1) інформатизація процесів адміністрування податків; 2) управління документами; 3) функціонування управлінської інформаційної системи; 4) розробка проектів нормативно-правових актів та впровадження законодавства; 5) матеріально-технічне та фінансове забезпечення; 6) управління персоналом; 7) внутрішній контроль тощо.

Об'єктом податкового ризику для платника є податкова політика держави, діяльність контролюючих органів, державні послуги щодо забезпечення сплати податків, подання звітності тощо. Прийнято вважати, що моніторинг та прогнозування цього процесу з позиції платника у вичерпній повноті на перспективу здійснити буде неможливо, тому завжди будуть існувати податкові ризики і невизначеність у діяльності господарюючих суб'єктів.

Запропонований підхід узгоджується із визначенням *операційного ризику* Базельським комітетом з питань банківського нагляду, що діє при Банку міжнародних розрахунків, і який визначається як «*ризик отримання збитків у результаті неадекватних або помилкових внутрішніх процесів, дій співробітників і систем або зовнішніх подій*».

За своїм змістом основні та забезпечуючі процеси є джерелами ризиків, що взаємопов'язані між собою і різною мірою впливають на якість результатів діяльності податкового відомства. Для систематизації ризиків операційної діяльності органів податкового відомства доцільно здійснити їх групування за напрямами діяльності та операційними процесами, які входять до них, і що дозволить ідентифікувати *типи, класи, групи, підгрупи та джерела ризиків.* Прийнято вважати, що ризики діяльності будь-якої організації поділяють на *два типи: зовнішні ризики та внутрішні* (екзогенні та ендогенні ризики відповідно). Підрозділи ДПС/ДФС України не є виключенням. Зовнішні ризики включають два класи ризиків: *некеровані* та *частково керовані ризики*, що вказує на можливість нейтралізації їх на стороні податкового відомства.

Внутрішні ризики можна умовно розподілити на два класи: ризики основних напрямів діяльності та ризики забезпечуючих (допоміжних) напрямів діяльності, і, відповідно, групи ризиків операційних процесів, що є їх складовою. Класифікація ризиків, що супроводжують діяльність органів податкового відомства України, подані на рис. 1.21.



Рис. 1.21. Ризики, властиві діяльності органів податкового відомства

Щодо груп *ризиків основних операційних процесів* органів ДПС України, як видно із табл. 1.11, вони можуть характеризуватися невизначеністю кінцевого результату основної операційної діяльності унаслідок неякісного їх виконання, неповного інформаційного забезпечення, можливого впливу (дії) на них об'єктивних та/або суб'єктивних факторів.

У табл. 1.11 подано групування ризиків основних напрямів діяльності та ризиків операційних процесів, що забезпечують їх.

Таблиця 1.11

Групування ризиків основних напрямів діяльності та ризиків операційних процесів податкового відомства

Групи ризиків основних напрямів діяльності	Ризики основних операційних процесів
1. Ризики обліково-реєстраційної діяльності	1. Реєстрація платників податків. 2. Обробка податкової звітності та платежів. 3. Облік платежів та ін.
2. Ризики контрольно-перевірочної роботи	1. Контрольно-перевірочна робота: 1.1. Камеральні перевірки. 1.2. Відбір платників податків для перевірок. 1.3. Виїзні перевірки. 1.4. Формування державного плану перевірок та ін.
3. Ризики юридично-правової діяльності	1. Супроводження апеляцій платників податків. 2. Вирішення справ у господарських судах за участі контролюючих органів. 3. Погашення податкового боргу. 4. Досудове слідство (кримінальні розслідування). 5. Організація боротьби з відмиванням доходів, одержаних злочинним шляхом та ін.

Ризики забезпечуючих напрямів діяльності або організаційно-управлінські ризики характеризуються можливістю (імовірністю) впровадження неефективної організаційної та інформаційної інфраструктури територіальних органів податкового відомства, впровадження неякісних прикладних інформаційних систем, методів прийняття управлінських рішень, неякісного планування, цілепокладання, прогнозування, що, у свою чергу, впливає на невизначеність результатів діяльності основних операційних процесів адміністрування податків. У табл. 1.12 подано групування ризиків за видами організаційно-управлінської діяльності.

Для управління ризиками важливим аспектом є виявлення причинно-наслідкових зв’язків ризиків та факторів, що їх породжують, розробка системи показників та критеріїв для вимірювання якості цих процесів.

Таблиця 1.12

Групування ризиків за видами організаційно-управлінської діяльності

Групування ризиків забезпечуючих процесів податкового відомства (організаційно-управлінські ризики)	
Групи ризиків	**Ризики забезпечуючи процесів**
1. Інформаційно-технологічні ризики	1. Неефективна інформатизація та автоматизація процесів адміністрування податків. 2. Відставання розвитку інформаційної інфраструктури від потреб операційної діяльності та ін.
2. Організаційно-управлінські ризики	1. Нерозвинена організаційна структура територіальних органів ДПС України. 2. Неефективне управління документами. 3.Неякісне управління персоналом (підготовка та підвищення кваліфікації). 4. Недостатнє матеріально-технічне та фінансове забезпечення та ін.
3. Ризики поглиблення недовіри до органів ДПС	1. Неякісний внутрішній контроль. 2. Неефективна антикорупційна діяльність 3. Недостатні зв’язки з громадськістю та ін.

Для прикладу у табл. 1.13 показано структурний опис ризику процесу «Відбір платників податків для перевірок», що належить до типу внутрішніх ризиків, класу ризиків основних операційних процесів, групи ризиків «Контрольно-перевірочної роботи», джерелами якого є: 1) неточна та недостовірна інформація про діяльність платників; 2) недостовірна і неякісна інформація про результати перевірок платників; 3) необ’єктивні показники для оцінювання рівня ризику платників; 4) неефективні методи класифікації платників за ознаками ризику. Власне, табл. 1.13 є паспортом даного ризику, на основі якого можна формувати показники для його вимірювання, спостерігати їхню зміну у динаміці, впливати відповідним чином на причини та фактори, що призводять до його появи та спрацювання.

Систематизація та формування об'єктивних показників ризиків за запропонованою системою класифікаційних ознак дозволить здійснювати комплексний аналіз вкладу у загальний ризик діяльності податкового відомства основних і забезпечуючих (допоміжних) операційних процесів, зовнішніх та внутрішніх факторів, визначати пріоритетні напрями для підвищення їхньої ефективності.

В управлінні ризиками особливо важливу роль відіграють методи кількісного вимірювання ризиків. Враховуючи накопичені масиви інформації в контролюючих органах, для оцінювання податкових ризиків можна застосовувати досить розвинені і апробовані на сьогодні методи і засоби інтелектуального аналізу даних: дискримінантний аналіз, кластерний аналіз, нейромережні алгоритми, класифікаційні методи розпізнавання та формування образів економічних явищ тощо.

Таблиця 1.13

Паспорт ризику процесу
«Відбір платників податків для перевірок»

Класифікаційна ознака ризику	**Зміст ознаки ризику**
Прояв дії ризику	1. Неефективне використання ресурсів контролюючих органів. 2. Збільшення перевірок законослухняних платників. 3. Поглиблення недовіри до контролюючих органів. 4. Зменшення надходжень до бюджету
Тип ризику	Внутрішній
Клас ризику	Основний напрям діяльності
Група ризику	Контрольно-перевірочна робота
Процес	Відбір платників податків для перевірок
Причини, джерела, (фактори) ризиків	1. Неточна та недостовірна інформація про діяльність платників. 2. Недостовірна інформація про результати перевірок платників. 3. Необ'єктивні показники для оцінювання рівня ризику платників. 4. Неефективні методи класифікації платників за ознаками ризику

Продовження табл. 1.13

Типові показники для розрахунку ймовірнісних параметрів ризику	1. Кількість перевірок платників податків вибраного для аналізу виду діяльності, кількість платників вибраного виду діяльності. 2. Сума надходжень за результатами перевірок у розрахунку на одного платника вибраного виду діяльності. 3. Кількість перевірок у розрахунку на одного платника податків вибраного виду діяльності та ін.

Узагальнено діяльність організації, направлену на врахування та управління ризиком, можна назвати *політикою ризику*. Вона включає весь комплекс дій та заходів, що мають на меті зменшити небезпеку на всіх стадіях функціонування організації, а саме: ідентифікацію та оцінку ризику (якісну та кількісну), вибір методів та механізмів управління ризиком, впровадження обраних методів та моніторинг результатів.

Для прикладу на рис. 1.22 подано управління ризиком за напрямом контрольно-перевірочної роботи у вигляді блок-схеми. Спираючись на результати ґрунтовного, якісного та кількісного аналізу, ризик-менеджер в управлінні ризиком може акцентувати увагу на зовнішніх, внутрішніх способах або їхній суперпозиції.

Необхідно зауважити, що якісний та кількісний аналіз ризику являє собою важливі самостійні складові у системі управління ризиками. Сучасні економічні умови мають високу ступінь невизначеності та вимагають від ризик-менеджера володіння широким спектром практичних і теоретичних знань, зокрема у сфері виявлення та ідентифікації ризику.

Отже, впровадження ризикоорієнтованого підходу в адмініструванні податків повинно базуватися на загальній теорії ризику, що забезпечить використання ряду апробованих методів для кількісного оцінювання ризиків. Податкові ризики необхідно розглядати як економічні категорії, що мають суб’єктивно-об’єктивну природу, і розподіляти на типи, класи та групи. Для ідентифікації ризиків діяльності податкового відомства доцільно групувати їх за основними напрямами операційної діяльності і оцінювати на основі об’єктивних показників результатів операційної діяльності.



Рис. 1.22. Узагальнена блок-схема процесу управління ризиками контрольно-перевірочної роботи

Завдання впровадження ризикоорієнтованого підходу в адміністрування податків є комплексним і включає вирішення організаційних, методологічних та інформаційних завдань і є актуальним особливо на етапі реформування податкової системи України.

Щодо *аналізу ризиків,* це етап, на якому застосовують методологію, за допомогою якої основні техніко-економічні параметри операційної діяльності піддаються аналізу з метою оцінювання впливу ризиків на результати діяльності податкового відомства.

Кількісна та якісна оцінка ризику ґрунтується на застосуванні імовірнісного підходу в оцінюванні ризику. Наприклад, імовірність порушення платниками податків податкового законодавства може визначатися на основі інформаційних масивів результатів перевірок типових господарюючих суб'єктів. Враховуючи, що імовірність Р може приймати значення від 0 до 1, якісні оцінки ризику можуть приймати, наприклад, такі значення: Р=0–0,33 «малий ризик»; Р=0,33–0,66 «середній ризик»; Р=0,66–1 «великий ризик». Власне, *імовірнісна оцінка ризику* ґрунтується на визначенні значення імовірності *Р*, що може використовуватися у визначенні *кількісної міри ризику*, вираженої у грошовій формі як очікуваний розмір втрат. *Якісна оцінка ризиків* здійснюється на основі значень кількісних оцінок, застосовуючи встановлені критерії або правила: великий, помірний, середній ризик тощо.

Значним етапом з впровадження ризикоорієнтованого адміністрування податків в органах ДПС/ДФС України є впровадження технологій відбору платників податків для перевірок на основі віднесення їх до категорій ризику або категорій уваги.

Категорія уваги – характеристика суб'єктів господарювання, сформована на основі їхніх обліково-реєстраційних даних, показників діяльності та історії порушень податкового законодавства.

Розрізняють такі категорії уваги:

Категорія 1 – сумлінні платники податків. Оцінюються позитивно. Найбільш істотними факторами, які впливають на віднесення платника податку до цієї категорії, є:

– відсутність випадків порушення термінів подачі податкової звітності;

– повнота та своєчасність сплати податків і зборів (обов'язкових платежів) до бюджетів та державних цільових фондів;

– більший за середній у групі (галузі) рівень сплати податків і зборів (обов'язкових платежів) до бюджетів та державних цільових фондів.

Категорія 2 – платники податків помірного ризику. У цілому оцінюються позитивно. Найбільш істотними факторами, які впливають на віднесення платника податку до даної категорії, є:

– низьке значення фіскальної важливості;

– середній рівень сплати податків і зборів (обов'язкових платежів) до бюджетів та державних цільових фондів у групі (галузі).

Категорія 3 – платники податків високого ризику. Оцінюються негативно в частині податкової поведінки. Найбільш істотними факторами, які впливають на віднесення платника податку до даної категорії, є:

– високий рівень значення фіскальної важливості, але рівень сплати податків і зборів (обов'язкових платежів) до бюджетів та державних цільових фондів у групі (галузі) є значно меншим за середні показники;

– наявність ознак серйозних порушень податкового законодавства, на які вказують фактори ризику.

Своєчасна ідентифікація ризику та станів, до яких він може призвести, дає можливість вчасно попередити небажані наслідки, обрати більш гнучку стратегію. У разі підтвердження (перевірки) за допомогою кількісних оцінок показників ефективності і ризикованості з'являється можливість формувати достовірні прогнози щодо майбутньої діяльності та планувати економічні результати. Якісний та кількісний аналіз ризиків являє собою важливі самостійні складові у системі аналізу податкового ризику.

1.3.7. Імовірнісне та кількісне оцінювання податкових ризиків

Розглянемо *імовірність* як один з підходів до оцінки податкового ризику. Оцінюючи ризик, на практиці часто обмежуються спрощеними підходами, спираються на один чи кілька основних показників (критеріїв), параметрів, які являють собою найважливіші узагальнені характеристики у цій конкретній ситуації. У ряді випадків, зокрема в оподаткуванні, величину (ступінь)

ризику визначають як ймовірність податкових втрат. У цьому випадку:

$$W = p_н ,$$

де $p_н$ – ймовірність настання небажаних наслідків,

W – величина податкового ризику.

Типовий графік кривої розподілу ймовірностей перевищення певного рівня податкових втрат зображено на рис.1.23 (крива $W(x)$).

Аналізуючи податкові втрати, у кожній із запропонованих зон ризику потрібно поставити у відповідність кількісні показники, критерії ризику. У прикладних проблемах економічного ризику для оцінки його величини широке використання має *ймовірність перевищення заданого рівня збитків.* Цю ймовірність обчислюють за формулою:

$$W(x) = P(X \geq x) = 1 - P(X < x) = 1 - F(x). \qquad (1.37)$$

У нашому випадку можна виділити три найважливіші базові показники ризику. Показник допустимого податкового ризику:

$$W_{дп} = W(x_{дп}) = P(X \geq x_{дп}),$$

тобто $W_{дп}$ – це ймовірність того, що податкові втрати виявляться більшими ніж їх гранично допустимий рівень $x_{дп}$.

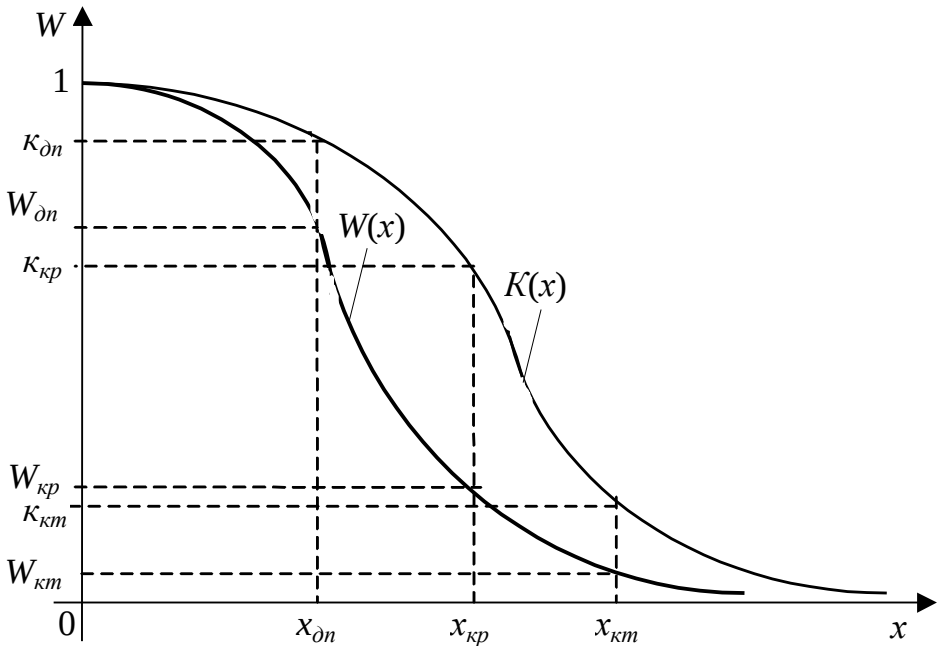


Рис. 1.23. Порівняння очікуваної ймовірності перевищення випадкових податкових втрат з гранично допустимою

Показник критичного податкового ризику:

$$W_{кр} = W(x_{кр}) = P(X \geq x_{кр}), \qquad (1.38)$$

тобто $W_{кр}$ – це ймовірність того, що втрата податків виявиться більшою ніж їхній гранично допустимий критичний рівень $x_{кр}$.

Показник катастрофічного податкового ризику:

$$W_{кт} = W(x_{кт}) = P(X \geq x_{кт}), \qquad (1.39)$$

тобто $W_{кт}$ – це ймовірність того, що податкові втрати виявляться більшими ніж їхній гранично допустимий катастрофічний рівень $x_{кт}$.

Знання цих показників дає змогу приймати обґрунтовані рішення щодо певних дій у здійсненні податкової політики: зміни ставок оподаткування, зміни процедур адміністрування податків, введення додаткових обмежень чи звільнень стосовно видів діяльності, категорій платників тощо. Але для остаточного прийняття рішення інформації про значення названих показників недостатньо – необхідно ще задати (встановити, прийняти) їхні граничні величини, щоб не потрапити в зону неприйнятного ризику. Такі величини називають *критеріями відповідно допустимого, критичного та катастрофічного ризику – $к_{дп}$, $к_{кр}$, $к_{кт}$*.

Отже, маючи значення трьох показників ризику та критеріїв граничного ризику, приходимо до таких найбільш загальних умов прийнятності рівня ризику в оподаткуванні:

$$W(x_{дп}) \leq к_{дп};$$

$$W(x_{кр}) \leq к_{кр};$$

$$W(x_{кт}) \leq к_{кт}.$$

Графічне пояснення основних умов прийнятності ризику зображено на рис. 1.23 (крива $K(x)$).

Приклад задачі

Під час розробки законодавчих ініціатив Міністерством фінансів аналізується пропозиція підприємців щодо зменшення ставки оподаткування певного виду підприємницької діяльності. Необхідно обчислити величину податкових втрат у вигляді відсотка відносно розрахункової суми бюджетних надходжень. Експертами було встановлено, що обчислена величина податкових втрат відповідає нормальному закону розподілу з параметрами m = 20 % (математичне сподівання) та σ = 4 % (середньоквадратичне відхилення). Експерти встановили для себе такі критерії ризику: $к_{дп}$ = 20 %; $к_{кр}$ = 5 %; $к_{кт}$ = 0,1 %.

Як бути Міністерству фінансів, якщо експерти від групи підприємців, що домагаються отримати зниження ставок оподаткування, вважає реальними такі показники ризику: $x_{дп}$ = 24 %; $x_{кр}$ = 28 %; $x_{кт}$ = 32 %?

Для вирішення цього завдання необхідно згадати теоретичні основи випадкових величин та їхні основні показники, закони розподілу.

Неперервною називають випадкову величину, що може набувати довільні значення з певного скінченого або нескінченного проміжку. Кількість можливих значень неперервної випадкової величини – нескінченна.

Інтегральною функцією розподілу випадкової величини називають функцію $F(x)$, яка визначає для довільного x ймовірність того, що випадкова величина X набуде значення, меншого за x:

$$F(x) = P(X < x) \qquad (1.40)$$

Ймовірність того, що випадкова величина набуде значення на проміжку (a,b), дорівнює приросту функції розподілу $F(x)$ на цьому проміжку:

$$P[X \in (a,b)] = F(b) - F(a) \qquad (1.41)$$

Оскільки ймовірність того, що випадкова величина X набуде конкретного можливого значення, завжди дорівнює нулю, тобто $P\ X = x_k\ = 0$, то мають місце рівності:

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b).$$

Диференціальною функцією розподілу або *щільністю ймовірностей* неперервної випадкової величини називають функцію $f(x)$, яка дорівнює похідній першого порядку від її інтегральної функції розподілу: $f(x) = F'(x)$. Ймовірність того, що неперервна випадкова величина X набуде будь-яке значення з проміжку (a,b), обчислюється за формулою:

$$P(X \in (a,b)) = \int_a^b f(x)dx \qquad (1.42)$$

Функцію розподілу випадкової величини визначають через її функцію щільності $f(x)$ таким чином:

$$F(x) = \int_{-\infty}^{x} f(t)dt. \qquad (1.43)$$

Умову нормування через функцію щільності записують таким чином: $\int_{-\infty}^{\infty} f(t)dt = 1$.

Якщо неперервна випадкова величина X визначена лише на проміжку (a,b), то умова формування має вигляд $\int_a^b f(t)dt = 1$.

У випадку неперервних випадкових величин математичне сподівання, дисперсія та середнє квадратичне відхилення мають такий самий зміст та властивості, як і для дискретних випадкових величин, але обчислюють їх за іншими формулами. Якщо неперервна випадкова величина приймає значення із інтервалу $(-\infty,-\infty)$ та має диференціальну функцію розподілу $f(x)$, то її математичне сподівання та дисперсія обчислюються за формулами:

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx \qquad D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - M^2(X).$$

Якщо неперервна випадкова величина приймає значення з інтервалу (a,b), то:

$$M(X)=\int_a^b xf(x)dx \qquad D(X)=\int_a^b (x-M(X))^2 f(x)dx=\int_a^b x^2 f(x)dx-M^2(X)$$

Числову характеристику закону розподілу випадкової величини

$$\sigma(X)=\sqrt{D(X)}\,, \tag{1.44}$$

називають середньоквадратичним відхиленням або стандартним відхиленням.

Основні закони розподілу неперервних випадкових величин розподіляють за виглядом їхніх диференціальних функцій розподілу (щільності ймовірностей) $f(x)$. Найчастіше використовуються такі закони розподілу.

Рівномірний закон розподілу. Розподіл ймовірностей називається *рівномірним,* якщо на інтервалі, якому належать усі можливі значення випадкової величини, функція щільності розподілу дорівнює константі. Випадкова величина X, розподілена рівномірно на проміжку $[a,b]$, має таку функцію щільності розподілу:

$$f(x)=\begin{cases} 0, & x\le a;\\ \dfrac{1}{b-a}, & a<x\le b;\\ 0, & x>b. \end{cases} \tag{1.45}$$

Функція розподілу ймовірностей має вигляд:

$$F(x)=\begin{cases} 0, & x\le a;\\ \dfrac{x-a}{b-a}, & a<x\le b;\\ 0, & x>b. \end{cases} \tag{1.46}$$

Числові характеристики рівномірного закону:

$$M(X)=\frac{a+b}{2};\quad D(X)=\frac{(b-a)^2}{12};\quad \sigma(X)=\frac{b-a}{2\sqrt{3}}. \tag{1.47}$$

Нормальний закон розподілу. Нормально розподіленою з параметрами a *та* σ називається випадкова величина X, функція щільності розподілу якої має вигляд:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}},\ -\infty<x<+\infty,\ -\infty<a<+\infty,\ \sigma>0 \qquad (1.48)$$

Підпорядкування випадкової величини X нормальному закону розподілу з параметрами a та σ позначають $N(a;\sigma)$. Графік функції щільності нормального розподілу називають *кривою Гауса або нормальною кривою.* Параметрами a та σ впливають на форму кривої розподілу $M(X)=a$ та $D(X)=\sigma^2$, де a та σ – параметри розподілу.

Інтегральна функція нормального розподілу має вигляд:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^2}}dt. \qquad (1.49)$$

Розподіл $N(0;1)$ називають *стандартним нормальним розподілом*. У цьому випадку:

$$f(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}},\qquad -\infty<x<+\infty; \qquad (1.50)$$

$$F(x)=\frac{1}{\sqrt{2\pi}}\cdot\int_{-\infty}^{x}e^{-\frac{t}{2}}dt=0{,}5+\frac{1}{\sqrt{2\pi}}\cdot\int_{0}^{x}e^{-\frac{t}{2}}dt=0{,}5+\Phi(x), \qquad (1.51)$$

де $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{t^2}{2}}dt$, – функція Лапласа ($\Phi(x)\approx 0{,}5$ при $x>5$; $\Phi(-x)=-\Phi(x)$.

Ймовірність того, що нормально розподілена випадкова величина X прийме значення на проміжку (α,β), визначається за формулою:

$$P(\alpha<X<\beta)=\int_{\alpha}^{\beta}f(t)dt=\frac{1}{\sigma\sqrt{2\pi}}\cdot\int_{\alpha}^{\beta}e^{-\frac{(t-a)^2}{2\sigma^2}}dt=\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi\left(\frac{\alpha-a}{\sigma}\right), \qquad (1.52)$$

Ймовірність заданого відхилення

$$P(|X-a|<\delta)=P(a-\delta<X<a+\delta)=$$
$$=\Phi\left(\frac{a+\delta-a}{\sigma}\right)-\Phi\left(\frac{a-\delta-a}{\sigma}\right)=2\Phi\left(\frac{\delta}{\sigma}\right). \quad (1.53)$$

Правило трьох сигм для нормального закону. Коли $\delta=3\sigma$, то

$$P(|x-a|<3\sigma)=2\Phi\left(\frac{3\sigma}{\sigma}\right)=2\Phi(3)=2\cdot 0{,}49865=0{,}9973. \quad (1.54)$$

Звідси випливає:

$$P(|x-a|>3\sigma)=1-P(|x-a|<3\sigma)=1-0{,}9973=0{,}0027.$$

Тобто ймовірність того, що внаслідок проведення експерименту випадкова величина X, яка має розподіл $N(a;\sigma)$, не потрапить у проміжок $[a-3\sigma;a+3\sigma]$, дорівнює 0,0027.

*Сутність правила трьох сигм: я*кщо випадкова величина розподілена нормально, то абсолютна величина її відхилення від математичного сподівання не перевищить потроєного середньоквадратичного відхилення.

✎ *Розв'язання задачі.* Оскільки у випадку нормального розподілу випадкової величини X інтегральна функція розподілу:

$$F(x) = 0{,}5 + \Phi\left(\frac{x-m}{\sigma}\right),$$

де $\Phi(t)$ – функція Лапласа, то

$$W_{дп}= P(X> x_{дп})= 1- P(X < x_{дп}) = 1-F(x_{дп})=$$

$$=1-0{,}5-\left(\Phi\left(\frac{x_{дп}-m}{\sigma}\right)\right)=0{,}5-\Phi\left(\frac{24-20}{4}\right)=0{,}5-\Phi(1)=0{,}5-0{,}3413=0{,}1587, \quad (1.55)$$

тобто $W_{дп} = 0{,}1587 < 0{,}2 = к_{дп}$.

Аналогічно знаходимо, що

$$W_{кр} = P(X > x_{кр}) = 0,5 – Ф(2) = 0,5 – 0,4772 = 0,0228,$$

тобто $W_{кр} = 0,0228 < 0,05 = к_{кр}$;

$$W_{кт} = P(X > x_{кт}) = 0,5 – Ф(3) = 0,5 – 0,49865 = 0,00135,$$

тобто $W_{кт} = 0,00135 > 0,001 = к_{кт}$.

Виходячи з того, що $W_{кт} > к_{кт}$, а також враховуючи, що Міністерство фінансів – відомство, не схильне до ризиків, робимо висновок, що ініціатива підприємців щодо зменшення ставок оподаткування їх діяльності не буде підтримана.

Питання для самоконтролю до теми 1.3

1. Особливості схеми управління органом податкового відомства.
2. Що означає багатокритеріальність та полісистемність середовища?
3. Інфраструктура органів податкового відомства України.
4. Назвіть три умови, у яких приймаються рішення.
5. У чому особливість прийняття рішень в умовах ризику?
6. У чому суть збалансованої системи показників?
7. У чому суть методології ключових показників ефективності?
8. У чому суть інтегральної згортки показників?
9. Для чого використовують вагові коефіцієнти?
10. Як розраховують інтегральний критерій якості операційних процесів?
11. Поясніть відмінність відносного показника затратності від відносного показника продуктивності.
12. Для чого здійснюється нормування показників?
13. Як ви поясните, що означає сігма-управління?
14. У чому суть методології Лаффера?
15. У чому відмінність двопараметричного методу від трипараметричного методу?
16. Поясніть економічний зміст точок Лаффера першого та другого роду.

17. Як розраховують ефективну ставку оподаткування ВРП?

18. Що характеризує індекс фіскальної ефективності?

19. Як можна розрахувати коефіцієнт результативності перевірочної роботи?

20. Як можна розрахувати коефіцієнт виконання бюджету з податкових надходжень?

21. Наведіть приклад використання критерію ефективності функціонування податкової служби.

22. У чому суть операційного ризик-менеджменту?

23. Дайте декілька визначень поняття «ризик».

24. Чому поняття «імовірність» тісно пов'язане із поняттям «ризик»?

25. Дайте визначення податкового ризику з позиції платника податків.

26. Дайте визначення податкового ризику з позиції держави.

27. Назвіть основні та забезпечуючі процеси податкового відомства.

28. Назвіть зовнішні та внутрішні ризики, що супроводжують діяльність органів податкового відомства.

29. У чому відмінність частково керованих ризиків та ризиків забезпечуючих напрямів діяльності податкового відомства?

30. Назвіть групи ризиків забезпечуючих процесів.

31. Що є причиною, джерелом (фактором) ризиків процесу «Відбір платників податків для перевірок»?

32. Які показники використовують для оцінювання ступеня ризику?

33. Охарактеризуйте категорії уваги платників податків: категорія 1, 2, 3.

34. Охарактеризуйте суть застосування нечітких множин у оцінюванні ризиковості платників податків.

35. Що характеризує диференціальна функція розподілу імовірностей?

36. Чим характеризується нормальний закон розподілу?

37. Суть функції Лапласа у визначенні ймовірності попадання випадкової величини у діапазон значень?

1.4. СКБД Oracle – базова платформа інформаційно-аналітичної системи податкового відомства України

1.4.1. Клієнт-серверна інфраструктура інформаційної системи податкового відомства України

Основний компонент інформаційної інфраструктури податкового відомства ІТС «Податковий блок» реалізований на платформі СКБД Oracle. Призначенням ІТС «Податковий блок» є централізація збору інформації, підвищення ефективності діяльності податкового відомства за рахунок розширення електронних сервісів, інформаційних та аналітичних технологій. ІТС «Податковий блок» побудована за трирівневою архітектурою «клієнт-сервер» з використанням як сервер задач WEB-сервера (рис. 1.24).

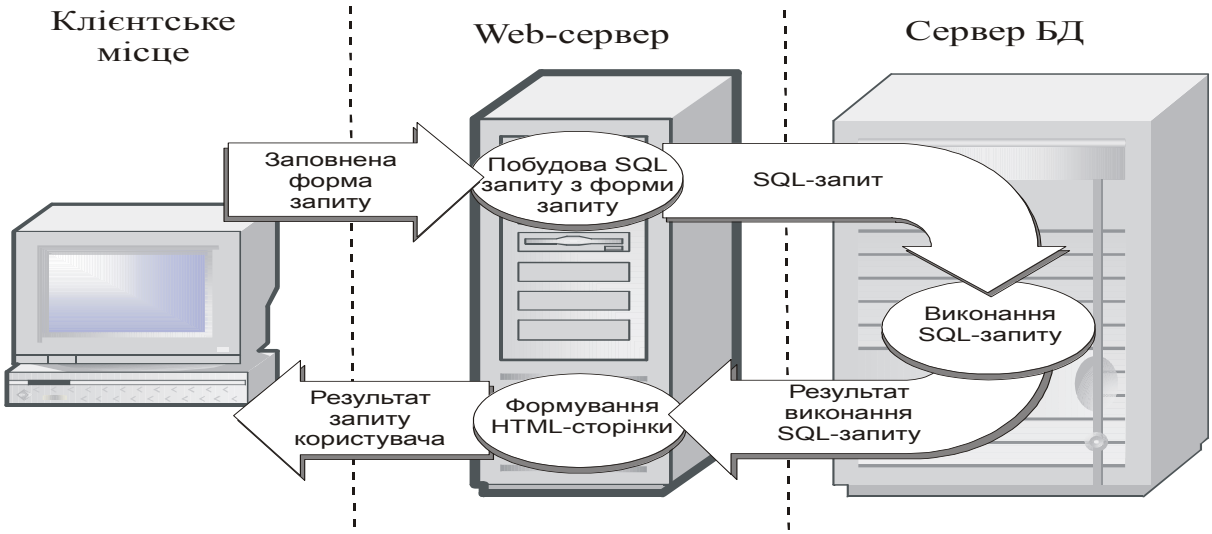


Рис. 1.24. Загальна схема обліку, накопичення, пошуку та надання інформації в інформаційній системі податкового відомства

Сервер баз даних – впорядкована сукупність даних, які призначені для зберігання, накопичення та обробки. Реляційна база даних податкового відомства функціонує на платформі Oracle і базується на реляційній моделі даних. Для адміністрування бази даних, забезпечення доступу до інформації, що в ній зберігається, її обробка та поновлення, використовується спеціальна мова і

набір програмних засобів, що у сукупності становлять *систему керування базою даних* (далі – СКБД). СКБД інтегрує відомості з різних джерел в одній реляційній базі даних (РБД).

Реляційна модель даних – модель даних, де текстова чи числова інформація подається у вигляді таблиць. Кожна таблиця, яка називається відношенням, складається з рядків, які називають *кортежами*, та стовпчиків, які називають *атрибутами*.

Web-сервер – ПО, що приймає HTTP-запити від клієнтів, зазвичай веб-браузерів, і видає їм HTTP-відповіді, як правило, разом із HTML-сторінкою, зображенням, файлом із медіапотоком або іншими даними.

Клієнтське місце – зазвичай веб-браузер, що передає веб-серверу запити на отримання ресурсів, позначених URL-адресами. У відповідь веб-сервер передає клієнту запитані дані. Обмін відбувається за протоколом HTTP.

Реляційна модель даних вперше була запропонована британським ученим, співробітником компанії IBM Едгаром Франком Коддом (E. F. Codd) 1970 року в статті «A Relational Model of Data for Large Shared Data Banks». Сьогодні ця модель є фактичним стандартом, на який орієнтуються практично всі сучасні комерційні СУБД.

У реляційній моделі досягається набагато вищий рівень абстракції даних, ніж в ієрархічній або мережевій. У статті Е. Ф. Кодда стверджується, що «реляційна модель надає засоби опису даних на основі тільки їх природної структури, тобто без потреби введення якої додаткової структури для цілей машинного представлення». Іншими словами, подання даних не залежить від способу їх фізичної організації. Це забезпечується за рахунок використання математичної теорії відносин (сама назва «реляційна» походить від англійського relation – «відношення»).

Реляційна модель орієнтована на організацію даних у вигляді двовимірних таблиць. Кожна реляційна таблиця являє собою двовимірний масив і має такі властивості:

– кожна клітинка таблиці – один елемент даних;

– усі стовпці в таблиці однорідні, тобто всі елементи в стовпці мають однаковий тин (числовий, символьний тощо) і однакову довжину;

– кожен стовпець має унікальне ім’я;

– однакові рядки в таблиці відсутні;

– порядок проходження рядків і стовпців може бути довільним.

Мова програмування, що використовується для доступу до реляційних баз даних, називається мовою структурованих запитів **(SQL)** – гнучка й ефективна мова, усі засоби якої служать для маніпулювання реляційними даними та для їхнього дослідження.

PL/SQL – мова програмування, розроблена корпорацією Oracle. ***PL/SQL*** розширює можливості *SQL*, додаючи у неї конструкції процедурних мов.

До баз даних, а також до СКБД пред’являють такі вимоги:

– висока швидкодія;

– простота відновлення даних;

– незалежність даних;

– можливість значного використання даних;

– безпека даних;

– стандартизація побудови й експлуатації БД;

– адекватність відображення даних відповідної до предметної області.

Висока швидкодія передбачає малий проміжок часу від моменту запиту до бази даних до моменту реального одержання даних.

Незалежність даних забезпечує мінімальні зміни структури бази даних при змінах стратегії доступу до даних і структури самих вихідних даних.

Безпека даних – це захист даних від навмисного або ненавмисного порушення таємності, викривлення або руйнування. Безпека включає два компоненти: цілісність і захист даних від несанкціонованого доступу.

Цілісність даних – стійкість збережених даних до руйнування й знищення, пов’язаного з несправностями технічних засобів, системними помилками й помилковими діями користувачів.

Захист даних від несанкціонованого доступу припускає обмеження доступу до певних даних бази й досягається введенням заходів безпеки: розмежуванням прав доступу до даних різних користувачів залежно від виконуваних ними функцій і/або посадових обов’язків; уведенням захисту у вигляді паролів.

Стандартизація забезпечує наступність поколінь конкретної СУБД, спрощує взаємодію баз даних одного покоління СУБД із однаковими й різними моделями даних.

Клієнт-сервер (англ. Client server) – мережева архітектура, в якій пристрої є або клієнтами, або серверами. Клієнтом є машина, яка здійснює запит, сервером – машина, яка відповідає на запит. Обидва терміни (клієнт і сервер) можуть застосовуватися як до фізичних пристроїв, так і до програмного забезпечення.

Технологія клієнт-сервер має низку переваг:

– забезпечується більш широкий доступ до існуючих баз даних;

– підвищується загальна продуктивність системи: оскільки клієнти і сервер знаходяться на різних комп'ютерах, їхні процесори здатні виконувати програми паралельно. Настроювання продуктивності комп'ютера з сервером спрощується, якщо на ньому виконується тільки робота з базою даних;

– знижується вартість апаратного забезпечення; досить потужний комп'ютер з великим пристроєм зберігання потрібен тільки серверу – для зберігання і управління базою даних;

– скорочуються комунікаційні витрати. Додатки виконують частину операцій на клієнтських комп'ютерах і надсилають через мережу тільки запити до баз даних, що дозволяє значно скоротити обсяг пересилання по мережі даних.

Технологію клієнт-сервер можна описати таким алгоритмом:

– клієнт формує і посилає запит до бази даних сервера (до програми, яка обробляє запити);

– ця програма проводить маніпуляції з базами даних, що знаходяться на сервері, відповідно до запиту формує результат і передає його клієнту;

– клієнт отримує результат, відображає його на дисплеї і чекає подальших дій користувача. Цикл повторюється доти, доки користувач не завершить роботу з сервером.

Стандартне програмне забезпечення, що реалізує технологію клієнт-сервер, має хорошу масштабованість (ефективне використання апаратного забезпечення), стійкість у роботі, захист від

несанкціонованого доступу і потужність при роботі з великими проектами в галузі баз даних.

Усе це підвищує швидкодію системи й знижує час очікування результату запиту. Під час виконання запитів сервером суттєво підвищується ступінь безпеки даних, оскільки правила цілісності даних визначаються в базі даних на сервері і є єдиними для всіх додатків, що використовують цю БД.

1.4.2. Логічні та фізичні об'єкти сервера бази даних Oracle

Сервер Oracle складається із *бази даних* Oracle і *екземпляра* бази даних Oracle (рис. 1.25). На одному сервері Oracle може бути відкрито декілька екземплярів баз даних.

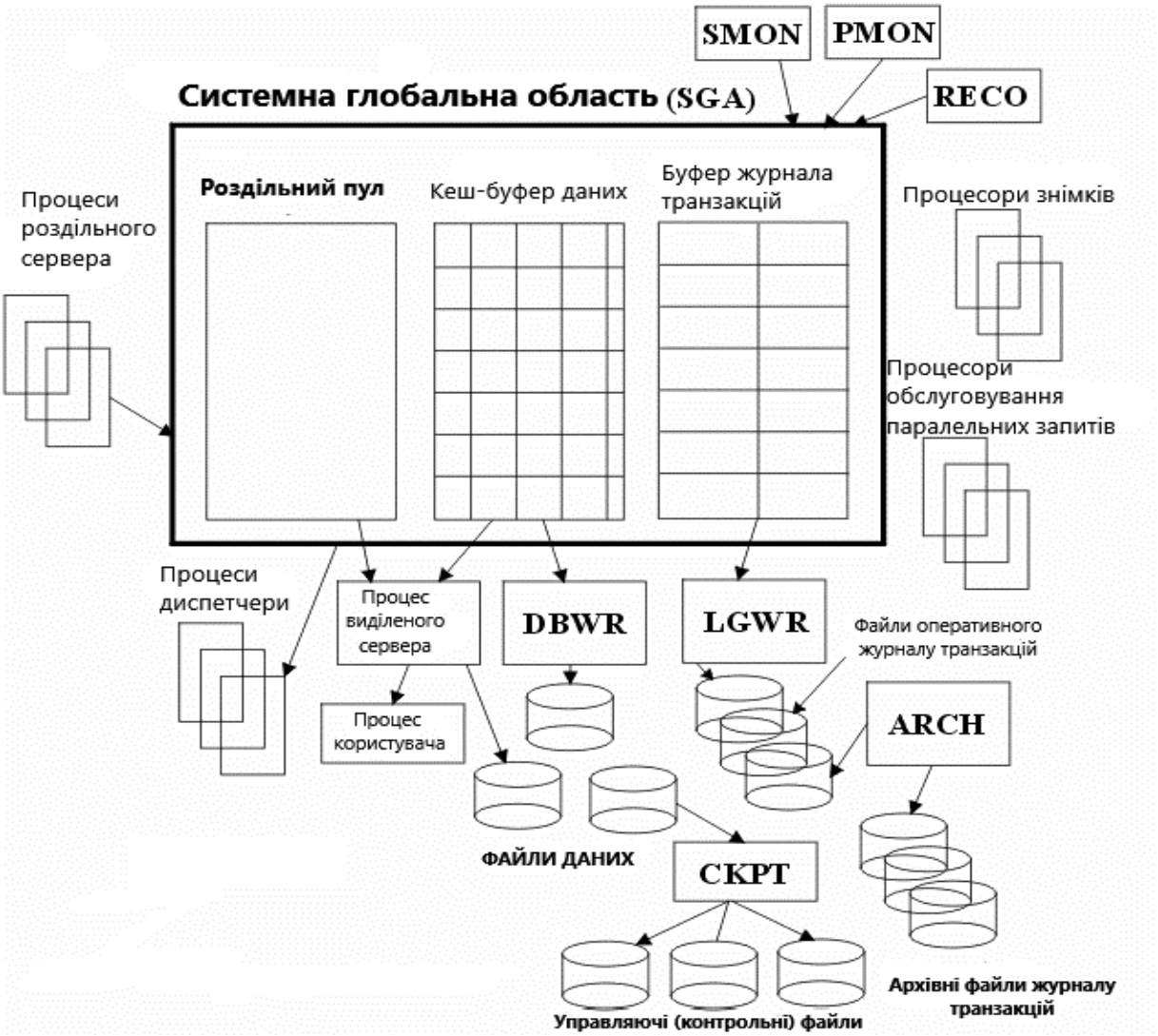


Рис. 1.25. Архітектура сервера Oracle (екземпляр + база даних)

Екземпляр Oracle складається із буферів пам'яті, що утворюють системну глобальну область (System Global Area – SGA), і фонових процесів, які контролюють і виконують більшу частину прихованої роботи екземпляра.

Екземпляр не існує до моменту його старту. При запуску читається файл параметрів ініціалізації і на його основі здійснюється конфігурація екземпляра.

Користувачі можуть з'єднуватись із БД після того, як екземпляр буде запущений і БД буде відкрито.

При інсталяції бази даних Oracle встановлюється програмне забезпечення СУБД. Крім СУБД, необхідна власне сама база даних. Одна система управління базою даних може працювати відразу з декількома базами даних на одному сервері. Кожна така база даних у термінології Oracle називається екземпляром бази даних (Data Base Instance). Кожен запущений екземпляр активно використовує ресурси процесора, оперативної та дискової пам'яті.

База даних Oracle. Кожна база даних Oracle містить такі фізичні об'єкти (рис. 1.26):

– управляючі (контрольні) файли (control files), які містять дані про структуру БД, про резервування БД та іншу інформацію;

– файли даних (data files), що власне містять дані;

– файли оперативних журналів (online redo log files), що дозволяють відновити БД після збою екземпляра. Якщо робота БД завершується аварійно і при цьому не втрачаються файли даних, екземпляр може відновити БД на основі даних таких файлів.

Файли, що не входять до БД, але належать до неї:

– файли параметрів ініціалізації, що використовуються для збереження параметрів екземпляра, з якими він стартує;

– файл паролів, який зберігає паролі адміністраторів БД. Ці файли дозволяють віддалено адміністраторам під'єднуватись до БД;

– архівні журнали (archive log files), які містять історію створених екземпляром оперативних журнальних файлів (їх автономні копії) і які використовуються для відновлення файлів БД.

Файли даних. Для накопичення та зберігання інформації, яка знаходиться у логічних структурах, створюється один або декілька файлів даних. Файл даних складається з *блоків бази даних*, які відображаються на *дискових блоках* операційної системи. Розмір блоку в Oracle варіюється від 2 до 32 Кбайт.

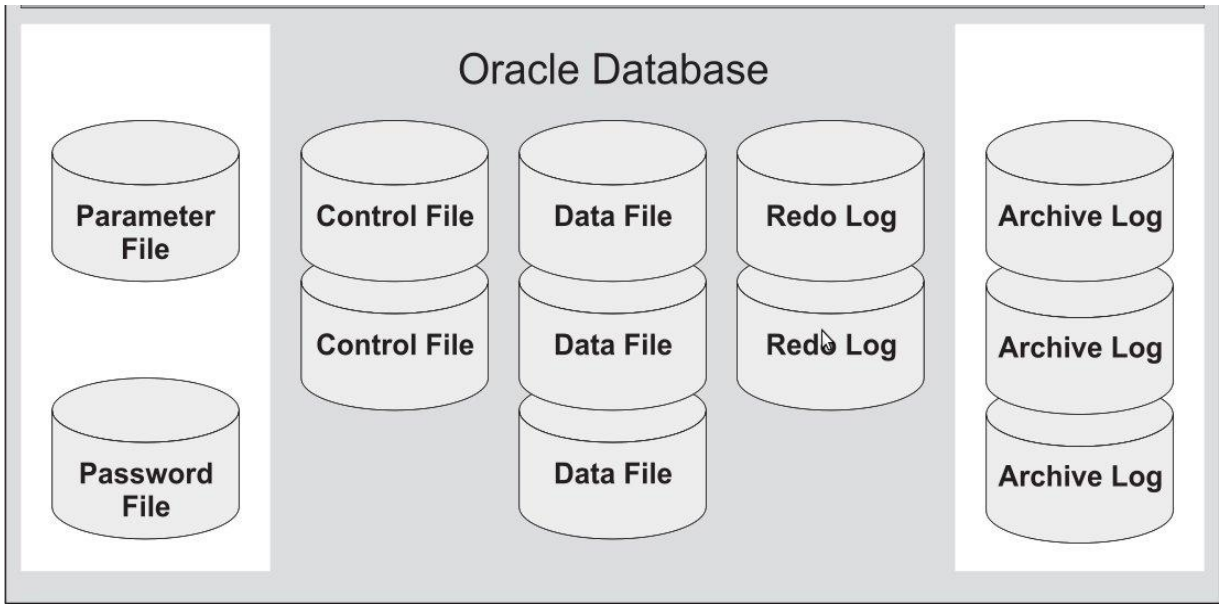


Рис. 1.26. Перелік компонентів бази даних Oracle

Кожен файл даних належить тільки одній базі даних і тільки одному табличному простору в ній. Дані зчитуються з дисків в оперативну пам'ять блоками Oracle. Блоки даних переписуються з пам'яті в файли даних на диску, коли це потрібно для гарантії збереження внесених користувачами змін.

Файли даних – це найнижчий рівень гранулярності взаємодій між Oracle і операційною системою. При розміщенні бази даних на фізичних пристроях найменша сутність, яку можна кудись помістити, – це файл. Оптимізація підсистеми введення/виведення для підвищення продуктивності Oracle зазвичай включає переміщення файлів даних з одного набору дисків на інший. З точки зору фізичної організації, файл даних являє собою набір блоків операційної системи. А з логічної – у файлах даних можна виділити три структурних рівні: блоки даних, екстенти і сегменти (рис. 1.27).

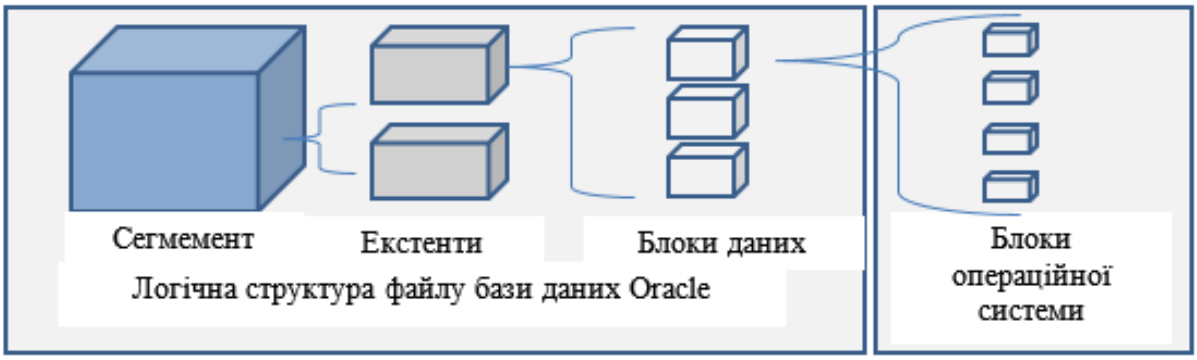


Рис. 1.27. Структура файлу бази даних Oracle

Сегмент – це об’єкт у базі даних Oracle, наприклад таблиця або індекс, що складається із одного або декількох екстентів.

Екстент – набір суміжних блоків даних у файлі даних.

Блок даних Oracle. Розмір блоку даних встановлюється при створенні БД. Розмір 8К підходить для багатьох БД. Якщо БД використовується як сховище даних з великими таблицями та індексами, то використання більших розмірів блоків дає виграш у продуктивності. Використовуючи БД в оперативних, трансакційних прикладних системах, можна задавати менші розміри блоків. Максимальний розмір блоку файлу даних Oracle залежить від ОС.

Виконуючи операцію оновлення, Oracle спочатку намагається оновити дані, не виходячи за межі вихідного блоку. Якщо в цьому блоці недостатньо місця для нової інформації, то Oracle записує дані в новий блок, який може знаходитися в іншому екстенті. В Oracle Database 10g доданий консультант Segment Advisor, який спрощує процедуру звільнення невикористаного простору.

1.4.3. Логічні структури екземпляра бази даних Oracle. Табличні простори і схеми

Екземпляр бази даних Oracle розподілений на логічні структури зберігання, що називаються табличними просторами (**TABLESPACES**) та схемами. Вони використовуються для об’єднання пов’язаних структур із прив’язкою до їхніх власників. Кожна база даних вміщує, як правило, декілька табличних просторів. Схема може мати об’єкти у різних табличних просторах.

Табличні простори (TABLESPACES). БД Oracle складається із однієї або більше логічних структур, що називаються табличними просторами. Кожен табличний простір складається із одного або декількох фізичних файлів даних. Найпростіша БД Oracle може мати один табличний простір, що складається із одного файлу даних. Максимальна кількість файлів даних 65 535. Як правило, БД за замовчуванням створюється із такими табличними просторами:

SYSTEM – табличний простір для управління БД, який містить словник даних і таблиці з адміністративною інформацією про БД. Усі ці об'єкти розміщуються у схемі SYS (адміністратору або іншим користувачам із адміністративними привілеями доступу до БД);

SYSAUX – табличний простір додатковий до SYSTEM. Деякі компоненти і продукти, які у попередніх версіях використовували табличний простір SYSTEM з версії Oracle10g, використовують табличний простір SYSAUX.

TEMP – табличний простір, що використовується для зберігання тимчасових таблиць та індексів під час обробки команд SQL, а також для виділення простору для сортування. У преконфігурованій БД TEMP – тимчасовий табличний простір за замовчуванням.

UNDOTBS1 – табличний простір, у якому сервер БД зберігає інформацію відміни трансакцій, змін.

USERS – табличний простір для зберігання об'єктів і даних користувачів. У преконфігурованій БД USERS – тимчасовий табличний простір за замовчуванням для усіх об'єктів, що належать користувачу, крім SYS і SYSTEM.

EXAMPLE – табличний простір, що містить об'єкти схем з прикладами, які можуть бути встановлені у процесі створення БД. Ці схеми використовують у документації Oracle.

Схема (SCHEMA) – це колекція об'єктів: таблиці, представлення, послідовності, процедури, синоніми, індекси, кластери і зв'язки баз даних, що належать користувачу. (Не існує жорсткого зв'язку між табличним простором і схемою; об'єкти однієї і тієї ж схеми можуть знаходитися у різних табличних просторах, один і той табличний простір може містити об'єкти різних

схем (користувачів)). Користувач може бути пов’язаний лише із однією схемою. Поняття «користувач» і «схема» часто використовуються як взаємозамінні.

Процедура створення користувача включає такі дії: вибирається профіль; метод аутентифікації; призначається робочий і тимчасовий табличний простори за замовчуванням; надаються потрібні привілеї; виділяються квоти на табличні простори.

1.4.4. Об’єкти та логічні структури зберігання і доступу до даних

Таблиця (TABLE) – це основна одиниця зберігання даних у базі даних Oracle. Таблиці бази даних зберігають усі дані, доступні користувачам. У реляційній моделі даних таблиця має такі основні властивості:

– ідентифікується унікальним ім’ям;

– має кінцеву (як правило, постійну) ненульову кількість стовпців;

– має кінцеве (можливо, нульове) число рядків;

– стовпці таблиці ідентифікуються своїми унікальними іменами й номерами;

– вміст усіх комірок стовпця належить одному типу даних (тобто стовпці однорідні);

– рядки таблиці не мають будь-якої упорядкованості й ідентифікуються тільки своїм вмістом (тобто поняття «номер рядка» не визначене).

У загальному випадку комірки таблиці можуть залишатися «порожніми» (тобто не містити будь-якого значення), такий їхній стан позначається як NULL.

Представлення (VIEW) – це налагоджене представлення даних з однієї або декількох таблиць. Представлення можна розглядати як «запит, що зберігається». Представлення не містять даних; замість цього вони отримують дані з тих таблиць, на основі яких вони створені.

Послідовність (SEQUENCE) – генерує унікальні порядкові номери, які можуть використовуватися як значення числових стовпців таблиць бази даних. Послідовності спрощують прикладне програмування, автоматично генеруючи унікальні числові значення для рядків однієї або декількох таблиць. Номери, що генеруються послідовністю, незалежні від таблиць, так що одну і ту ж послідовність можна використовувати для декількох таблиць. Після її створення до послідовності можуть звертатися різні користувачі, щоб отримувати дійсні порядкові номери.

Процедура (PROCEDURE) або **функція (FUNCTION)** – це сукупність команд SQL і PL/SQL, згрупованих разом як програмна одиниця, що виконує специфічне завдання. Процедури і функції поєднують легкість і гнучкість SQL з процедурними можливостями мови структурного програмування PL/SQL. За допомогою PL/SQL такі процедури і функції можна визначати і зберігати в базі даних для тривалого використання.

Процедури і функції подібні між собою, з тією різницею, що функція завжди повертає програмі єдине значення, тоді як процедура в загальному випадку не повертає значення, проте існують специфічні методики для повернення значення процедури.

Пакети (PACKAGE) дають метод інкапсулювання і зберігання взаємозв'язаних процедур, функцій та інших конструктів пакету як одиниці в базі даних. Надаючи адміністраторові бази даних або розробникові додатків організаційні переваги, пакети, водночас розширюють функціональні можливості і збільшують продуктивність бази даних.

Синонім (Synonym) – це аліас (додаткове ім'я) для таблиці, представлень, послідовності або програмної одиниці. Синонім не є об'єктом, він – пряме посиланням на об'єкт.

Синоніми використовуються для:

– маскування дійсного імені і власника об'єкта;

– забезпечення загального доступу до об'єкта;

– досягнення прозорості місця розташування для таблиць, представлень або програмних одиниць віддаленої бази даних.

Індекси, кластери – це необов'язкові структури, що асоціюються з таблицями, які можна створювати для підвищення продуктивності операцій отримання даних. Так само, як індекс у книзі дозволяє вам швидше відшукувати потрібну інформацію, індекс ORACLE надає швидкий доступ до даних таблиці. Індекси корисні, коли додатки часто опитують інтервали рядків таблиці або окремі рядки.

Кластери (CLASTER) надають необов'язковий спосіб зберігання даних таблиць. Кластер – це група з однієї або декількох таблиць, що фізично зберігаються разом.

Транзакція (TRANSACTION) – модуль фіксації, логічний модуль, який складається з набору змін (вставок, оновлень і видалень). Трансакції або мають бути збережені в базі даних, або повинен бути виконаний їх відкат. Фіксуються або всі зміни в трансакції, або жодна з них. Для фіксації в базі даних результатів трансакції використовується команда COMMIT, для відновлення змінених даних – команда ROLLBACK.

Трансакція починається з моменту підключення користувача до бази даних і починає з нею працювати.

Трансакція завершується після виконання оператора COMMIT та ROLLBACK.

1.4.5. Структури оперативної пам'яті екземпляра БД Oracle

Як видно з рис. 1.28, системна глобальна область (**SGA**) складається з декількох областей. Основні структури пам'яті, пов'язані з екземпляром Oracle, такі:

– **Системна глобальна область (SYSTEM GLOBAL AREA – SGA)**, що розподіляється усіма серверними і фоновими процесами; **SGA** – область, у якій містяться дані і управляюча інформація екземпляра.

– **Програмна глобальна область (Program Global Area-PGA).** Окрема для кожного серверного і фонового процесу область пам'яті.

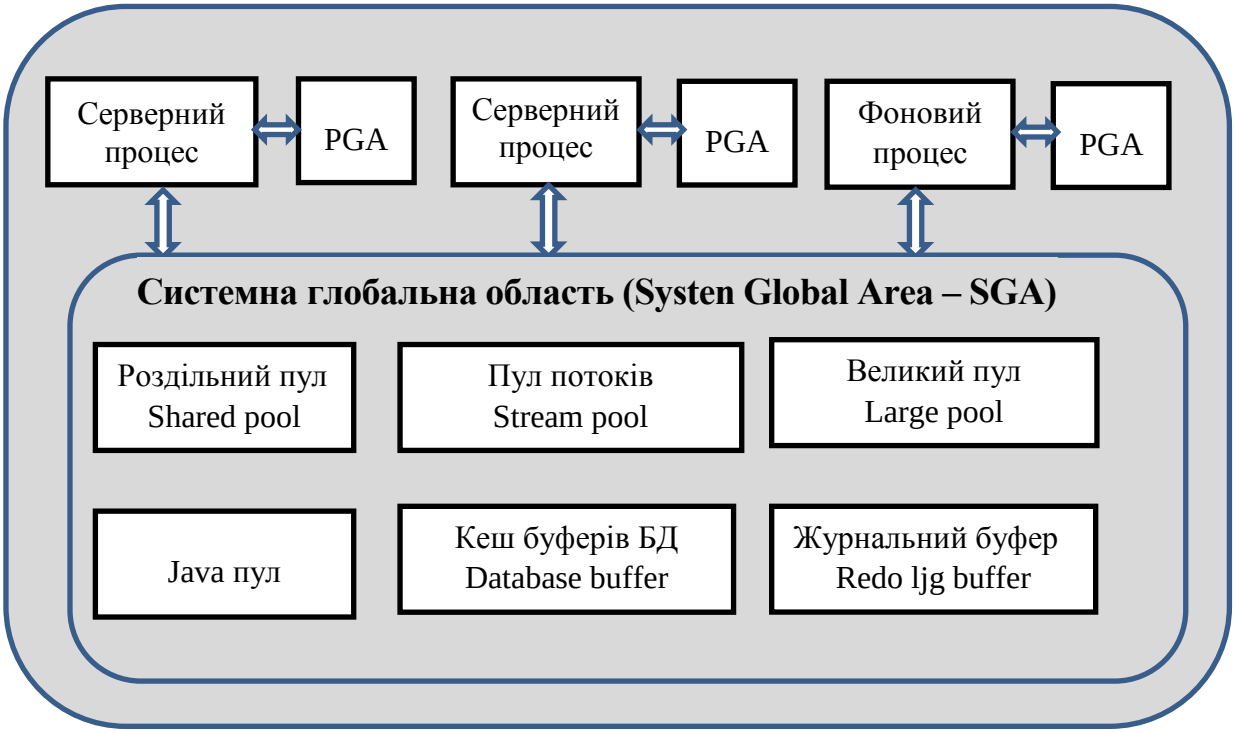


Рис. 1.28. Структури оперативної пам'яті екземпляра БД Oracle

– Кеш буферів БД (Database buffer cache) пам'ять для блоків даних, що вибираються із БД.

– **Журнальний буфер (Redo log buffer)** для кешування інформації повторного виконання (що використовується при відновлені екземпляра) до моменту їх запису в журнальні файли.

– **Роздільний пул (Shared pool)** для кешування різноманітних структур, що використовуються користувачами разом.

– **Великий пул (Larg pool) –** необов'язкова область, у якій відводиться пам'ять для буферів, що вимагають великих операцій вводу/виводу.

– **Java pool** використовується для Java сеансів і даних віртуальної машини Java Virtual Machine – JVM.

– **Пул потоків (Stream pool)** використовується Oracle Streams.

1.4.6. Фонові процеси екземпляра сервера Oracle

Найчастіше використовують фонові процеси, показані на рис. 1.26, хоча їхній склад змінюється від версії до версії. Нижче описані деякі процеси.

Процес запису в базу даних (Database Writer, DBWn). Записує блоки даних з кеша буферів SGA в файли даних, розташовані на диску. Для одного примірника Oracle може існувати до 20 процесів DBW, обробних введення / виведення в різні файли даних, звідси і позначення DBW/n.

Процес запису в журнал (Log Writer, LGWR) записує інформацію з журнального буфера в SGA в усі копії поточного журнального файлу на диску. Поки транзакція обробляється, необхідна для її повторного виконання інформація зберігається у журнальному буфері в SGA. Після фіксації транзакції Oracle відправляє цю інформацію на постійне зберігання, викликаючи процес LGWR для її запису на диск.

Системний монітор (System Monitor, SMON). Відповідає за загальну справність і безпеку примірника. SMON відновлює екземпляр під час запуску після збою. Він координує доступ і реалізує відновлення екземпляра, коли до бази даних звертаються відразу кілька примірників, як буває при роботі з Real Application Clusters. SMON також упорядковує суміжні області вільного простору в файлах даних, об’єднуючи їх, і позбавляється від простору, використовуваного для сортування рядків, коли в ньому вже немає необхідності.

Монітор процесів (Process Monitor, PMON). Стежить за призначеними для користувача процесами, що звертаються до бази даних. У разі аварійного припинення, призначеного для користувача про пам’ять і процеси примірника процесу PMON, відповідає за очистку залишків зайнятих ресурсів (таких як оперативна пам’ять) і за зняття всіх блокувань, встановлених збійним процесом.

Архіватор (Archiver, ARCH). Читає заповнені файли журналу і копіює їх в один або кілька архівних журналів. Підтримується до 10 процесів архівації, яким присвоюються імена ARC0-ARC9. За необхідності LGWR запускає додаткові архіватори

залежно від навантаження. Максимально дозволена кількість процесів архівації задається параметром ініціалізації LOG_ARCHIVE_MAX_PROCESSES. За замовчуванням цей параметр дорівнює 2, необхідність змінювати це значення виникає рідко.

Контрольна точка (Checkpoint, CKPT). Оновлює заголовки файлів даних після установки контрольної точки.

Процес відновлення (Recover, RECO). Автоматично очищає невдалі і відкладені розподілені трансакції.

Диспетчер (**Dispatcher**).

Ці необов'язкові фонові процеси запускаються, якщо розгорнено роздільний сервер.

Служба глобального кешу (**Global Cache Service, LMS**).

Управляє ресурсами, необхідними Real Application Clusters, а також координує використання ресурсів декількома екземплярами.

Черга завдань (**Job Queue**).

Служба виконання команд і процедур PL / SQL за заданим розкладом.

Монітор черг (**Queue Monitor, QMNn**).

Здійснює моніторинг черг повідомлень підсистеми Oracle Streams. Підтримується до 10 таких моніторів.

1.4.7. Утиліти для написання запитів до БД ORACLE

Для написання запитів до ORACLE (використовуючи SQL) необхідно визначитися з робочим інструментом, який буде використовуватися для написання запитів до бази даних Oracle. На відміну, наприклад, від Microsoft SQL Server, з яким поставляються дуже потужні засоби для написання запитів (Query Analyzer, SQL Server Management Studio), з Oracle поставляється лише SQL*Plus, який важко назвати зручним і повнофункціональним рішенням. Звичайно, знати SQL*Plus потрібно, і в багатьох ситуаціях без нього не обійтися, однак складні запити зручніше створювати за допомогою інших програм.

Найбільш часто використовуються такі програмні засоби:

– **Quest Software** фірми Toad;

– **SQL Navigator** тієї ж фірми;

– **PL/SQL Developer** фірми Allround Automations;

– **Oracle SQL Developer** виробництва Oracle.

Вважається, що Toad більш зручний для адміністраторів, а PL/SQL Developer – для розробників, які створюють складні пакети на PL/SQL. Однак функціональні можливості всіх перерахованих програмних засобів схожі і орієнтуватися потрібно на той програмний продукт, який прийнято використовувати на вашому підприємстві (або до якого ви самі більше звикли).

У цьому списку можна відзначити Oracle SQL Developer, у якого є ряд переваг:

– він офіційно безкоштовний;

– він розроблений копорацією Oracle, тому, зазвичай, з попаданням у список затверджених програмних продуктів на підприємстві проблем не виникає;

– він написаний на мові Java, тому один і той же графічний інтерфейс можна використовувати як з-під Windows, так і під Unix.

Його можна використовувати також для написання й налагодження запитів до інших СУБД, наприклад, Microsoft SQL Server і Microsoft Access, тому з’являється можливість використовувати єдине середовище для написання запитів до різних баз даних підприємства.

Питання для самоконтролю до теми 1.4

1. Опишіть клієнт-серверну інфраструктуру інформаційної системи податкового відомства, призначення її елементів.

2. Опишіть основні властивості реляційної моделі даних.

3. Опишіть основні вимоги до баз даних та СКБД.

4. Опишіть переваги клієнт-серверної технології.

5. Назвіть основні логічні та фізичні об’єкти бази даних Oracle.

6. Опишіть архітектуру сервера Oracle. Які основні компоненти сервера Oracle?

7. Екземпляр бази даних Oracle. Основні компоненти та функції;

8. База даних Oracle. Складові та їх призначення.

9. Файли даних бази даних Oracle. Поняття оптимізації БД.

10. Логічні структури файлу бази даних Oracle.

11. Логічні структури бази даних Oracle. Табличні простори і схеми.

12. Основні функції табличних просторів SYSTEM, SYSAUX, TEMP.

13. Основні функції табличних просторів UNDOTBS1 EXAMPLE.

14. Опишіть основні об’єкти та логічні структури зберігання і доступу до даних БД Oracle: представлення (View); послідовність (Sequence).

15. Основні компоненти програм на PL/SQL: процедура (Procedure); функція (Function); пакети (Package).

16. Поняття транзакції та команд COMMIT, ROLLBACK.

17. Опишіть структури оперативної пам’яті екземпляра БД Oracle.

18. Опишіть фонові процеси екземпляра сервера Oracle.

19. Особливості процесу запису в базу даних (Database Writer, DBWn).

20. Особливості процесу запису в журнал (Log Writer, LGWR).

21. Опишіть основні утиліти для написання запитів до БД ORACLE. Основні функціональні можливості Oracle SQL Developer.

Лабораторний практикум 1

Лабораторна робота 1.1
Оцінювання результативності підрозділів податкового відомства на основі інтегральної згортки показників

Мета: отримання практичних навичок використання методології інтегральної згортки показників для оцінювання результативності діяльності організації.

Завдання роботи

1. Провести аналіз результативності діяльності територіальних органів податкового відомства відповідно до варіанта (як варіант може бути місце народження або проживання).

2. Визначити показники, що покращують/погіршують результативність територіального підрозділу податкового відомства.

3. Сформувати висновки та пропозиції щодо покращання результативності територіального підрозділу податкового відомства. Підготувати звіт і розмістити на СДН за адресою: http://moodle.nusta.edu.ua/ .

Порядок виконання лабораторної роботи

Завдання 1. Аналіз результативності діяльності територіальних органів податкового відомства відповідно до варіанта.

1. Ознайомитися із теоретичним матеріалом параграфа 1.3.2.

2. Завантажити у розділі курсу на http://moodle.nusta.edu.ua файл Excel. Відкрити файл і ознайомитись із листами файлу (п.4-6).

3. «1.Зведтабл_відн_показників» – аналіз фактичних якісних показників, на основі яких розраховують усі інші показники та інтегральні критерії.

4. «2.Зведтабл_у_розрізі_податків» – зведена таблиця для вибору показників виконання завдання у розрізі податків.

5. «3.Звед_табл_інтегральних_кр» – виведення значень інтегральних критеріїв та критеріїв, на основі яких вони розраховані.

6. **Зверніть увагу!** У наданому файлі наведений приклад розрахунку інтегральних критеріїв за даними ДПС України 2009–2010 років.

7. Ознайомитися із інтегральними критеріями та показниками діяльності територіальних підрозділів податкового відомства на листах книги Excel відповідно до варіанта (рис. 1.29).

8. Для виведення на екран даних органів ДПС, вказаного у варіанті регіону, необхідно вибирати у закладці «Назва регіону» – відповідно потрібний регіон. При цьому у закладках «Номер-рег» і «ДПІ» повинно бути значення «Все».

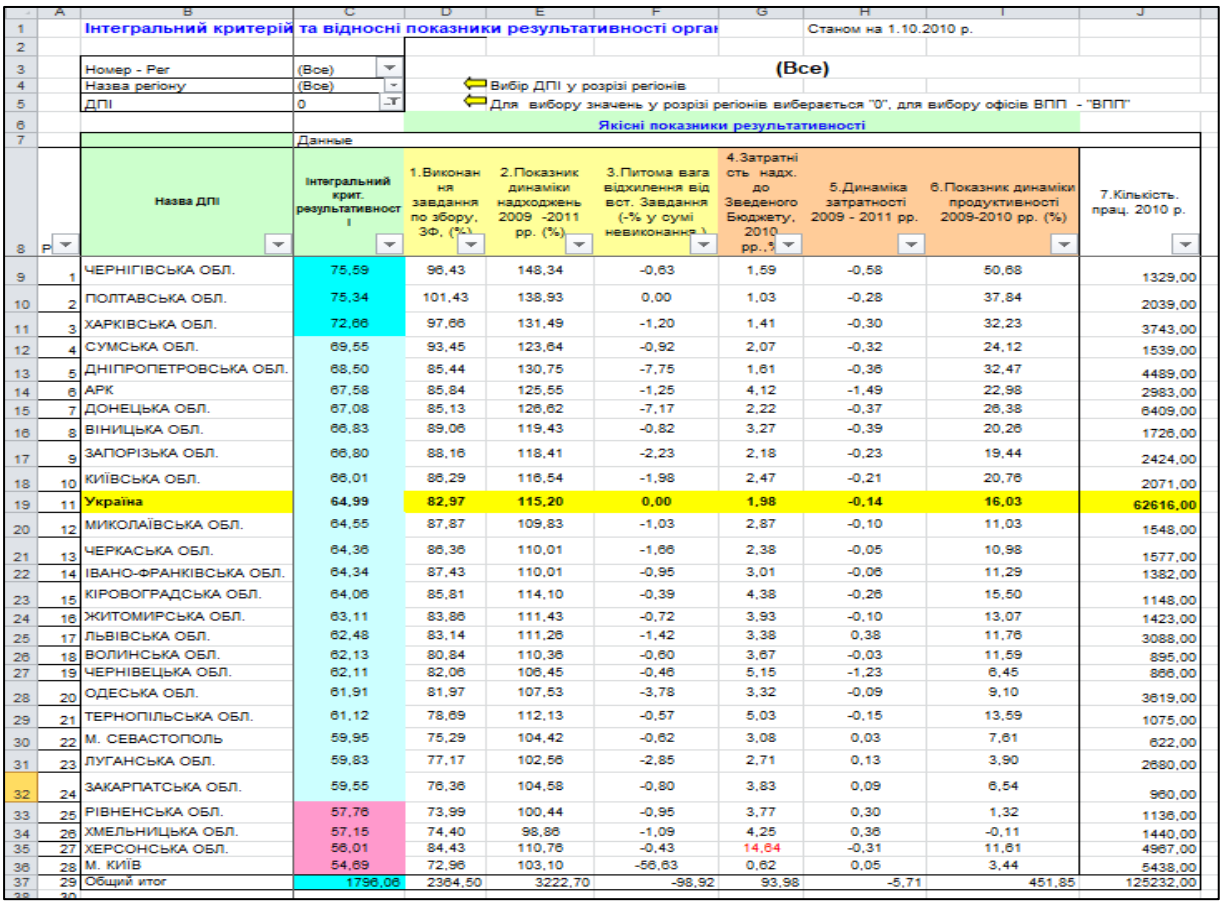
9. Для виведення на екран даних у розрізі регіонів у закладках «Номер-рег» і «Назва регіону» повинно бути значення «Все», а у закладці «ДПІ» – «0». («0» – ознака ДПА в області).

10. Для виведення на екран даних по офісах ВПП у закладках «Номер-рег» і «Назва регіону» повинно бути значення «Все», а у закладці «ДПІ» – «ВПП».

11. Інтегральний критерій результативності органу ДПС відображається у крайньому лівому стовпці. Для сортування їхніх значень по убуванню – поставте маркер миші у колонку «Інтегральний крит. результативності» і виберіть у панелі інструментів (А/Я) – «Сортировка по убыванию».

12. Виберіть лист «1.Звед_табл_відн_показників». Виведіть дані по регіону, вказаному у варіанті. У відсортованій зведеній таблиці на листі визначте рейтинг вказаного у варіанті підрозділу, фактичне значення інтегрального показника, колір, яким він відображається, та групу, до якої він повинен належати за рівнем результативності (низька, помірна, добра, висока).

13. Проаналізуйте показники регіонального підрозділу за рівнем результативності порівняно з обласним та центральним рівнями?



Інтегральний критерій та відносні показники результативності орган — Станом на 1.10.2010 р.

Номер - Рег: (Все)
Назва регіону: (Все)
ДПІ: 0

(Все)

⇦ Вибір ДПІ у розрізі регіонів
⇦ Для вибору значень у розрізі регіонів вибирається "0", для вибору офісів ВПП - "ВПП"

Якісні показники результативності

Данные

Р	Назва ДПІ	Інтегральний крит. результативності	1.Виконання завдання по збору, ЗФ, (%)	2.Показник динаміки надходжень 2009 -2011 рр. (%)	3.Питома вага відхилення від вст. Завдання (-% у сумі невиконання.)	4.Затратність надх. до Зведеного Бюджету, 2010 рр.,%	5.Динаміка затратності 2009 - 2011 рр.	6.Показник динаміки продуктивності 2009-2010 рр. (%)	7.Кількість. прац. 2010 р.
1	ЧЕРНІГІВСЬКА ОБЛ.	75,59	96,43	148,34	-0,63	1,59	-0,58	50,68	1329,00
2	ПОЛТАВСЬКА ОБЛ.	75,34	101,43	138,93	0,00	1,03	-0,28	37,84	2039,00
3	ХАРКІВСЬКА ОБЛ.	72,66	97,66	131,49	-1,20	1,41	-0,30	32,23	3743,00
4	СУМСЬКА ОБЛ.	69,55	93,45	123,64	-0,92	2,07	-0,32	24,12	1539,00
5	ДНІПРОПЕТРОВСЬКА ОБЛ.	68,50	85,44	130,75	-7,75	1,61	-0,36	32,47	4489,00
6	АРК	67,58	85,84	125,55	-1,25	4,12	-1,49	22,98	2983,00
7	ДОНЕЦЬКА ОБЛ.	67,08	85,13	126,62	-7,17	2,22	-0,37	26,38	6409,00
8	ВІННИЦЬКА ОБЛ.	66,83	89,06	119,43	-0,82	3,27	-0,39	20,26	1726,00
9	ЗАПОРІЗЬКА ОБЛ.	66,80	88,16	118,41	-2,23	2,18	-0,23	19,44	2424,00
10	КИЇВСЬКА ОБЛ.	66,01	86,29	116,54	-1,98	2,47	-0,21	20,76	2071,00
11	**Україна**	**64,99**	**82,97**	**115,20**	**0,00**	**1,98**	**-0,14**	**16,03**	**62616,00**
12	МИКОЛАЇВСЬКА ОБЛ.	64,55	87,87	109,83	-1,03	2,87	-0,10	11,03	1548,00
13	ЧЕРКАСЬКА ОБЛ.	64,36	86,36	110,01	-1,66	2,38	-0,05	10,98	1577,00
14	ІВАНО-ФРАНКІВСЬКА ОБЛ.	64,34	87,43	110,01	-0,95	3,01	-0,06	11,29	1382,00
15	КІРОВОГРАДСЬКА ОБЛ.	64,06	85,81	114,10	-0,39	4,38	-0,26	15,50	1148,00
16	ЖИТОМИРСЬКА ОБЛ.	63,11	83,86	111,43	-0,72	3,93	-0,10	13,07	1423,00
17	ЛЬВІВСЬКА ОБЛ.	62,48	83,14	111,26	-1,42	3,38	0,38	11,76	3088,00
18	ВОЛИНСЬКА ОБЛ.	62,13	80,84	110,36	-0,60	3,67	-0,03	11,59	895,00
19	ЧЕРНІВЕЦЬКА ОБЛ.	62,11	82,06	106,45	-0,46	5,15	-1,23	6,45	866,00
20	ОДЕСЬКА ОБЛ.	61,91	81,97	107,53	-3,78	3,32	-0,09	9,10	3619,00
21	ТЕРНОПІЛЬСЬКА ОБЛ.	61,12	78,69	112,13	-0,57	5,03	-0,15	13,59	1075,00
22	М. СЕВАСТОПОЛЬ	59,95	75,29	104,42	-0,62	3,08	0,03	7,61	622,00
23	ЛУГАНСЬКА ОБЛ.	59,83	77,17	102,56	-2,85	2,71	0,13	3,90	2680,00
24	ЗАКАРПАТСЬКА ОБЛ.	59,55	76,36	104,58	-0,80	3,83	0,09	6,54	960,00
25	РІВНЕНСЬКА ОБЛ.	57,76	73,99	100,44	-0,95	3,77	0,30	1,32	1136,00
26	ХМЕЛЬНИЦЬКА ОБЛ.	57,15	74,40	98,86	-1,09	4,25	0,36	-0,11	1440,00
27	ХЕРСОНСЬКА ОБЛ.	56,01	84,43	110,76	-0,43	14,64	-0,31	11,61	4967,00
28	М. КИЇВ	54,69	72,96	103,10	-56,63	0,62	0,05	3,44	5438,00
29	Общий итог	1796,06	2364,50	3222,70	-98,92	93,98	-5,71	451,85	125232,00

Рис. 1.29. Книга Excel з розрахованими критеріями результативності територіальних органів ДПС України

Завдання 2. Визначення показників, що понижують/підвищують результативність територіальних підрозділів.

1. Визначте показники, що покращують або понижують показник результативності, а для цього проаналізуйте в їхніх листах «1.Звед_табл_відн_показників», та 2.Звед_табл_у_розрізі_-податків».

2. Поясніть можливі причини низьких/високих значень показників, що вплинули на загальну результативність територіального підрозділу ДПС України.

Завдання 3. Формування висновків та пропозицій

3.1. Сформувати висновки та пропозиції щодо покращання результативності територіального підрозділу податкового відомства.

3.2. Підготувати звіт і розмістити на СДН за адресою: http://moodle.nusta.edu.ua/.

Лабораторна робота 1.2
Оцінка ймовірнісних показників податкових ризиків

Мета: придбання навичок у оцінюванні ймовірності, як показника податкового ризику.

Вихідні дані для роботи: Нормально розподілена випадкова величина податкових втрат X має математичне сподівання M(X) = a та середньоквадратичне відхилення σ(X) = σ (a, σ задані за варіантами).

Завдання роботи

1. Записати вигляд диференціальної функції розподілу $f(x)$.
2. У середовищі Excel побудувати графік диференціальної функції розподілу податкових втрат $f(x)$.
3. У середовищі Excel побудувати графік інтегральної функції розподілу податкових втрат $F(x)$.
4. Знайти ймовірність того, що внаслідок реалізації рішення випадкова величина податкових втрат x набуде значення в інтервалі (s, t) при (a-3σ, a+3σ), тобто знайти P(s<X<t) і P(a-3σ<X<a+ 3σ).
5. Сформуйте висновки до отриманих результатів. Підготуйте звіт про виконання лабораторної робити і надішліть для оцінювання в середовище СДН.

Порядок виконання лабораторної роботи

Завдання 1. Записати вигляд диференціальної функції розподілу податкових втрат $f(x)$.

Розглянемо рішення задачі на прикладі, у якому математичне сподівання $M(X) = 30$ та середньоквадратичне відхилення $\sigma(X) = 10$.

Згідно з умовою задачі $a = 30$, $\sigma = 10$, тому відповідно

$$f(x) = \frac{1}{10\sqrt{2\pi}} \cdot \int_{\alpha}^{\beta} e^{-\frac{(t-30)^2}{2\cdot 10^2}} dt = \frac{1}{10\sqrt{2\pi}} \cdot \int_{\alpha}^{\beta} e^{-\frac{(t-30)^2}{200}} dt.$$

Завдання 2. У середовищі Excel побудувати графік диференціальної функції розподілу податкових втрат f (x).

Для побудови графіка функції $f(x)$ знайдемо значення $f(x)$ у точках $x = 5, 10, \ldots, 50$, використовуючи функцію пакета Excel, яка має вигляд:

НОРМРАСП(x; Среднее; Стандартное_откл; Интегральная),

де ***x***-значення, для якого будується розподіл;

Среднее-математичне сподівання *M(X);*

Стандартное_откл – середнє квадратичне відхилення **σ**(*X*);

Интегральный-логічне значення 0 або 1. (Якщо **«Интегральный»=0,** отримаємо значення диференціальної функції розподілу $f(x)$ в точці ***x***, якщо **«Интегральный»=1** – функція повертає значення інтегральної функції розподілу $F(x)$ в точці *x).*

Завдання 3. У середовищі Excel побудувати графік *інтегральної функції* розподілу величини податкових втрат $F(x)$.

Для побудови графіка функції $F(x)$ знайдемо значення $F(x)$ у точках $x = 5, 10, \ldots, 50$, використовуючи функцію **НОРМРАСП (x;** *M(X)***; σ**(*X*)**; 1).**

Завдання 4. Знайти ймовірність того, що в наслідок реалізації рішень випадкова величина негативних втрат набуде значення в інтервалі (10,50), $(a - 3\sigma, a + 3\sigma)$.

Використовуючи формулу (1.55), одержимо:

$$P(10 < X < 50) = \int_{10}^{50} f(t)dt = \Phi\left(\frac{50-30}{10}\right) - \Phi\left(\frac{10-30}{10}\right) = 2\Phi(2) = 0{,}9544\ ,$$

де $\Phi(2)$ знаходимо за таблицею значень функції Лапласа.

Цю ймовірність можна обчислити, якщо скористатися функцією «НОРМРАСП», використовуючи вираз «НОРМРАСП (50;30;10;1)–НОРМРАСП(10;30;10;1)» і дорівнює 0,9545.

Результат застосування функції **«НОРМРАСП»** та графіки функції $f(x)$ і $F(x)$ наведено на рис. 1.30.

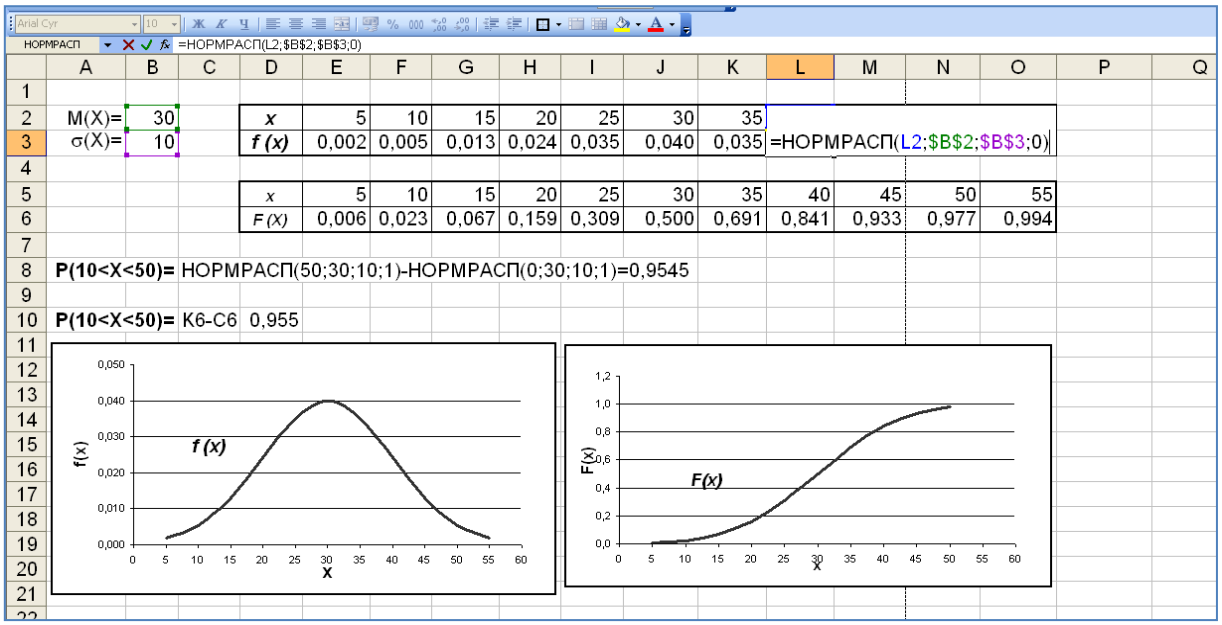


Рис. 1.30. Диференціальна та інтегральна функції випадкової величини

Для інтервалу (a – 3σ, a + 3σ)=(0,60) шукана ймовірність обчислюється аналогічно за допомогою виразу =НОРМРАСП (60;30;10;1)–НОРМРАСП(0;30;10;1) і дорівнює 0,9973, що збігається з результатами, одержаними за формулою (1.53),

$$P\left(\frac{\pi}{6} < x < \frac{\pi}{2}\right) = \frac{1}{2}\int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \sin t \cdot dt = \frac{1}{2}(-\cos t)\Big|_{\frac{\pi}{6}}^{\frac{\pi}{2}} = -\frac{1}{2}\left(0 - \frac{\sqrt{3}}{2}\right) = \frac{\sqrt{3}}{4}$$

За формулою будемо мати:

$$P\left(\frac{\pi}{6} < x < \frac{\pi}{2}\right) = F\left(\frac{\pi}{2}\right) - F\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{4}.$$

Завдання 5. Звіт про виконання роботи повинен вміщувати змістовні висновки, що описують отримані результати, рекомендації щодо прийняття рішень при оцінюванні і врахуванні податкових ризиків.

Лабораторна робота 1.3
Організація робочого простору на хмарному ресурсі Oracle Application Express (APEX)

Мета: використання хмарного ресурсу APEX apex.oracle.com для організації самостійної роботи, отримання навичок у формуванні аналітичних запитів до баз даних Oracle 10-11g (XE) та розробці Web орієнтованих додатків на платформі Oracle.

Завдання роботи: створити персональний обліковий запис на хмарному ресурсі *apex.oracle.com*, завантажити дані схеми HR та відпрацювати базові запити до схеми HR на мові SQL.

Порядок виконання лабораторної роботи

1. Реєстрація на сайті Oracle APEX.

1.1. У вікні браузера введіть http://apex.oracle.com, виберіть Sign Up.

Студенти та викладачі Oracle Academy можуть увійти до Oracle Application Express з реєстраційними даними, які надаються інструктором Oracle Academy.

1.2. Виберіть *Get Started*, в поле *First Name і Last Name* введіть своє ПІБ, в поле Email введіть адресу електронної пошти, в полі Workspace введіть ім'я робочого простору, краще використовувати ваші дані, наприклад, sokolenko, натисніть Next, в поле New Schema to .. введіть ім'я схеми бази даних, наприклад можна також sokolenko, задайте пароль, Next, далі …

1.3. Дайте відповідь на питання анкети, виберіть *I accept the terms*, виберіть *Next.*

1.4. На вказану поштову адресу прийде лист *Approved: account request for*, відкрийте його. Структура листа буде такою:

Your request for an account has been approved. Workspace: SOKOLENKO User ID: *mi@gmail.com*.

Please click on the link: http://apex.oracle.com/pls/apex/f? p = 4700: 73 ::::: PROVISION_ID, E_ID: 62549341213783696278, *mi@gmail.com* to complete the approval process and receive your credentials.

1.5. Виберіть посилання з листа, щоб закінчити створення робочого простору. На вказану поштову адресу прийде лист Approved: account request for, відкрийте його. Структура листа буде такою:

Your request for an account has been approved. Workspace: SOKOLENKO User ID: *mi@gmail.com* Password: (вказаний пароль) Go to http://apex.oracle.com/pls/apex/ to get started.

1.6. Виберіть посилання з листа і увійдіть у робочий простір, ввівши отримані облікові дані (рис. 1.32): 1) назва простору; 2) ім'я користувача; 3) пароль користувача.

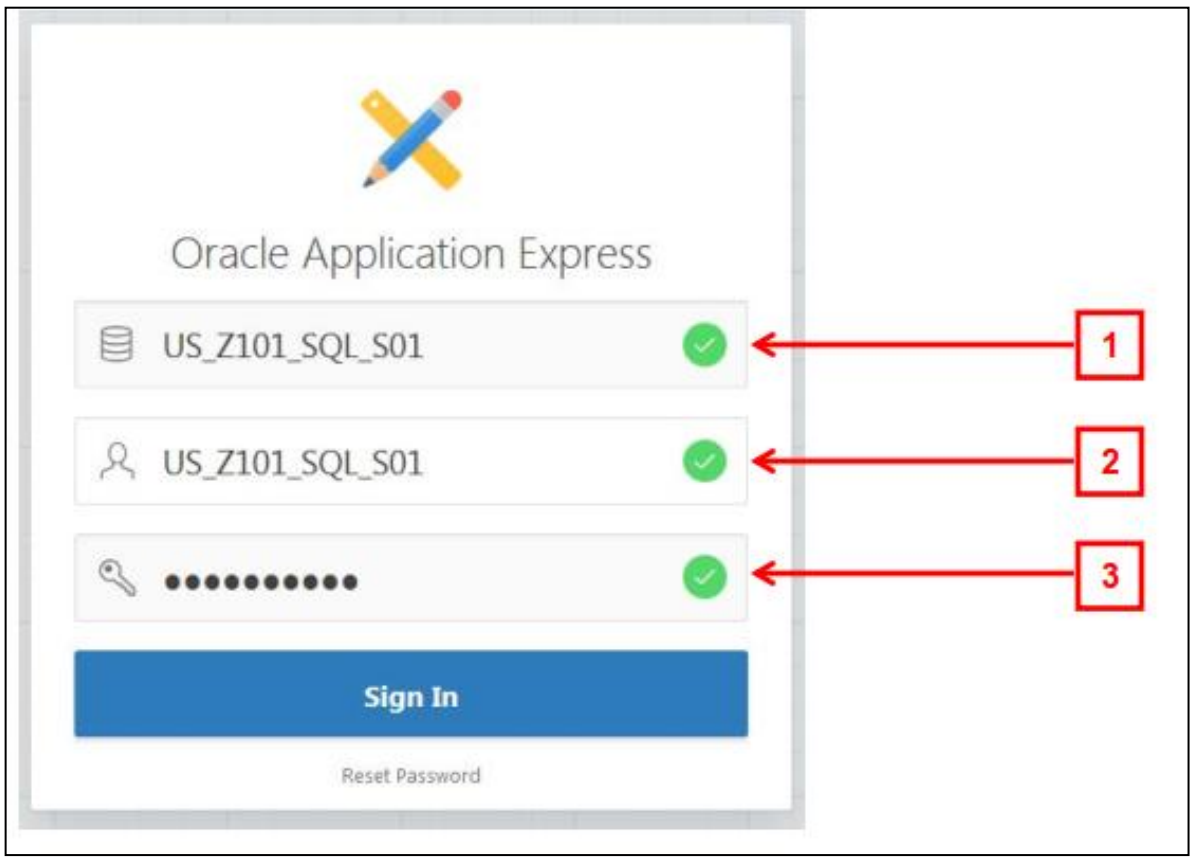


Рис. 1.31. Вікно введення облікових даних для входу до APEX

2. Знайомство із компонентами Apex. Завантаження даних та включення навчальної схеми HR.

2.1. Після входу в Oracle Application Express побачите домашню сторінку Oracle Application Express. Ця сторінка відображає усі компоненти Oracle Application Express: Application Builder, SQL Workshop Team Development та упаковані програми (рис. 1.32).

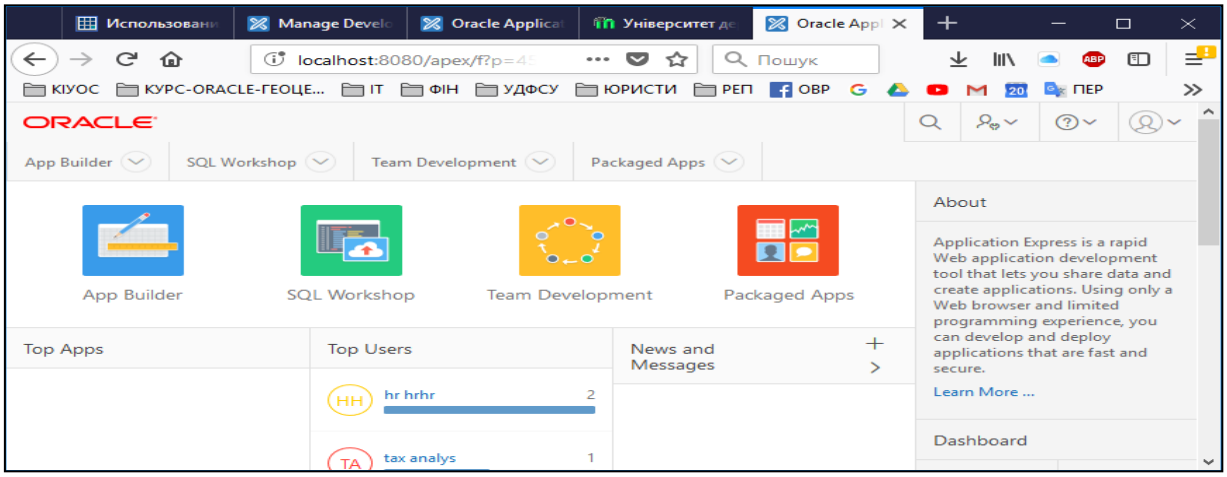


Рис. 1.32. Домашня сторінка Oracle Application Express (APEX)

– **Application Builder:** дозволяє створювати, переглядати або контролювати програми.

– **SQL Workshop**: дозволяє створювати, керувати та переглядати об’єкти бази даних з веб-браузера за допомогою SQL.

– **Team development:** розробка команди: полегшує управління процесом розробки додатків.

– **Packaged Apps**: набір додатків для підвищення продуктивності розробок групою розробників.

SQL Workshop є основним компонентом, який використовують для програмування баз даних з програмою курсу Oracle Academy: DB Designer, SQL, PL / SQL.

Вкладки у верхній частині сторінки забезпечують швидкий доступ до цих компонентів.

2.2. Додавання таблиць і даних до облікових записів APEX

Для того щоб мати доступ до таблиць і даних, що використовують протягом усього курсу, файл сценарію, який можна отримати з Oracle iLearning, повинен бути запущений в обліковому записі всіх облікових записів студентів. Якщо ви є слухачем Oracle Academy, файл сценарію можна завантажити з Oracle iLearning для чого:

2.2.1. Перейдіть у розділ «Ресурси курсів», натисніть «Сценарії APEX та посібники користувача», натисніть «Сценарій для створення таблиць і дані для цього курсу», розархівуйте ZIP-файл і збережіть файл сценарію локально на вашому комп’ютері.

2.2.2. Увійдіть в обліковий запис на Apex і виберіть «SQL Workshop», а потім «SQL Scripts» (рис. 1.33). Виберіть «Upload» (рис. 1.34)

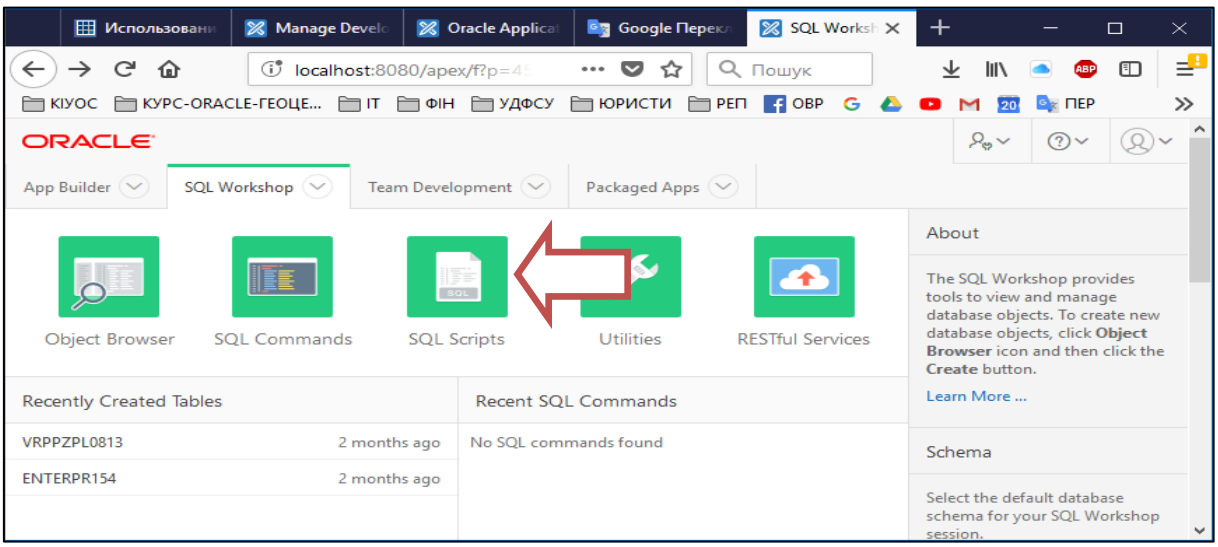


Рис. 1.33. Закладка «SQL Workshop»

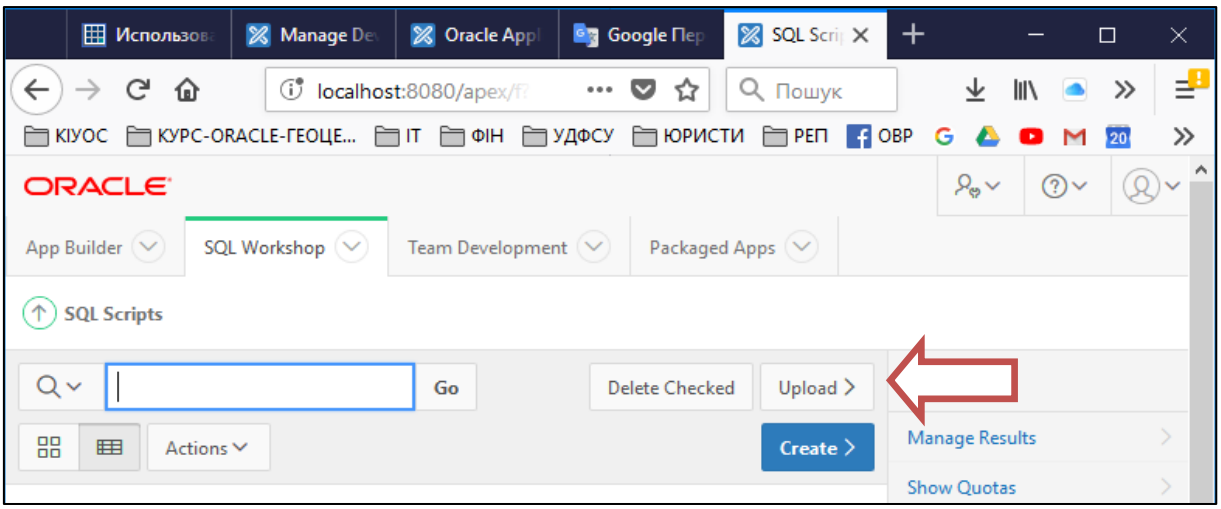


Рис. 1.34.Сторінка завантаження скриптів

2.2.3. Натисніть кнопку *«Browse»* і перейдіть до файлу, завантаженого з Oracle iLearning на кроці 2.2.1

2.2.4. Додайте ім'я сценарію – «SQL add all tables» або «PLSQL add all tables», залиште «File Character Set» як типовий, залиште «File Character Set» як типовий (Unicode UTF-8) і натисніть «Upload».

2.2.5. Натисніть піктограму «Run», після чого побачите список сценаріїв. Натисніть піктограму «Run now» (рис. 1.35).

2.2.6. Натисніть «View Results». На сторінці, що з'явиться, будете повідомлені про успіше (неуспішне) виконання скрипта.

2.3. Прогляд об'єктів бази даних Oracle в Apex.

2.3.1. Перейдіть в «SQL Workshop» і виберіть піктограму «Objects Browser» (рис. 1.36).

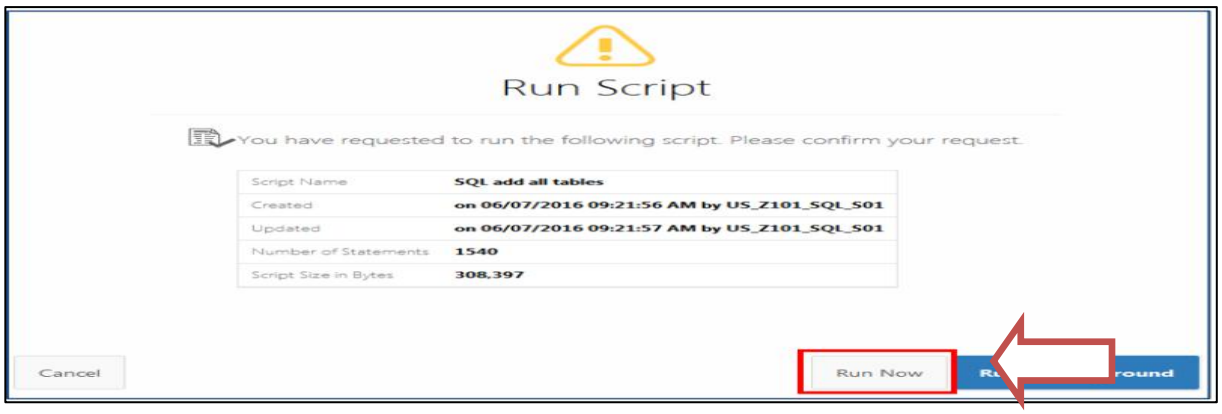


Рис. 1.35. Сторінка запуску скриптів

2.3.2. Ознайомтеся із об'єктами БД, що відображаються у лівому вікні. Виберіть типи об'єктів у меню зліва. Познайомтеся із даними таблиць для формування осмислених запитів до даних (рис. 1.36, 1.37).

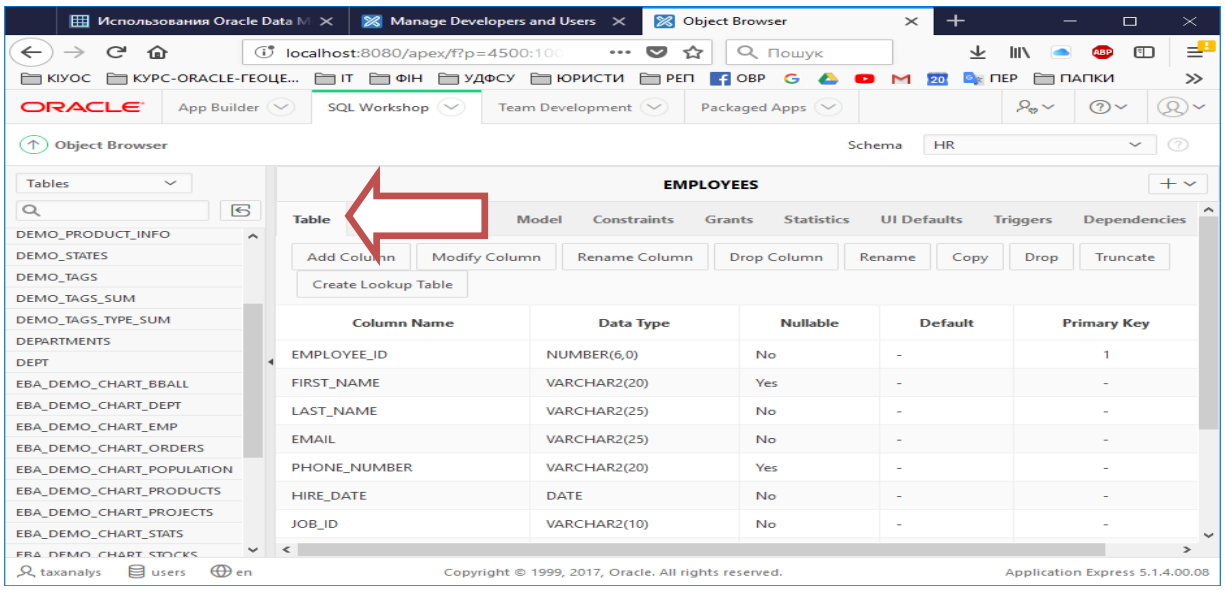


Рис. 1.36. Вікно «Table» відображення структури таблиці EMPLOYEES

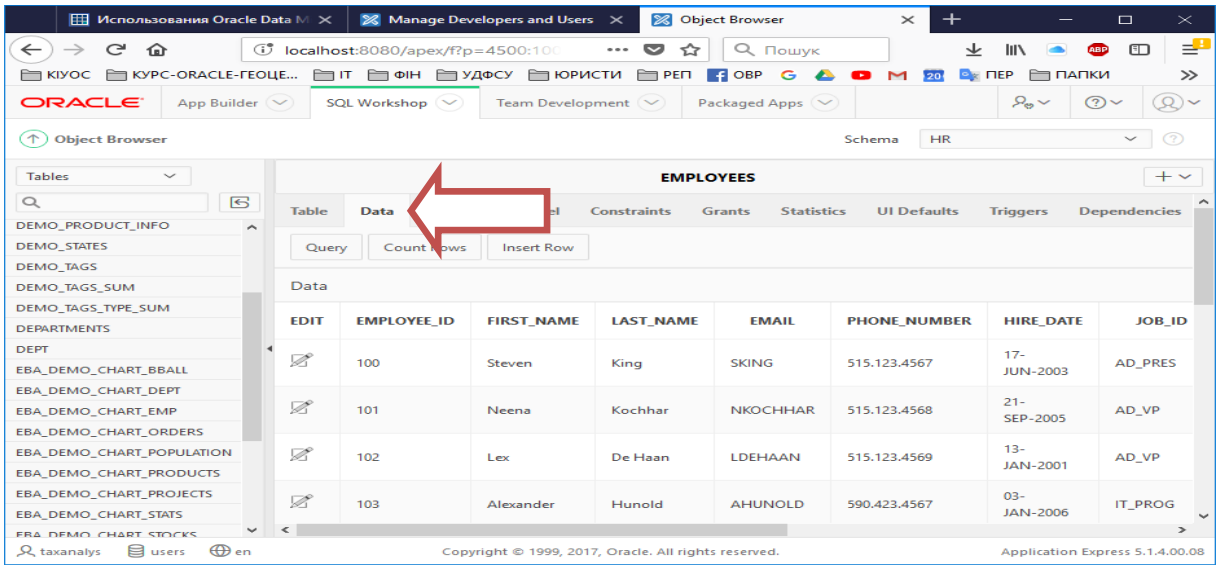


Рис. 1.37. Вікно «Data» відображення даних таблиці EMPLOYEES

2.3. Виконайте запити до БД Oracle в Apex.

2.3.1. Виконайте в середовищі APEX apex.oracle.com запити лабораторної роботи до таблиць схеми HR, вказаних у варіантах.

2.3.2. Підготуйте звіт про виконання самостійної роботи і розмістіть його на системі ДН http://moodle.nusta.edu.ua

РОЗДІЛ 2
ТЕХНОЛОГІЇ АНАЛІЗУ ПОДАТКОВОЇ ІНФОРМАЦІЇ. МОВА SQL ТА СКБД ORACLE

2.1. Пакети прикладних програм для аналізу інформації та технології доступу до баз даних

2.1.1. Огляд прикладних пакетів статистичного та інтелектуального аналізу

Сучасний фахівець з аналізу інформації повинен мати не тільки теоретичні знання, а й практичні навички використання інформаційних інструментів і ресурсів для вирішення професійних задач. Статистичний аналіз вхідної і вихідної інформації є важливою складовою процесу підвищення ефективності діяльності підприємств і організацій та відіграє основну роль при прийнятті управлінських рішень.

Як зазначалося вище, технології інтелектуального аналізу даних об'єднують ідеї із галузі комп'ютерних наук, створених на перетині штучного інтелекту, статистики та теорії баз даних, і позначаються як *KDD* або *DM (Knowledge Discovery in Database, Data Mining* – добування знань із баз даних (дейтамайнінг).

З огляду накопичених інформаційних ресурсів у будь-якій галузі діяльності людини все актуальнішою стає наука *Data Science* – наука про методи аналізу даних і вилучення з них цінної інформації, знань. Вона тісно перетинається з такими галузями, як машинне навчання (Machine Learning) і наукою про мислення (Cognitive Science) і, звичайно, технологіями роботи з великими даними (Big Data). Ці технології вже активно використовують у великому бізнесі, у банківській сфері, у соціальних мережах.

Крім MS Excel, який за своїми аналітичними функціями є достатнім для фахівця, що уміє ним користуватися, важливими інформаційними інструментами прийняття управлінських рішень є комп'ютерні статистичні пакети.

Статистичні пакети, як правило, мають модульну структуру. У базовому або додаткових модулях прикладних статистичних пакетів реалізовані основні процедури статистичного аналізу даних, зокрема:

– описова статистика (аналіз вибіркових даних);
– проста і множинна лінійна регресія;
– кореляційно-регресійний аналіз;
– аналіз часових рядів і прогнозування;
– дисперсійний аналіз;
– факторний аналіз;
– кластерний аналіз;
– методи контролю якості;
– обробка результатів анкетування;
– розвідувальний аналіз (data mining);
– нейромережеве прогнозування.

До найбільш поширених статистичних пакетів, що застосовують як у практиці бізнес-аналізу, так і в економічній освіті, відносять такі:

– SPSS (Statistical Package for the Social Sciences) – статистичний пакет для соціальних наук). Сайт: http://www.spss.com. 2009 року компанія SPSS куплена компанією IBM;

– статистичний пакет Statgraphics компанії Statpoint Technologies, Inc. Сайт: http://www.statgraphics.com. Є онлайн-версія пакета («хмарний сервіс»): Statgraphics Online;

– http://www.statgraphics.com/product/online.aspx);

– статистичний пакет Statistica. Сайт: http://www.statsoft.com. 2014 року компанія StatSoft куплена компанією Dell Software;

– засоби бізнес-аналітики SAS (Statistical Analysis System). Компанія SAS Institute, Inc (сайт: http://www.sas.com). Є версія SAS University Edition (http://www.sas.com/en_us/software/university-edition.html);

– безкоштовний статистичний пакет R (http://www.r-project.org). Призначений для проведення статистичних розрахунків та візуалізації;

– універсальна статистична діалогова система Stadia (http://statsoft.msu.ru/products.htm). Вітчизняний програмний продукт широко використовується у російських вузах;

– Python – мова програмування, яку можна розглядати як окремим продуктом з великим набором бібліотек інтелектуального аналізу, можливість створення завершених Web орієнтованих аналітичних систем.

Для ефективного застосування статистичних пакетів у підрозділах податкового відомства необхідно дотримуватися компетентнісної моделі підготовки фахівців аналітичних підрозділів. Зокрема, вони повинні:

– знати джерела податкової інформації, структуру даних для здійснення нерегламентованих запитів та баз даних, основні показники діяльності платників податків, основи оцінювання податкового навантаження галузі, регіону і економіки в цілому, критерії соціально-економічної ефективності суб'єктів господарювання, установ державного управління тощо;

– вміти обґрунтовувати актуальність, теоретичну і практичну значущість обраного методу, факторів та зрізу даних для побудови відповідних моделей;

– вміти представляти результати отриманих результатів моделювання у вигляді презентації, статті або доповіді;

– вміти здійснювати підготовку завдань і розробляти проектні завдання для виконуючих організацій та проектних груп;

– оцінювати ефективність виконання проектів з урахуванням фактора невизначеності;

– розробляти стратегії поведінки економічних агентів, суб'єктів податкових відносин у різних ситуаціях;

– розраховувати прогнозні показники фінансової діяльності підприємства, галузі, регіону й економіки в цілому;

– розробляти варіанти управлінських рішень і обґрунтовувати їх вибір на основі критеріїв ефективності;

– володіти здатністю удосконалювати і розвивати свій інтелектуальний та загальнокультурний рівень і самостійно освоювати нові методи дослідження.

Загальною тенденцією розвитку багатьох статистичних пакетів є доповнення функцій класичного статистичного аналізу можливостями інтелектуального аналізу даних (*Data Mining*), що дозволяє використовувати статистичні пакети як окремі елементи інформаційно-аналітичної системи установи (корпоративної системи підтримки прийняття рішень). Можливості статистичних пакетів роблять їх незамінним помічником економістів, маркетологів та інших фахівців, які використовують інструменти аналізу бізнес-інформації.

У розділі 3 розглядаються приклади застосування Statistica, Matlab, Deduktor, Oracle Data Miner у різних аспектах аналізу податкової інформації. Проте важливим у цьому контексті є також знання технологій доступу до баз даних ODBC, DAO, ADO, OLE DB, RDO, оскільки більшість статистичних пакетів передбачають під'єднання до баз даних на основі саме цих технологій.

2.1.2. Технології доступу до баз даних (ODBC, DAO, ADO, OLE DB, RDO)

Якщо програмні засоби, які пишуть дані і які їх читають, підтримують різні формати, то виникає проблема, пов'язана із сумісністю форматів, а отже, і з доступом до даних для читання. Одним із рішень, яке можна запропонувати – використання програмних засобів, які підтримують єдиний формат доступу баз даних. Найбільш простим рішенням є використання стандартних технологій доступу до баз даних. Серед таких технологій можна виділити використання СУБД з підтримкою стандартної мови запитів SQL, а також стандартних інтерфейсів доступу до баз даних ODBC та OLEDB. Крім наведених технологій, є й інші відомі і можливо більш прогресивні рішення, однак на сьогодні в системах ІАСУ вони поки що не знайшли великої популярності.

ODBC (Open DataBase Connectivity) – один із стандартних інтерфейсів доступу до реляційних баз даних, які засновані на мові SQL. Архітектура ODBC показана на рис. 2.1.

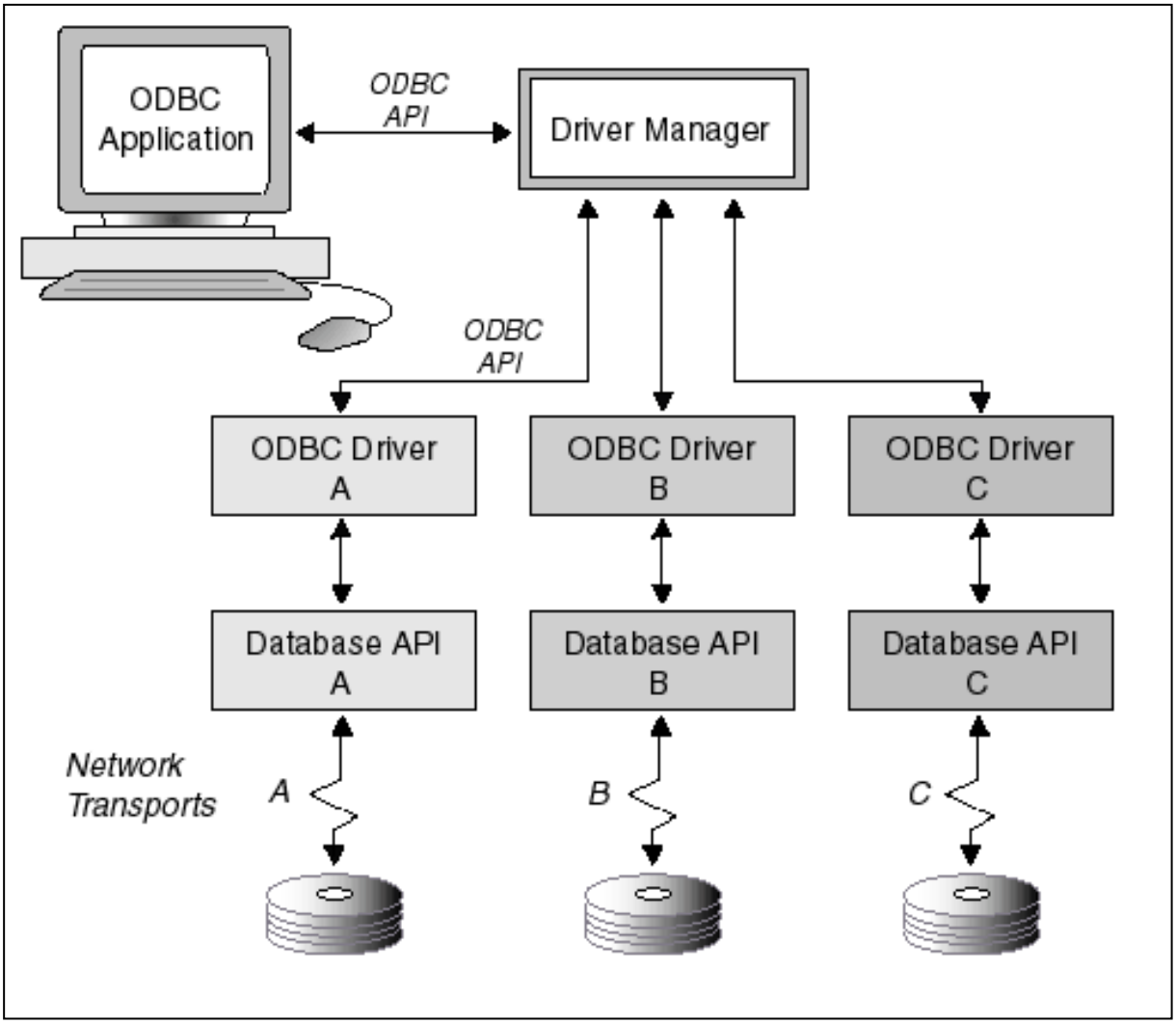


Рис. 2.1. Архітектура стандартного інтерфейсу відкритого доступу до реляційних баз даних (ODBC)

Зв'язок з різними джерелами даних (базами даних, файловими ресурсами тощо) проводиться за допомогою драйверів ODBC, які оформлені у вигляді динамічних бібліотек (DLL) і підтримують єдиний інтерфейс ODBC API. Даний API базується на виконанні SQL-запитів. Тобто незалежно від типу джерела даних прикладна програма генерує SQL-запити, які відправляються потрібному драйверу, що працює з даним джерелом. Драйвер забезпечує потрібне перетворення цих запитів у мову джерела, або сам виконує необхідні операції, що вказані в запиті.

У клієнтській прикладній програмі як джерело даних вказується ім'я DSN. З'єднання прикладної програми з необхідним джерелом даних забезпечиться диспетчером драйверів (Driver Manager), який реалізований у вигляді бібліотеки ODBC.DLL

(рис. 2.1). Під час написання прикладних програм з використання ODBC більш зручний доступ до ODBC через RDO та DAO (рис. 2.2).

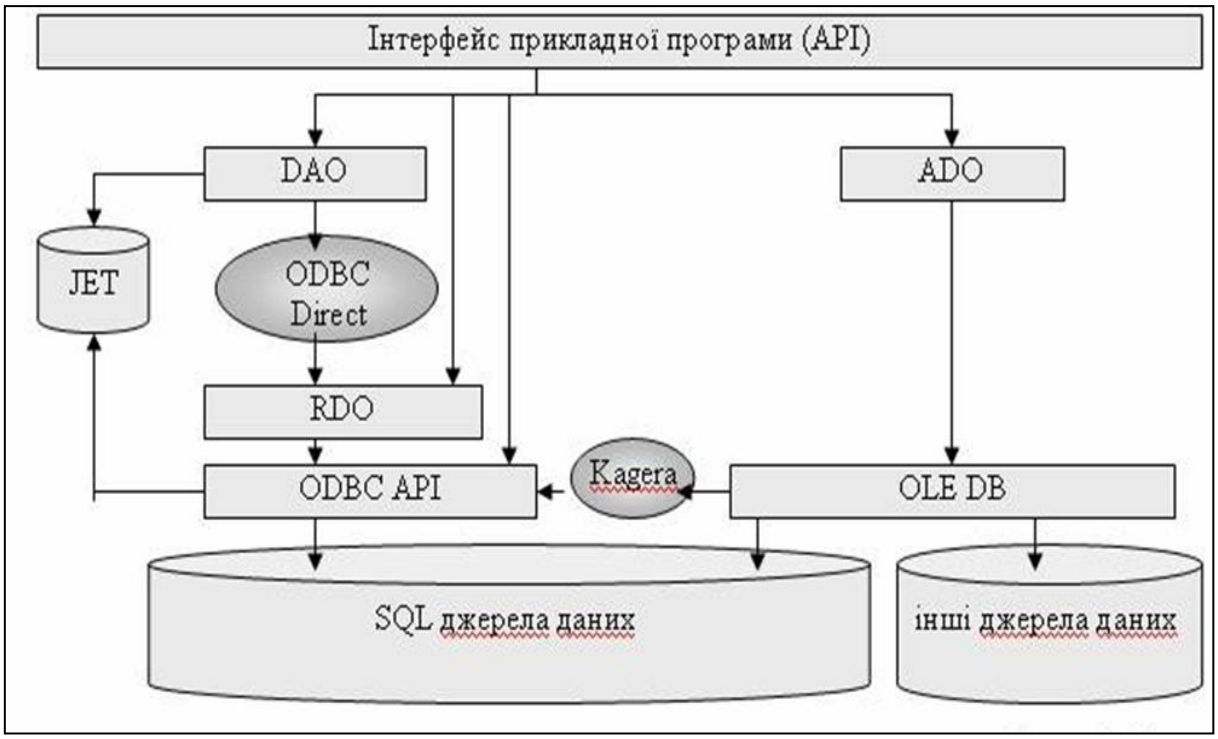


Рис. 2.2. Принцип доступу до джерела даних через стандартні інтерфейси

RDO (Remote Data Object) – є об'єктним інтерфейсом доступу до ODBC джерел даних.

DAO* (*Data Access Object*)** – є об'єктним COM-інтерфейсом до процесора баз даних ***Jet (*Joint Engine Technology database engine*), а також до надбудови над RDO-інтерфейсом. Режим ODBC Direct перетворює усі об'єкти та методи DAO в еквівалент RDO.

Kagera – менеджер драйвера ODBC, що включає модифікації, які відповідають специфікації OLE.

2.1.3. Особливості з'єднання з БД Oracle через ODBC

Для з'єднання з БД Oracle на основі технології ODBC потрібно інсталювати клієнтську версію Oracle, із якою встановлюється актуальна версія драйвера ODBC. При інсталяції *Net Configuration Assistant* формує файл *tnsnames.ora*, у якому прописуються аліаси з'єднання із конкретною БД Oracle (*TNS service name*).

При інсталяції на одному сервері БД і клієнта Oracle шляхи, де знаходяться дані файли, можуть виглядати так:

C:\Oracle\app\product\11.2.0\dbhome_1\NETWORK\ADMIN\ tnsnames.ora
C:\Oracle\app\product\11.2.0\client_1\network\admin\ tnsnames.ora

Це потрібно враховувати під час формування аліасів із іншими серверами БД Oracle. У різних програмних продуктах *TNS service name* використовується як *ім'я сервера* або *ім'я джерела даних*.

Нижче показано зміст файлу *tnsnames.ora із двома аліасами* ORCL2 і ORCL5 для з'єднання з двома навчальними серверами Oracle за адресами IP, відповідно 192.168.4.2 і 192.168.4.5:

```
ORCL2 =
  (DESCRIPTION =
   (ADDRESS_LIST =
    (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.4.2)
(PORT = 1521))
  )
  (CONNECT_DATA =
   (SERVICE_NAME = ORCL)
  )
  )
ORCL5 =
  (DESCRIPTION =
   (ADDRESS_LIST =
    (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.4.5)
(PORT = 1521))
  )
(CONNECT_DATA =
   (SERVICE_NAME = ORCL)
  )
  )
```

Екземпляр бази даних Oracle складається із системної області пам'яті SGA і процесів Oracle. Ім'я екземпляра бази даних вказується у файлі ініціалізації *init.ora* параметром *instance_name.* Кожній БД Oracle відповідає один екземпляр. Системний ідентифікатор

SID (System IDentifier – системний ідентифікатор) є унікальним ім'ям, яке однозначно ідентифікує екземпляр / базу даних. SID як ORACLE_SID прописується у *tnsnames.ora* і використовується утилітами і мережевими компонентами для доступу до бази даних.

Дамо деякі пояснення до змісту файлу на прикладі ORCL5.
ORCL5 = -- TNS service name. Ім'я з'єднання (аліас) з БД Oracle.
POTOCOL = TCP -- мережевий протокол TCP
HOST = 192.168.4.5 -- IP адреса сервера БД Oracle
PORT = 1521 -- порт для з'єднання з БД Oracle
SERVICE_NAME = ORCL -- SID унікальне ім'я екземпляра БД Oracle

Як видно вище із змісту файлу *tnsnames.ora,* два сервери мають активні екземпляри БД із однаковими унікальними іменами (SID) – *ORCL*.

2.1.4. Об'єктно-орієнтована технологія доступу до даних OLE DB

Альтернативою ODBC є ***OLE DB*** – це об'єктно-орієнтована технологія доступу до даних, яка базується на COM. Спеціалістами Microsoft запропонована стратегія універсального доступу до будь-яких джерел даних ***UDA*** (*Universal Data Access*), де драйвери OLE DB займають нижній щабель ієрархії (рис. 2.3). У цій архітектурі виділяються ***OLE DB Consumers*** (***OLE DB Споживачі***) – це будь-яка частина прикладної програми, яка *користується OLE DB-інтерфейсами* та ***OLE DB Providers*** (***OLE DB Провайдери***) – це частина прикладної програми, яка *надає* свої послуги через OLE DB-інтерфейси.

Розрізняють два види OLE DB провайдерів: ***OLE DB провайдер даних*** (***OLE DB Data Provider***), ***OLE DB провайдер сервісів*** (***OLE DB Service Provider***). OLE DB провайдер даних надає послуги доступу до власних баз даних у вигляді таблиць. Провайдер сервісів не має власних даних, однак як провайдер – він

надає набір послуг (сервісів). Споживачі користуються цими сервісами, частина з яких використовує провайдерів даних для доступу до даних. Тобто провайдер сервісів щодо провайдерів даних виступає як споживач (рис. 2.3.).

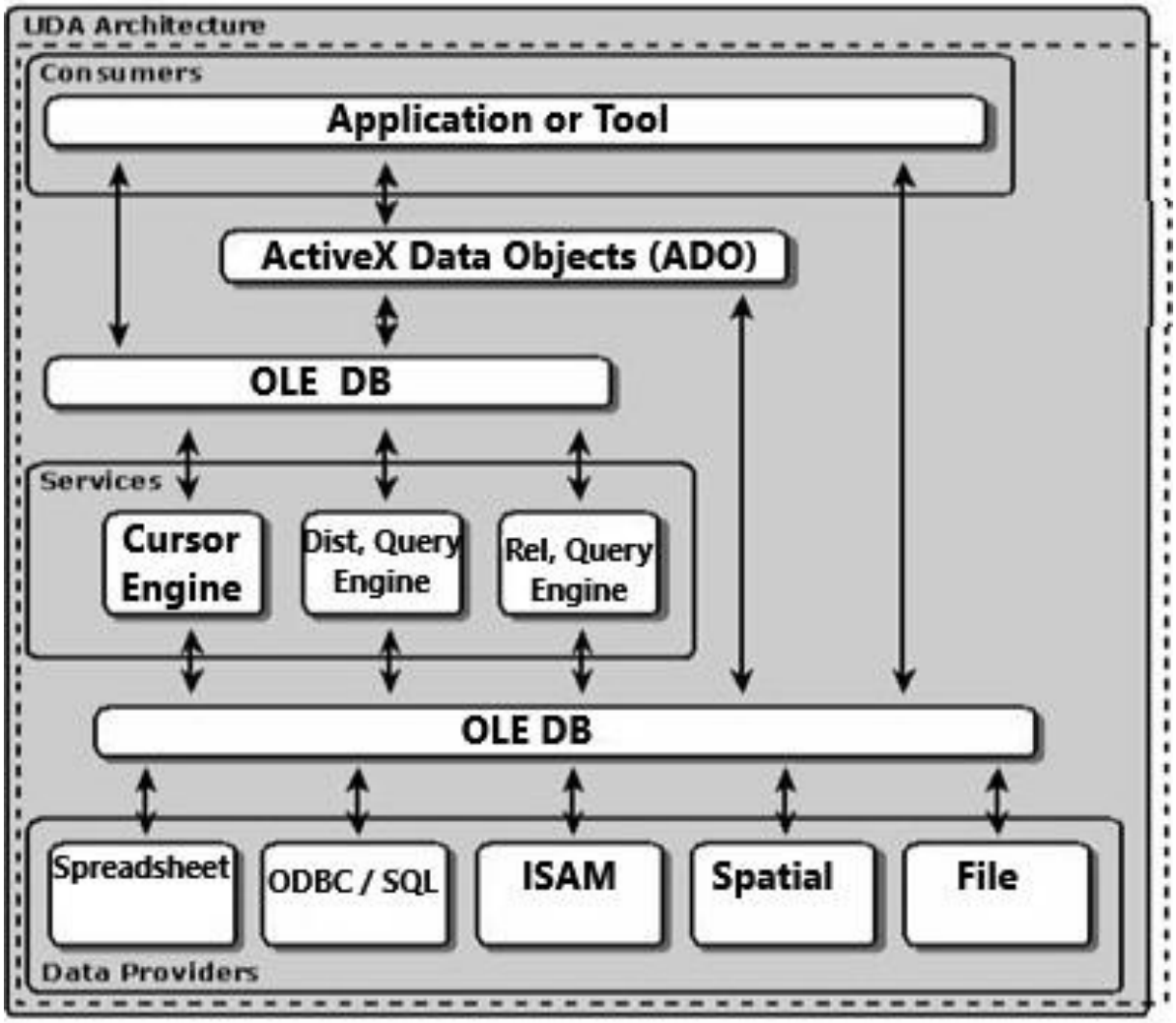


Рис. 2.3. Архітектура універсального доступу до джерел даних UDA (Universal Data Access)

На відміну від ODBC, який надає доступ тільки до реляційних даних, OLE DB може надати доступ до будь-якого джерела даних, оскільки базується на об'єктній моделі, яка спроектована для представлення даних довільного формату.

Для сумісності з розробленими ODBC драйверами, OLE DB має інтерфейс доступу до них через провайдера ODBC. Інтерфейс OLE DB – складний в реалізації і призначений для розробки драйверів для нового типу джерела даних. Для кінцевого користувача необхідно мати зручний і легко зрозумілий прикладний інтерфейс (API), яким служить технологія ADO.

2.1.5. Технологія ActiveX Data Object (ADO)

ADO (ActiveX Data Object) визначає модель програмування – послідовність дій, які необхідні для отримання доступу та модифікації джерела даних. Це об'єктна модель, яка являє собою набір об'єктів зі своїми методами, властивостями та подіями.

Модель базується на таких основних поняттях:

1. ***Connection*** (***Підключення***) – це джерело даних, з яким необхідно з'єднатися, може бути представленим у вигляді символьного ***Рядка підключення (Connection String)*** або ***Uniform Resource Locator*** (***URL***). У рядку вказується уся послідовність (транзакція) підключення, частини (кроки) якої розділені спеціальними роздільниками. У *Рядку підключення* спочатку вказується *провайдер OLE DB*, а потім усі інші параметри, які відрізняються залежно від *провайдера*. Потрібно зазначити, що транзакція виконується повністю або не виконується взагалі, тобто виконання підключення до частини, вказаної в *Рядку підключення,* неможливе.

2. ***Command*** (***Команда***) – вказує операцію, яку необхідно зробити з джерелом даних, визначених *Підключенням* (добавити, видалити, модифікувати, знайти дані за заданими параметрами). Розрізняють декілька типів команд:

– за замовчанням: тип команди визначається провайдером;

– текст (SQL): команда записується у вигляді SQL-виразу;

– таблиця: команда являє собою назву таблиці, всі колонки якої повертаються із запитом;

– збережена процедура (stored procedure) – назва процедури, яка повинна бути викликана;

– файл: вказується ім'я файлу, з яким необхідно з'єднатися.

1. ***Parameters*** (***Параметри***) – це параметри команди, які можуть змінюватися. Команда може виступати як функція, тобто частина команди незмінна, а інша частина виступає як параметри (параметричні запити).

2. ***Recordset*** – це об'єкт, який надає доступ до рядків таблиці, що повертає Команда-запит.

3. ***Field*** (***Поле***) – об'єкт, який надає доступ до полів об'єкта Recordset.

4. ***Error*** (***Помилка***) – об'єкт, який містить у собі інформацію про помилку. Ці об'єкти асоціюються з кожним підключенням.

5. ***Property*** (***Властивість***) – кожний ADO-об'єкт має набір динамічних та статичних властивостей.

6. ***Record*** (***Запис***) – це об'єкт, який надає доступ до даних, які подаються у вигляді ***контейнера*** (***container***) та ***місткості*** (***content***). Для прикладу в файловій системі каталоги – це контейнери, які можуть вміщувати інші каталоги (контейнери) або файли (місткість). Цей об'єкт дає можливість працювати з нереляційними даними.

7. ***Stream(Потік***) – це об'єкт, за допомогою якого можна працювати з потоками байтів файлів чи буферів пам'яті (дані content).

8. ***Collection*** (***Колекція***) – це об'єкт, який вміщує декілька об'єктів одного типу, до яких є доступ за іменем чи індексом.

9. Event (Подія).

Таким чином для доступу до даних за допомогою ADO та їх модифікації необхідно провести таку послідовність:

1. Вказати джерело даних для *підключення*, тобто його розміщення в *Рядку підключення* або в *URL*. Виконати з'єднання з вказаним джерелом.

2. Визначити *команду* для доступу до джерела даних, її тип та за необхідності параметри. Виконати команду.

3. Для реляційних баз даних результати виконання команди у вигляді таблиці записують у кеш. За необхідності їх можна модифікувати за допомогою об'єкта *Recordset.*

Технологія ADO як API інтерфейс OLE DB надає стандартні сервіси та методи роботи з даними незалежно від *провайдера даних*. Однак у зв'язку зі специфікою останніх можливі деякі особливості у процесі їх використання. Відмінності, як правило, стосуються Рядка підключення (*Connection String*), використання *команди* та об'єкта *Recordset.* Проаналізуємо основні правила формування *Рядка підключення* та *Команди*.

Рядок *підключення (Connection String)* формується з послідовності виразів типу *аргумент=значення* розділених крапкою з комою. Типові аргументи наведені в табл. 2.1.

Таблиця 2.1

Типові аргументи ADO

Аргумент	Пояснення
Provider=	Назва провайдера даних
File Name=	Назва файлу, де зберігаються наперед визначені налаштування підключення
Remote Provider=	Назва віддаленого провайдера (тільки для Remote Data Service)
Remote Server=	Шлях до віддаленого серверу (тільки для Remote Data Service)
URL=	URL-шлях до файлу чи папки

Команда (Command) вказується за допомогою текстового рядка. Правила його формування залежать від типу команди. Зазвичай – це команда SQL, діалект якої залежить від Провайдера даних, однак це може бути інший текст.

Популярними є такі *провайдери даних*:

– Microsoft OLE DB Provider for ODBC;

– OLE DB Provider for Microsoft Jet;

– Microsoft OLE DB Provider for SQL Server.

Усе більшого використання набуває технологія .NET, яка лягла в основу стандартного механізму доступу до даних під назвою ***ADO.NET***. Так само як ADO базується на OLE DB, ADO.NET базується на .NET Framework. Принципи застосування залишилися однакові. Відомі також такі провайдери від Microsoft:

– Провайдер даних NET Framework для SQL-сервера;

– Провайдер даних .NET Framework для OLE DB;

– Провайдер даних .NET Framework для ODBC;

– Провайдер даних .NET Framework для Oracle.

Як видно зі списку, технологія ADO.NET підтримує сумісність із OLE DB та ODBC.

Питання для самоконтролю до теми 2.1

1. Охарактеризуйте основні пакети статистичного та інтелектуального аналізу, які можна використовувати для аналізу податкової інформації.

2. Охарактеризуйте основні призначення тегнологій доступу до баз даних ODBC, DAO.

3. Охарактеризуйте основні призначення тегнологій доступу до баз даних ODBC, DAO.

4. Охарактеризуйте основні призначення тегнологій доступу до баз даних ADO, OLE DB, RDO.

5. Опишіть архітектуру стандартного інтерфейсу відкритого доступу до реляційних баз даних (ODBC).

6. Опишіть принцип доступу до джерела даних через стандартні інтерфейси.

7. У чому полягають особливості з'єднання з БД Oracle через ODBC?

8. Запишіть зміст файлу tnsnames.ora, якщо для з'єднання з двома навчальними серверами Oracle використовуються аліаси ORCL2 і ORCL5 за адресами IP, відповідно 192.168.8.3 і 192.168.8.4 і з іменами екземплярів відповідно orcl3, orcl4.

9. У чому полягає альтернатива OLE DB для використання її замість ODBC ?

10. Опишіть архітектуру універсального доступу до джерел даних UDA (Universal Data Access).

11. Опишіть порядок формування джерела даних БД Oracle.

12. У чому полягає відмінність формування джерела даних до файлової системи .dbf і БД Oracle?

13. Що означають терміни «SID», «аліас» при описі з'єднання за БД Oracle?

14. Що означають терміни ADO, ODBC, DODB?

2.2. SQL – мова структурованих запитів до баз даних

2.2.1. Загальні відомості та категорії команд SQL

SQL (*Structured Query Language* – мова структурованих запитів) – декларативна мова програмування для взаємодії користувача з базами даних, що застосовується для формування запитів, оновлення і керування реляційними БД, створення схеми бази даних та її модифікації, системи контролю за доступом до бази даних. 1970-х років для перевірки можливостей практичної реалізації реляційної моделі даних IBM розробила експериментальну СУБД System R. Для неї була створена мова SEQUEL (англ. *Structured English QUEry Language*). Мова отримала досить широку популярність і дещо пізніше, через існування зареєстрованої торгової марки SEQUEL, була перейменована в SQL. Підтримка мови SQL, незалежно реалізована в СУБД ряду виробників, привела до необхідності розробки стандартів цієї мови. Розробка специфікації SQL була розпочата Американським національним інститутом стандартів (ANSI) 1982 року. Стандарт був прийнятий ANSI 1986 року і 1987 року схвалений Міжнародною організацією зі стандартизації (ISO). Ця версія стандарту SQL, зазвичай, позначається як SQL-86. Наступна версія випущена 1989 року (SQL-89). У ній, зокрема, були чітко стандартизовані синтаксис і семантика операторів вибірки даних і маніпулювання даними, визначені засоби обмеження цілісності БД.

Версія стандарту SQL-92, також відома як SQL2, значно розширила попередню версію, визначивши кошти маніпулювання схемою БД, управління транзакціями і сесіями, підключення до БД тощо.

У випущену 1999 року версію SQL: 1999 (SQL3) додана підтримка регулярних виразів, рекурсивних запитів, підтримка тригерів, базові процедурні розширення, нескалярні типи даних і деякі об’єктно-орієнтовані можливості. Подальші версії – SQI-: 2003, SQL: 2006, SQL 2008 – внесли підтримку даних у форматі XML, спільного використання в запитах SQL і мови запитів до XML-даними XQuery (SQL: 2006) і ряд інших нововведень.

Починаючи з SQL: +1999 стандарт розподіляється на частини, кожна з яких оформляється у вигляді окремого документа. Після затвердження в ISO ці документи випускаються також у вигляді стандартів цієї організації (серія 9075). Наприклад, друга частина SQL: 2008 випущена як ISO / IEC 9075-2: 2011 Information technology – Database languages – SQL – Part 2: Foundation (SQL / Foundation) («Інформаційні технології. Мови бази даних. Мова структурованих запитів (SQL). частина 2. Основа»).

На сьогодні діє стандарт, який був прийнятий 2003 року (SQL:2003), а надалі зазнав незначних модифікацій.

Сама по собі SQL не є ані системою керування базами даних, ані окремим програмним продуктом. На відміну від дійсних мов програмування (С або Pascal), SQL може формувати інтерактивні запити або виступати як інструкції для керування даними. Крім того, стандарт SQL містить функції для визначення зміни, перевірки та захисту даних. SQL – це діалогова мова програмування для здійснення запиту і внесення змін до бази даних, а також керування базами даних. Багато баз даних підтримує SQL з розширеннями до стандартної мови. Ядро SQL формує командна мова, яка дозволяє здійснювати пошук, вставку, оновлення і вилучення даних за допомогою використання системи керування і адміністративних функцій SQL.

Основні переваги мови SQL полягають у такому:

Стандартність – використання мови ЗС^Ь у програмах стандартизовано міжнародними організаціями.

Незалежність від конкретних СУБД – усі поширені СУБД використовують, оскільки реляційні бази даних можна перенести з однієї СУБД на іншу з мінімальними доробками.

Можливість перенесення з однієї обчислювальної системи на іншу – СУБД може бути орієнтована на різні обчислювальні системи, однак програми, створені за допомогою SQL, допускають використання як для локальних БД, так і для великих багатокористувацьких систем.

Реляційна основа мови – SQL є мовою реляційних БД, тому вона стала популярною тоді, коли отримала широке розповсюдження реляційна модель представлення даних. Таблична

структура реляційної БД добре зрозуміла, а тому мова SQL проста для вивчення.

Можливість створення інтерактивних запитів SQL забезпечує користувачам швидкий доступ до даних, при цьому в інтерактивному режимі можна отримати результат запиту без написання складної програми.

Можливість програмного доступу до БД – мову SQL легко використовувати у програмах, яким необхідно звертатися до баз даних. Одні й ті ж оператори SQL вживаються як для інтерактивного, так і програмного доступу, тому частини програм, що містять звернення до БД, можна спочатку перевірити в інтерактивному режимі, а потім вбудовувати в програму.

Забезпечення різного подання даних – за допомогою SQL можна уявити таку структуру даних, що той чи інший користувач буде бачити різні їхні уявлення. Крім того, дані з різних частин БД можуть бути скомбіновані і представлені у вигляді однієї простої таблиці, а отже, представлення придатні для посилення захисту БД і її налаштування під конкретні вимоги окремих користувачів.

Можливість динамічної зміни і розширення структури БД дозволяє маніпулювати структурою БД, тим самим забезпечуючи гнучкість з позиції пристосованості БД до мінливих вимог предметної галузі.

Основні категорії команд мови SQL призначені для виконання різних функцій, включаючи побудову об'єктів бази даних і маніпулювання ними, навчальне завантаження даних у таблиці, оновлення та видалення існуючої інформації, виконання запитів до бази даних, керування доступом до неї та її загальне адміністрування.

Основні категорії команд мови SQL:

- **DDL** – мова визначення даних;
- **DML** – мова маніпулювання даними;
- **DQL** – мова запитів;
- **DCL** – мова керування даними;
- команди адміністрування даних;
- команди управління трансакціями.

Визначення структур бази даних (DDL – Data Definition Language).

Мова визначення даних дозволяє створювати та змінювати структуру об'єктів бази даних, наприклад, створювати і видаляти таблиці. Основними командами мови DDL є такі: CREATE TABLE, ALTER TABLE, DROP TABLE, CREATE INDEX, ALTER INDEX, DROP INDEX.

Маніпулювання даними (DML – Data Manipulation Language)

Мова маніпулювання даними використовується для маніпулювання інформацією всередині об’єктів реляційної бази даних за допомогою трьох основних команд: INSERT, UPDATE, DELETE.

Вибірка даних (DQL – Data Query Language)

Мова запитів найбільш відома користувачам реляційної бази даних, незважаючи на те, що вона містить лише одну команду SELECT. Ця команда разом зі своїми численними параметрами та пропозиціями використовується для формування запитів до реляційної бази даних.

Мова керування (адміністрування) даними (DCL – Data Control Language)

Команди керування даними дозволяють керувати доступом до інформації, що знаходиться всередині бази даних. Як правило, вони використовуються для створення об’єктів, пов’язаних з доступом до даних, а також служать для контролю над розподілом привілеїв між користувачами. Команди керування даними такі: GRANT, REVOKE.

За допомогою команд адміністрування даних користувач здійснює контроль за виконуваними діями і аналізує операції бази даних; вони також можуть виявитися корисними під час аналізу продуктивності системи. Не потрібно плутати адміністрування даних з адмініструванням бази даних, яке являє собою загальне управління базою даних і передбачає використання команд усіх рівнів.

Команди управління трансакціями

Існують такі команди, що дозволяють керувати трансакціями бази даних: COMMIT, ROLLBACK, SAVEPOINT, SET TRANSACTION.

2.2.2. Основні типи даних

У мові SQL є засоби, що дозволяють для кожного атрибута вказувати тип даних, якому відповідатимуть усі значення цього атрибута.

Символьні типи даних – містять літери, цифри та спеціальні символи.

CHAR або **CHAR(n)** – символьні рядки фіксованої довжини. Довжина рядка визначається параметром n. CHAR без параметра відповідає CHAR(l). Для зберігання таких даних завжди приділяється п байт поза залежністю від реальної довжини рядка.

VARCHAR(n) – символьний рядок змінної довжини. Для зберігання даних цього типу відводиться число байт, що відповідає реальній довжині рядка.

Цілі типи даних – підтримують тільки цілі числа (дробові частини і десяткові точки не допускаються). Над цими типами дозволяється виконувати арифметичні операції і застосовувати до них функції збору (визначення максимального, мінімального, середнього та сумарного значення стовпця реляційної таблиці).

INTEGER або INT-ціле, для зберігання якого відводиться, як правило, 4 байти. (Зауваження: число байт, яке відводиться для зберігання того чи іншого числового типу даних, залежить від СУБД, яка для цього використовується, та апаратної платформи, тут наводяться найбільш «типові» значення). Інтервал значень від: «-2147483647» до «+2147483648» **SMALLINT** – коротке ціле (2 байтів), інтервал значень від «-32767» до «+32768».

Речові типи даних – описують числа з дрібною частиною.

FLOAT та **SMALLFLOAT** – числа з плаваючою точкою (для зберігання приділяється зазвичай 8 і 4 байтів відповідно).

DECIMAL(p) – це тип даних FLOAT з кількістю значущих цифр **p.**

DECIMAL(p,n) – це DECIMAL(p), де **n** – кількість цифр після десяткової коми.

Грошові типи даних – описують природно грошові величини. Якщо ваша система такого типу даних не підтримує, то використовуйте DECIMAL(p,n).

MONEY(p,n) – відповідає типу DECIMAL(p,n). Вводиться тільки тому, що деякі СУБД передбачають для нього спеціальні методи форматування.

Дата і час – використовуються для зберігання дати, часу та їхніх комбінацій. Більшість СУБД вміє визначати інтервал між двома датами, а також зменшувати або збільшувати дату на певну кількість часу.

DATE – тип даних для зберігання дати.

TIME – тип даних для зберігання часу.

INTERVAL – тип даних для зберігання інтервалу часу.

DATETIME – тип даних для зберігання моментів часу (рік + місяць + день + годинник + хвилини + секунди + частки секунд).

Двійкові типи даних – дозволяють зберігати дані будь-якого обсягу в двійковому коді (зображення, файли тощо). Визначення цих типів найбільш різняться від системи до системи, часто використовуються ключові слова: BINARY; BYTE; BLOB.

Послідовні типи даних – використовуються для представлення зростаючих числових послідовностей.

SERIAL – тип даних на основі INTEGER, що дозволяє сформувати унікальне значення (наприклад, для первинного ключа). Додаючи запис, СУБД автоматично присвоює полю даного типу значення, одержуване зі зростаючої послідовності цілих чисел.

Невизначені або пропущені дані (NULL)

Для позначення відсутніх, пропущених або невідомих значень атрибуту в SQL використовується ключове слово NULL. Досить часто можна зустріти словосполучення «атрибут має значення NULL». NULL не є значенням у звичайному розумінні, а саме: використовується для позначення того факту, що дійсне значення атрибуту пропущене або невідоме. Це призводить до ряду особливостей, що потрібно враховувати під час використання значень атрибутів, які можуть перебувати в стані NULL.

У функції, що дозволяють одержувати зведену інформацію серед нескінченості значень атрибуту, наприклад сумарне або середнє значення, для забезпечення точності і однозначності тлумачення результатів відсутні або NULL – значення атрибутів ігноруються.

Умовні оператори від двозначної логіки TRUE/FALSE розширюються до TRUE/FALSE/UNKNOWN.

Усі оператори, за винятком оператора конкатенації строк «||», повертають порожнє значення (NULL), якщо значення кожного з операндів відсутнє (має «значення NULL»).

Для перевірки на порожнє значення потрібно використовувати оператори «is NULL» та «is NOT NULL» (використання з цією метою оператора порівняння «=» є помилкою).

Функції перетворення типів, які мають NULL як аргумент, повертають порожнє значення (NULL).

2.2.3. Створення та модифікація таблиць БД Oracle

Для створення, зміни та видалення таблиць у SQL БД використовуються оператори CREATE TABLE, ALTER TABLE і DROP TABLE.

Створення таблиці

Щоб створити таблицю за допомогою оператора CREATE TABLE, потрібно вказати такі дані:

– ім’я нової таблиці; воно вводиться після ключового слова CREATE TABLE;

– імена та визначення стовпців таблиці, розділені комами.

Приклад 2.1

```
CREATE TABLE ZREGDOC
COD_REGDOC NUMBER (17) NOT NULL,
COD_DOCUM NUMBER (6),
TIN NUMBER (10),
C_DOC VARCHAR2 (3 Byte),
D_GET DATE;
```

Оновлення таблиць

Для того, щоб оновити таблиці, потрібно скористатися оператором ALTER TABLE. Хоча всі СУБД підтримують цей оператор, те, що вони при цьому дозволяють вам робити, значною мірою залежить від реалізації СУБД. Нижче наведено кілька міркувань з приводу застосування оператора ALTER TABLE.

Усі СУБД дозволяють додавати у вже існуючі таблиці стовпці, але деякі обмежують типи даних, які можуть бути додані (а заодно і використання значень NULL і DEFAULT).

Щоб змінити таблицю за допомогою оператора ALTER TABLE, потрібно ввести таку інформацію.

– Ім'я таблиці, що підлягає зміні, після ключових слів ALTER TABLE (таблиця з такою назвою має існувати, інакше буде видано повідомлення про помилку).

– Список змін, які повинні бути зроблені. Оскільки додавання стовпців у таблицю – єдина операція, підтримувана усіма СУБД, саме її ми розглянемо як приклад.

Приклад 2.2

```
ALTER TABLE ZDATA_C
ADD CARD NUMBER (15).
```

За допомогою цього оператора в таблицю ZDATA_C додається стовпчик, названий CARD. Повинен бути визначений тип даних. Інші операції зміни, наприклад, зміна або видалення стовпців, введення обмежень або ключів, вимагають схожого синтаксису.

Приклад 2.3:

```
ALTER TABLE ZDATA_C DROP COLUMN
 COD_REGDOC;
```

Оновлення таблиць

Видалення таблиць (мається на увазі видалення саме таблиць, а не їхнього вмісту) – дуже простий процес. Таблиці видаляють за допомогою оператора DROP TABLE.

Приклад 2.4:

```
DROP TABLE CARD
```

Цей оператор видаляє таблицю CARD. У цьому випадку не потрібно ніякого підтвердження, неможливо повернутися до попереднього стану – в результаті застосування цього оператора таблицю буде безповоротно вилучено.

Перейменування таблиць

У різних СУБД перейменування таблиць здійснюється порізному. Не існує жорстких, усталених стандартів на виконання цієї операції. Користувачі СУБД DB2, MySQL, Oracle та PostgreSQL можуть застосовувати для цього оператор RENAME. Основний синтаксис для всіх операцій перейменувати вимагає вказівки старого і нового імен.

Додавання рядків у таблицю

Для додавання рядків у таблицю SQL бази даних використовується оператор INSERT INTO.

Основні його синтаксичні форми описуються нижче.

1. Додавання рядка перерахуванням значень усіх її комірок

INSERT INTO ім'я_табл VALUES (знач, . ..),

де «знач» – константні значення комірки рядка.

Значення осередків у списку повинні відповідати порядку перерахування специфікацій стовпців таблиці в операторі CREATE TABLE. Можливо як «знач» вказувати ключове слово NULL, що означає відсутність значення для відповідної комірки рядка. Перед додаванням нового рядка в таблицю СУБД перевіряє допустимість перерахованих значень, використовуючи опис стовпців таблиці з оператора CREATE TABLE.

2. Додавання рядка з використанням списку імен стовпчиків

INSERT INTO імя_табл (ім'я_стовпця, - - -)
VALUES (знач, . .),

Тут списки імен стовпців і значень комірок додаються в рядки, які повинні бути узгоджені, хоча немає ніяких вимог до їхнього порядку. Можна пропускати в списках інформацію про деякі комірки рядка, при цьому комірки, що відповідають стовпчикам зі специфікацією NULL в операторі CREATE TABLE, будуть порожніми.

Комірки, що відповідають стовпчикам зі специфікацією NOT NULL в операторі CREATE TABLE, заповнюються значеннями «за замовчуванням».

3. Додавання рядків за результатами запиту до БД

INSERT INTO ім'я_табл [(ім'я_стовбця, ...)] SELECT ...

Такий оператор дає можливість додати в таблицю 0, 1 або відразу декілька нових рядків, отриманих у результаті запиту до бази даних, що реалізується оператором SELECT.

Приклад 2.5

--Додавання нової інформації в таблицю ZDATA_N:
INSERT INTO ZDATA_N VALUES (25, 6.3, 1.8).

Зазначимо, що додавання нового рядка буде вдалим лише в тому випадку, якщо рядка з таким же ідентифікатором у таблиці ZDATA_N ще немає – справа в тому, що стовпець id цієї таблиці оголошений первинним ключем і, отже, усі його значення повинні бути унікальними.

Визначення структури таблиці

Створюючи свою власну таблицю, користувач знає її структуру до певного періоду. Потім вносяться зміни іншими користувачами, і тоді втрачається уявлення про її структуру. Отже, потрібно мати можливість визначати структуру існуючої таблиці. В Oracle є команда, призначена саме для цієї мети. Вона називається **DESCRIBE** (скорочено – DESC) і має такий синтаксис:

DESC «ім'я_таблиці»

Це одна з небагатьох команд, які не потрібно завершувати крапкою з комою. З іншого боку, наявність крапки з комою нічому не пошкодить, так що можете її ставити просто для закріплення звички. Щоб побачити, як працює команда DESC, введіть такий код: DESC CARD.

Вихідними даними команди DESC є три стовпці: «Name», «Null» і «Type». Під заголовком «Name» перераховані імена всіх стовпців таблиці в порядку їхньої появи в цій таблиці. Призначення стовпця «Null?» буде пояснено в наступному розділі, а стовпець Type показує тип даних і довжину кожного із стовпців.

2.2.4. Команда SELECT. Основи запитів на мові SQL

Основний засіб для здобуття інформації з бази даних – це запити, тобто команда SELECT. Оскільки у основних базах даних підтримується стандарт ANSI SQL, то базові можливості і в Oracle, і в інших базах даних (Microsoft SQL Server, IBM Db2, Informix, Sybase, Microsoft Access) схожі.

Оператор SELECT мови SQL є найбільш важливим і найбільш часто використовуваним оператором. Він призначений для вибірки інформації з таблиць бази даних. Спрощений синтаксис оператора SELECT виглядає таким чином:

Приклад 2.6.

SELECT [DISTINCT] <список атрибутів>
FROM <список таблиць>
[WHERE <умова вибірки>]
[ORDER BY <список атрибутів>]
[GROUP BY <список атрибутів>]
[HAVING <умова>]
[UNION <вираз з оператором SELECT>;

У квадратних дужках зазначені елементи, які можуть бути відсутні у запиті.

Ключове слово **SELECT** повідомляє базі даних, що така команда є запитом на одержання інформації.

Після слова SELECT через кому перераховуються найменування полів (список атрибутів), вміст яких запитується.

Обов'язковим ключовим словом у запиті SELECT є слово FROM (з). За ключовим словом FROM вказується список розділених комами імен таблиць, з яких отримується інформація.

Важливо пам'ятати! Будь-який SQL-запит повинен закінчуватися символом «;» **(крапка з комою).**

Приклад 2.7

```
SELECT EDRPOU, NAME
FROM ORG;
```

Для виключення з результату SELECT-запиту записів, що повторюються, використовується ключове слово **DISTINCT** (відмінний).

Якщо запит SELECT повинен дати вибірку із багатьох полів, то DISTINCT виключає дублікати рядків, в яких значення усіх обраних полів ідентичні. Попередній запит можна записати в такому вигляді.

Приклад 2.8

```
SELECT DISTINCT EDRPOU, NAME
FROM ORG;
```

У результаті отримаємо таблицю, в якій дублікати строк виключені.

Вибірка окремих стовпців

Приклад 2.9

```
SELECT CODE
FROM CARD;
```

У наведеному вище операторі використовується оператор SELECT для вибірки одного стовпця під назвою CODE з таблиці CARD. Ім’я стовпця вказується відразу після ключового слова SELECT, а ключове слово FROM вказує на ім’я таблиці, з якої вибираються дані.

Для вибірки з таблиці декількох стовпців використовується той самий оператор SELECT. Відмінність полягає у тому, що після ключового слова SELECT необхідно через кому вказати кілька імен стовпчиків.

Приклад 2.10

SELECT TIN, C_DOC, C_DOC_SUB
bestzvit.ZREGDOC;

Використання псевдоніму

Ім’я стовпця, що виводиться, можна поміняти за допомогою псевдоніма **(alias). Псевдонім** – це альтернативне ім’я для поля чи значення. Для цього досить просто написати ім’я псевдоніма відразу після назви стовпця або використовувати ключове слово AS (обидва синтаксиси повністю рівнозначні).

Приклад 2.11

SELECT C_RAJ Код_району
FROM tax_l.T21VIDSTR
або:
SELECT C_RAJ AS Код_району
FROM tax_l.T21VIDSTR

Особливо зручно використовувати псевдоніми для стовпців, що обчислюються, і у яких ім’я не передбачене:

Приклад 2.12

--Визначити суму податкових зобов’язань (рядок 9 декларації з ПДВ)
-- платника податків з кодом 00292988 за 2017 рік
SELECT SUM(R02G4) AS SUM9R2017
FROM bestzvit.DOD5VAL0
WHERE FIRM YO=00292988;
SUM9R2017 132029005;

Псевдонімом може служити як одне слово, так і цілий рядок. Якщо в псевдонімі використовуються пробіли або якісь службові символи, то такий псевдонім потрібно помістити в подвійні лапки.

Таким чином, найбільш часто псевдоніми використовуються для перейменування багатослівних назв стовпців у однослівне.

Арифметичні оператори в запитах Oracle SQL.

Оператори – це синтаксичні конструкції мови, які призначені для виконання певних дій.

У Oracle SQL передбачена велика кількість операторів – як звичайних (наприклад, арифметичні або логічні), так і екзотичних (оператори для роботи з ієрархічними наборами записів, з безліччю записів – MULTISET).

Усі оператори Oracle SQL можна поділити на два основних типи:

1) унарні;
2) бінарні.

Унарні – оператори працюють лише з одним значенням.

До унарних операторів відносять оператора (-): якщо для нього використовувати лише одне значення, він змінює знак числа.

Бінарні – оператори працюють з двома значеннями.

У виразі «10-7» оператор (-) – це бінарний оператор, який міняє знак числа.

До арифметичних також відносять унарні оператори (+) і (-), які визначають знак числа. Пріоритет арифметичних операторів також звичний: найвищий пріоритет в унарних операторів, потім виконується множення і ділення, а після цього – складання і віднімання. Порядок застосування операторів регулюється круглими дужками.

Застосування унарного оператора до значення типа NULL дасть у результаті NULL. Оператори конкатенації (злиття строкових значень), злиття строкових даних різних типів. У реальній роботі досить часто доводиться виробляти злиття рядкових значень. Така операція називається конкатенацією.

У Oracle SQL передбачений оператор конкатенації – дві вертикальні межі (||). Крім того, в Oracle SQL передбачена ще і вбудована функція CONCAT, яка також виробляє конкатенацію і працює на будь-яких платформах.

Приклад 2.13

SELECT EDRPOU || ' ' || NAME AS "Код і назва підприємства"
FROM bestzvit.org
where EDRPOU LIKE 20543599;

Результат

Код і назва підприємства
20543599 Державна податкова інспекція в Тисменицькому районі.

У цьому випадку у нас зливається два значення: код, пробіл (у вигляді літерала) та назва підприємства.

Оператора конкатенації можна застосовувати як для рядкових значень і значень CLOB (великих рядкових значень), так і для чисел та дат.

Під час виконання конкатенації рекомендується подумати над обмеженнями типів даних, які використовуються для конкатенації. Oracle використовує для результату конкатенації більший тип даних з тих значень, які використовуються для конкатенації. Наприклад, при злитті двох значень з типом даних CHAR (максимальний розмір – 2000 символів) повернеться CHAR, при злитті CHAR і Varchar2 (максимальний розмір – 4000 символів) повернеться Varchar2, а при злитті будь-якого типа даних з CLOB повернеться CLOB. Важливо пам’ятати! Пробіли при конкатенації завжди зберігаються.

Фільтр WHERE і логічні оператори

Якщо звернутися до базового синтаксису команди SELECT, то можна побачити, що після списку стовпців повинно йти ключове слово FROM з вказівкою джерела інформації (таким джерелом може бути таблиця, представлення, підзапит, функція і тому подібне). Відразу після вираження FROM повинно йти ключове слово WHERE з виразом, який визначає фільтр для отримання записів з джерела. Повертаються лише ті записи, для яких вираз WHERE повертає true.

Приклад 2.14

```
SELECT *
FROM bestzvit.ZREGDOC
WHERE D_GET < '01.01.2008';
```

Прості умови у фільтрі WHERE (прості оператори порівняння)

Умова	Значення
=	Дорівнює
<	Менше
>	Більше
< =	Менше або дорівнює
> =	Більше або дорівнює
<>	Не дорівнює

Зазначимо лише про два моменти:

– інколи у користувачів бувають проблеми з тим, як саме пишеться більше або рівно (або менше або рівно) – що йде спочатку: символ рівності або знак більше/менше. Тут можна просто запам’ятати принцип: як говоримо, так і пишемо;

– «не дорівнює» в Oracle може позначатися відразу чотирма різними способами, у тому числі досить екзотичними. Проте рекомендується за можливості використовувати стандартне позначення вигляду (о).

Приклад 2.15

```
SELECT *
FROM bestzvit.ZREGDOC
WHERE D GET <> '01.01.2017';
```

Групові умови (оператори порівняння) в запитах, ключові слова ALL, ANY, SOME

Для групових умов порівняння в Oracle SQL використовуються три ключові слова: **ALL, SOME і ANY**. Всі ці ключові слова застосовуються для порівняння вказаного вами значення з набором значень, який повертає підзапит (або чітко вказаний набір значень).

ALL – порівняння проводитиметься зі всіма записами, які повертає підзапит (або просто зі всіма значеннями в набір). **True** повернеться лише в тому випадку, якщо всі записи, які повертає підзапит, задовольнятимуть вказаній вами умові. Крім того, в Oracle значення **True** повернеться за ситуації, коли підзапит не поверне жодного запису.

Приклад 2.16

```
SELECT * FROM tax_1.T21PENNOTI
WHERE ZN <= ALL
(SELECT ZN FROM tax_l.T21PENNOTI WHERE yearoper
='2017');
```

Він поверне записи для всіх підприємств, у яких сума пені менша або дорівнює найменшій пені у 2017 році.

ANY – порівняння поверне True, якщо умові задовольнятиме будь-який запис з набору (або підзапиту). Наприклад, такий запит поверне всі підприємства, КВЕД діяльності яких збігається із обраним підприємством.

Приклад 2.17

```
SELECT tin, full__name, kved, c_stan, c_sti, d_reg_sti
FROM aisr.r21taxpay where kved = ANY(SELECT kved
FROM aisr.r21taxpay WHERE tin = 00300015)
```

SOME – порівняння поверне True, якщо умові задовольнятимуть деякі записи з набору (або підзапиту).

Логічні умови в Oracle SQL дозволяють визначити потрібну комбінацію з декількох звичайних умов. Таких умов в Oracle SQL передбачено три:

AND – логічне «І». Щоб вираз повернув значення True, потрібно, щоб були істинними обидві умови.

Приклад 2.18

```
SELECT *
FROM bestzvit.ZREGDOC
WHERE C_DOC='J02'
AND C_DOC_SUB=151;
```

OR – логічне АБО. Щоб вираз повернув значення True, достатньо, щоб істинною була лише одна умова.

Приклад 2.19

```
SELECT *
FROM bestzvit.ZREGDOC
WHERE D_ENTER IN ( '18.06.2009' , 49.06.2009')
OR C_DOC= 'JO2'
AND C_DOC_SUB=151
AND C_DOC VER=1;
```

NOT – при додаванні NOT вираз буде повертати True, коли без нього буде повертатись False і навпаки.

Приклад 2.20

```
SELECT *
FROM ZREGDOC
WHERE NOT TIN = 22186593;
```

Умови для роботи зі значеннями типу NULL (порожні значення).

Якщо якась комірка в таблиці бази даних не має значення, то вважається, що в цій комірці знаходиться спеціальне значення NULL.

Важливо пам’ятати!

NULL – це не «0» і не порожній рядок. NULL вважається спеціальним значенням, для якого існують спеціальні прийоми роботи.

Застосувати спеціальні оператори порівняння для значень типу NULL не вийде: NULL ніколи не дорівнює другому значення NULL. Тому для роботи з ним передбачені дві спеціальні умови:

IS NULL – ця умова поверне True, якщо значення, що перевіряється, дорівнює NULL;

IS NOT NULL – це умова поверне NULL, якщо значення, що перевіряється, не дорівнює NULL

Приклад 2.21

Повернути всі записи з таблиці Locations, для яких не вказано поштовий індекс, можна так:

```
SELECT *
FROM ZREGDOC
WHERE N_REG IS NULL;
-- Повернути всі інші записи :
SELECT *
FROM bestzvit.ZREGDOC
WHERE N_REG IS NOT NULL;
```

Умова BETWEEN для порівняння діапазонів

Іноді необхідно повернути всі значення, які потрапляють у будь-який діапазон. Для цієї мети дуже зручно використовувати спеціальний оператор для порівняння діапазонів BETWEEN. Наприклад, щоб повернути інформацію про суму нарахованої пені за несвоєчасну сплату по всіх підприємствам району, і сума нарахованої пені у яких в діапазоні від 100 до 500 грн, можна використовувати запит такого вигляду.

Приклад 2.22

```
SELECT *
FROM tax_l.T21PENNOTI
WHERE ZN BETWEEN 10 0 AND 500;
```

Зверніть увагу, що під час використання ключового слова BETWEEN граничні значення також потрапляють у діапазон, тому запит поверне в тому числі підприємства із сумою нарахованої пені 100 і 500.

Можемо замінити запит з ключовим словом BETWEEN на запит зі звичайними операторами порівняння.

Приклад 2.23

```
SELECT *
FROM tax_I.T21PENNOTI
WHERE ZN >= 100 AND ZN <= 500;
```

Робота зі списками значення в запитах, умова IN

Якщо умова BETWEEN призначене для роботи з діапазонами, то умова IN використовується для перевірки значення на відповідність списку значень. Наприклад, повернути всі записи для підприємств, які подали додаток № 5 «Розшифровки в розрізі контрагентів» 18 та 19 червня 2009 року, можна так:

Приклад 2.24

```
SELECT *
FROM bestzvit.ZREGDOC
WHERE D_ENTER IN ('18.06.2017' , ' 19.06.2017')
AND C_DOC=J0 2
AND C_DOC_SUB=151
AND C_DOC_VER=1;
```

Щоб повернути всі значення, які не потрапляють у зазначений список, можна використовувати ключове слово NOT. Наприклад, щоб повернути інформацію про всіх співробітників з іншими посадами, можна використовувати запит.

Приклад 2.25

```
SELECT *
FROM bestzvit.ZREGDOC
WHERE D_ENTER NOT IN ('18.06.2017' , '19.06.2017')
and C_DOC=J02
and C_DOC_SUB=151
and C_DOC_VER=1;
```

Звичайно ж, як і у випадку з BETWEEN, синтаксичні конструкції з IN можна замінити на звичайні оператори порівняння з OR. Однак порівняно з великим списком значень синтаксис з IN зручніший.

Впорядкування вихідних даних за допомогою оператора ORDER BY

Як уже зазначалося, записи в таблицях реляційної бази даних не впорядковані. Однак дані, що виводяться у результаті виконання запиту, можуть бути впорядковані. Для цього використовується оператор ORDER BY, який дозволяє організовувати записи, що виводяться відповідно до значень одного або декількох обраних стовпців. При цьому можна задати зростаючу (ASC) або спадну (DESC) послідовність сортування для кожного зі стовпців. За замовчуванням прийнята зростаюча послідовність сортування.

Приклад 2.26

– Запит, що дозволяє вибрати всі дані з таблиці SUBJECT з упорядкуванням за ЄДРПОУ платника податків, виглядає таким чином:

```
SELECT *
FROM bestzvit.ZREGDOC
ORDER BY tin;
```

Приклад 2.27

– Той самий список, але впорядкований у зворотному порядку, можна отримати запитом:

```
SELECT *
FROM bestzvit.ZREGDOC
ORDER BY tin DESC;
```

Приклад 2.28

– Можна впорядкувати дані з таблиці за кодом платника, а всередині кодів платника – за найменуваннями реєстраційним номером.

```
SELECT *
FROM bestzvit.ZREGDOC
ORDER BY tin, N_REG;
```

Команда ORDER BY може використовуватися з GROUP BY для впорядкування груп записів. При цьому оператор ORDER BY у запиті завжди повинен бути останнім.

При впорядкуванні замість найменувань стовпців можна вказувати номери, маючи на увазі, що в даному випадку це номер b стовпців, зазначені при визначенні вихідних даних у запиті, а не номери стовпців у таблиці. Поле з номером 1 є першим полем, вказаним у команді ORDER BY – незалежно від його розташування у таблиці.

Приклад 2.29

```
SELECT tin, n_reg
FROM bestzvit.ZREGDOC
ORDER BY 1 DESC;
```

У цьому запиті записи будуть впорядковані у полі TIN.

Важливо пам'ятати! Якщо в полі, що використовується для впорядкування, існують NULL-значення, то всі вони розміщуються в кінці або передують усім іншим значенням цього поля.

Об'єднання декількох таблиць. Оператори Inner Join та Outer (left, right, full) Join в SQL (Oracle)

Інструкція **JOIN** дозволяє об'єднати стовпчики із декількох таблиць в одну таблицю. Є три типи JOIN-виразів:

```
INNER JOIN;
OUTER JOIN;
CROSS JOIN;
```

У свою чергу, OUTER JOIN може бути LEFT, RIGHT та FULL. Конструкція **JOIN** використовується після оператора **FROM** та має такий вигляд:

```
SELECT * FROM table_name_l join_type JOIN table_name_2
ON CONDITION
 ...
SELECT * FROM table_name_l join_type JOIN table_name_2
ON USING (KEY).
```

– де join_type - тип join-виразу;

– table_name_l, table_name_2 – назви таблиць, які об'єднуються;

– CONDITION, USING(KEY) – умова об'єднання таблиць.

Таблиця 2.2

Типи об’єднань та їхні характеристики

№ з/п	Тип об’єд-нання	Синтаксис	Отриманий результат
1.	**INNER JOIN**	Таблицяі INNER JOIN Табли-ця2 ON (Таблиця1.ключове поле = Таблиця2.ключове поле) або Таблиця1 INNER JOIN Таблиця2 USING (ключове поле, спільне для двох таблиць)	Об’єднуються рядки Таблиці1 та Таблиці2, для яких має місце повна відповідність ключових полів
2.	**LEFT OUTER JOIN**	Таблицяі LEFT OUTER JOIN Таблиця2 ON (Таблиця1. ключове поле = Таблиця2.ключове поле) або Таблиця1 LEFT OUTER JOIN Таблиця2 USING (ключове поле, спільне для двох таблиць)	Як результат, отримаємо всі рядки з Таблиці1 та тільки ті рядки з Таблиці2, для яких має місце відповідність ключових полів (Таблиці1 та Таблиці2)
3.	**RIGHT OUTER JOIN**	Таблиця1 RIGHT OUTER JOIN Таблиця2 ON (Таблиця1.ключове поле = Таблиця2.ключове поле) або Таблиця1 RIGHT OUTER JOIN Таблиця2 USING (ключове поле, спільне для двох таблиць)	Як результат, отримаємо всі рядки з Таблиці2 та тільки ті рядки з Таблиці1, для яких має місце відповідність ключових полів (Таблиці1 та Таблиці2)
4.	**FULL OUTER JOIN**	Таблиця1 FULL OUTER JOIN Таблиця2 ON (Таблиця1.ключове поле = Таблиця2.ключове поле) або Таблиця1 FULL OUTER JOIN Таблиця2 USING (ключове поле, спільне для двох таблиць)	Як результат, отримаємо всі рядки з Таблиці1 та всі рядки з Таблиці2, для яких має місце відповідність ключових полів (Таблиці1 та Таблиці2) та для яких немає такої відповідності. Іншими словами, отримаємо об’єднаний результат виконання запитів з об’єднанням типу left outer join та right outer join

Нижче наведено приклад використання одного із об’єднань у роботі підрозділів економічного аналізу ДФС України.

Приклад 2.30

Необхідно порахувати суму податку на додану вартість задекларованого у червні 2017 року, належну до сплати до бюджету (за мінусом переплат, наявних на перше число звітного місяця). Для цього об’єднаємо як дві таблиці два підзапити:

2) інформація про нараховані суми ПДВ за травень 2017 року в розрізі СГД:

```
SELECT c_sti, tin, r27
FROM dekl_pdv_062017
```

3) інформація про переплати по ПДВ станом на 01 червня 2017 року в розрізі СГД

```
SELECT c_reg*100+c_raj, c_sti, kod,
sum(decode(idparm,53,zn))
AS perepl
FROM galuz.vg306
WHERE rdata='01-06-2017' AND idparm=53 AND
shot=14010100
GROUP BY c_reg*100+c_raj, kod
```

У результаті виконання запиту отримаємо таблицю:

Код ДПІ	Код СГД	Нараховано ПДВ за декларацією	Переплата з ПДВ	Сума ПДВ до сплати
12*1	12***3224	123245656	546456	234234
23*4	34***2345	45645646	6777	5567567

Питання для самоконтролю до теми 2.2

1. Опишіть основні переваги мови SQL.

2. Перерахуйте основні категорії команд мови SQL.

3. Наведіть приклади застосування команд групиDML - Data Manipulation Language.

4. Наведіть приклади застосування команд групи DQL - Data Query Language.

5. Наведіть приклади застосування команд групи DCL - Data Control Language.

6. Наведіть приклади застосування команд групи управління транзакціями.

7. Перерахуйте і наведіть приклади основних типів даних SQL.

8. Опишіть тип даних «Невизначені або пропущені дані», (NULL). Наведіть приклади.

9. Наведіть приклади застосування умовних операторів TRUE/FALSE/UNKNOWN.

10. Наведіть приклад використання оператора для перевірки на порожнє значення «is NULL» та «is NOT NULL».

11. Напишіть групу команд створення простої таблиці із двома полями числового типу даних та двома полями текстового типу даних.

12. Напишіть скрипт заповнення простої таблиці даними (наприклад таблиці, створеної у п.11).

13. Напишіть скрипт оновлення даних (наприклад, даних таблиці, заповненої у п.12).

14. Напишіть команди SQL для визначення суми податкових зобов'язань (рядок 9 декларації з ПДВ) платника податків з кодом 002923458 за 2019 рік.

15. Напишіть запит, що дозволяє вибрати всі дані з таблиці SUBJECT з упорядкуванням за ЄДРПОУ платника податків.

16. Напишіть запит із об'єднанням таблиць, використовуючи оператори: INNER JOIN; OUTER JOIN; CROSS JOIN. Поясніть їхній зміст.

17. Напишіть запит для розрахунку суми податку на прибуток підприємств, задекларованого у червні 2019 року, що належить сплатити до бюджету (за мінусом переплат, наявних на перше число звітного місяця.

2.3. Особливості застосування аналітичних функцій ORACLE

2.3.1. Огляд аналітичних функцій SQL Oracle

Загальні відомості про аналітичні функції

Починаючи з версії СУБД Oracle 8.1.6 з'являється новий клас із 26 функцій, що мають назву аналітичні, та отримали подальший розвиток у версії 9. На відміну від звичайних скалярних функцій аналітичні функції беруть за аргумент SQL-таблицю, що являє собою логічний проміжний результат обробки SQL-оператора, де використано звернення до такої функції, і повертають як свій результат, зазвичай, теж SQL-таблицю.

Мета введення аналітичних функцій в Oracle

Технічна мета введення аналітичних функцій – дати лаконічне формулювання і прискорити виконання «аналітичних запитів» до БД, тобто запитів, які мають на меті виявлення внутрішніх співвідношень і залежностей між даними. Використання аналітичних функцій може дати такі переваги перед звичайними SQL-операторами:

– **Лаконічне й просте формулювання**. Багато аналітичних запитів до БД традиційними засобами складно формулюються, а тому важко осмислюються.

– **Зниження навантаження на мережу.** Те, що раніше могло формулювати тільки серію запитів, згортається в один запит. По мережі тільки відправляється запит і виходить остаточний результат.

– **Перенесення обчислень на сервер**. З використанням аналітичних функцій немає потреби організовувати розрахунки на ПК клієнта; вони повністю проводяться на сервері, ресурси якого більш реалізовані для швидкої обробки великих обсягів даних.

– **Кращу ефективність обробки запитів.** Аналітичні функції мають алгоритми обчислення, нерозривно пов'язані зі спеціальними планами обробки запитів, оптимізованими для більшої швидкості отримання результату.

Стратегічна мета введення в Oracle аналітичних функцій – дати базовий засіб для побудови інформаційних систем «аналітичного характеру» або OLAP-систем. За інформацією розробників, набір таких базових засобів, крім аналітичних функцій, формують ще й інші засоби Oracle, такі як:

– конструкції ROLLUP, CUBE та пов'язані з ними в пропозиціях з GROUP BY;

– матеріалізовані таблиці (materialized views).

Класифікація видів аналітичних функцій в Oracle.

Згідно з класифікацією документації з Oracle аналітичні функції можуть бути таких видів:

– функції ранжування – дозволяють «роздати» рядками «місця» залежно від наявних у них значень;

– статистичні функції для плаваючого інтервалу функції підрахунку частки – дозволяють одній SQL-операції отримати для кожного рядка її «питому вагу» в таблиці відповідно до її значень;

– статистичні функції LAG / LEAD;

– статистичні функції (лінійна регресія тощо).

Основні технічні особливості

Аналітичні функції використовують як аргумент стовпець проміжного результату обчислення SQL-запиту і повертає теж стовпець. Тому місцем їхнього використання в SQL-запиті можуть бути тільки фрази ORDER BY і SELECT, що виконують завершальну обробку логічного проміжного результату.

Синтаксис аналітичних функцій

Синтаксис виклику аналітичної функції має таку форму:

```
ІМ'Я_ФУНКЦІЇ (<аргумент>, <аргумент>, ...)
OVER (<конструкція_фрагментації>
   <конструкція_впорядкування>
   <конструкція_вікна>)
```

Виклик аналітичної функції може містити до чотирьох частин: аргументи, конструкція фрагментації, конструкція впорядкування та конструкція, що задає вікно.

Таблиця 2.3

Опис аналітичних функцій Oracle

AVG([DISTINCT | ALL] вираз)	Використовується для обчислення середнього значення виразу в межах групи і вікна. Для пошуку середнього після видалення значень, що дублюються, можна указувати ключове слово DISTINCT
CORR (вираз, вираз)	Видає коефіцієнт кореляції для пари виразів, що повертають числові значення. Це скорочення для виразу: *COVAR_POP<вираз1, вираз2) / STDDEV_POP(вираз1)* STDDEV_POP(вираз2)).* У статистичному сенсі, кореляція – це ступінь зв'язку між змінними. Зв'язок між змінними означає, що значення однієї змінної можна певною мірою передбачити за значенням іншої. Коефіцієнт кореляції представляє ступінь кореляції у вигляді числа в діапазоні від -1 (висока зворотна кореляція) до 1 (висока кореляція). Значення «0» відповідає відсутності кореляції
COUNT([DISTINCT] [*] [вираз])	Ця функція рахує рядки в групах. Якщо вказати * або будь-яку константу, крім NULL, функція COUNT рахуватиме всі рядки. Якщо вказати вираз, функція count рахуватиме рядки, для яких вираз має значення не NULL. Можна задавати модифікатор DISTINCT, щоб рахувати рядки в групах після видалення рядків, що дублюються
COVAR_POP(вираз, вираз)	Повертає коваріацію генеральної сукупності (population covariance) пари виразів з числовими значеннями
COVAR_SAMP (вираз, вираз)	Повертає вибіркову коваріацію (sample covariance) (вираз, вираз) пари виразів з числовими значеннями

Продовження табл. 2.3

CUME_DIST	Обчислює відносну позицію рядка в групі. Функція CUME_DIST завжди повертає число більше 0 і менше або рівне 1. Це число являє собою «позицію» рядка в групі з N арок. У групі з трьох рядків, наприклад, повертаються наступні значення кумулятивного розподілу: 1/3, 2/3 і 3/3
DENSE_RANK	Ця функція обчислює відносний ранг кожного рядка, що повертається запитом, відносно інших рядків, ґрунтуючись на значеннях виразів у конструкції ORDER BY. Дані в групі сортуються відповідно до конструкції ORDER BY, а потім кожному рядку по черзі привласнюється числовий ранг, починаючи з 1. Ранг збільшується при кожній зміні значень виразів, що входять у конструкцію ORDER BY. Рядки з однаковими значеннями отримують один і той самий ранг (при цьому порівнянні значення NULL вважаються однаковими). «Щільний» ранг, що повертається цією функцією, дає рангові значення без проміжків. Порівняйте її з представленою далі функцією RANK
FIRST_VALUE	Повертає перше значення у групі
LAG(вираз,< зсув >, <стандартне значення>)	Функція LAG дає доступ до інших рядків результуючої множини, позбавляючи від необхідності виконувати самоз'єднання. Вона дозволяє працювати з курсором як з масивом. Можна посилатися на рядки, передуючі поточному рядку в групі. Про те, як звертатися до наступних рядків в групі, див. в описі функції LEAD. Зсув – це позитивне ціле число із стандартним значенням 1 (попередній рядок). Стандартне значення повертається, якщо індекс виходить за межі вікна (для першого рядка групи буде повернено стандартне значення)

Продовження табл. 2.3

LAST_VALUE	Повертає останнє значення в групі
LEAD(вираз, <зсув>, <стандартне значення>)	Функція LEAD протилежна функції LAG. Якщо функція LAG дає доступ до попередніх рядків групи, то функція LEAD дозволяє звертатися до рядків, наступних за поточною. Зсув – це позитивне ціле число із стандартним значенням 1 (наступний рядок). Стандартне значення повертається, якщо індекс виходить за межі вікна (для останнього рядка групи буде повернено стандартне значення)
MAX(вираз)	Знаходить максимальне значення виразу в межах вікна в групі
MIN(вираз)	Знаходить мінімальне значення виразу в межах вікна в групі
NTILE(вираз)	Ділить групу на фрагменти щодо значення виразу. Наприклад, якщо вираз = 4, то кожному рядку в групі привласнюється число від 1 до 4 відповідно до фрагмента, з якого вона потрапляє. Якщо в групі 20 рядків, перші 5 набудуть значення 1, наступні 5 – значення 2 і так далі. Якщо кількість рядків в групі не ділиться на значення виразу без залишку, рядки розподіляють так, що ні в одному фрагменті кількість рядків не перевершує мінімальну кількість в інших фрагментах більш ніж на 1, причому додаткові рядки будуть в групах з меншими номерами фрагмента. Наприклад, якщо знову вираз = 4, а кількість рядків = 21, в першому фрагменті буде 6 рядків, в другому і подальших – 5
PERCENT RANK	Аналогічна функція CUME_DIST (кумулятивний розподіл). Обчислює ранг рядка в групі мінус 1, ділений на кількість оброблюваних рядків мінус 1. Ця функція завжди повертає значення в діапазоні від 0 до 1 включно

Продовження табл. 2.3

RANK	Ця функція обчислює відносний ранг кожного рядка, що повертається запитом, на основі значень виразів, що входять у конструкцію ORDER B Y. Дані в групі групуються відповідно до конструкції ORDER BY, а потім кожному рядку по черзі привласнюється числовий ранг, починаючи з 1. Рядки з однаковими значеннями виразів, що входять в конструкцію ORDER BY, отримують однаковий ранг, але якщо два рядки отримають однаковий ранг, наступне значення рангу пропускається. Якщо два рядки отримали ранг 1, рядків з рангом 2 не буде; наступний рядок в групі отримає ранг 3. У цьому відмінність від функції DENSE_RANK, яка не пропускає значень
RATIO_TO_REPORT (вираз)	Ця функція обчислює значення виразу (sum(вираз)) по рядках групи. Це дає відсоток, який складає значення поточного рядка щодо sum(вираз)
REGR_xxxxxxx (вираз, вираз)	Ці функції лінійної регресії застосовують стандартну (вираз, вираз) лінійну регресію за методом найменших квадратів до пари виразів. Пропонується 9 різних функцій регресії
ROW NUMBER	Повертає зсув рядка відносно початку впорядкованої групи. Може використовуватися для послідовної нумерації рядків, впорядкованих за певними критеріями
STDDEV(вираз)	Обчислює стандартне (середньо квадратичне) відхилення (standard deviation) поточного рядка відносно групи
STDDEV_POP (вираз)	Ця функція обчислює стандартне відхилення генеральної сукупності (population standard deviation) і повертає квадратний корінь з дисперсії генеральної сукупності (population variance). Вона повертає значення, що збігається з квадратним коренем з результату функції VAR_POP

Продовження табл. 2.3

STDDEV_SAMP (вираз)	Ця функція обчислює накопичене стандартне відхилення вибірки (cumulative sample standard deviation) і повертає квадратний корінь вибіркової дисперсії (sample variance). Вона повертає значення, що збігається з квадратним коренем з результату функції VAR_SAMP
SUM (вираз)	Обчислює загальну суму значень виразу для групи
VAR_TOP (вираз)	Ця функція повертає дисперсію генеральної сукупності для набору числових значень (значення NULL ігноруються). Функція VAR_POP обчислює значення: *(SUM(вираз * вираз) – SUM(вираз)*SUM(вираз) / COUNT(вираз)) /COUNT(вираз)*
VAR_SAMP(вираз)	Ця функція повертає вибіркову дисперсію для набору числових значень (значення NULL ігноруються). Вона обчислює значення: *(SUM(вираз * вираз) – SUM(вираз)*SUM(вираз) / COUNT(вираз)) / (COUNT(вираз) - 1)*
VARIANCE(вираз)	Повертає дисперсію . Обчислює дисперсію як: 0, якщо кількість рядків в групі = 1; VAR_SAMP, якщо кількість рядків у групі > 1

Приклад

SUM (sal) OVER (PARTITION BY deptno
ORDER BY ename) department_total,
де, SUM – ім'я функції.

(SAL) – аргумент аналітичної функції. Аналітичні функції приймають від нуля до трьох аргументів. Як аргументи, передаються вираження, тобто цілком можна було б використовувати SUM(SAL+COMM).

OVER – ключове слово, що ідентифікує цю функцію як аналітичну. Без цього ідентифікатора синтаксичний аналізатор не міг би відрізнити функцію агрегування SUM() від аналітичної функції SUM(). Конструкція після ключового слова OVER описує зріз даних, за яким буде обчислюватися аналітична функція.

PARTITION BY – необов'язкова конструкція фрагментації. Якщо конструкція фрагментації не задана, уся результуюча множина вважається одним великим фрагментом. Це використовується для розподілу результуючої множини на групи, так що аналітична функція застосовується до груп, а не до всієї результуючої безлічі.

ORDER BY – необов'язкова конструкція ORDER BY; для деяких функцій вона обов'язкова, для інших – ні. Функції, що залежать від впорядкування даних, наприклад LAG і LEAD, які дозволяють звернутися до попередніх і наступних рядків у результуючій множині, вимагають обов'язкової вказівки на конструкцію **ORDER BY.** Інші функції, наприклад AVG, не вимагають. Ця конструкція обов'язкова, якщо використовується будь-яка функція роботи з вікном. Конструкція ORDER BY визначає, як впорядковуються дані в групах під час обчислення аналітичної функції.

Конструкція фрагментації

Конструкція **PARTITION BY** логічно розподіляє результуючу множину на N груп за критеріями, що задаються виразами фрагментації. Слова «фрагмент» і «група» у цьому документі й у документації Oracle використовуються як синоніми. Аналітичні функції застосовуються до кожної групи незалежно – для кожної нової групи вони скидаються. Якщо не вказати конструкцію фрагментації, уся результуюча множина вважається однією групою.

Синтаксис конструкції фрагментації схожий на синтаксис конструкції

GROUP BY у звичайних SQL-Запитах:
PARTITION BY вираз [, вираз] [, вираз]

Конструкція впорядкування

Конструкція **ORDER BY** задає критерій групування даних у кожній групі (в кожному фрагменті). Це, без сумніву, впливає на результат виконання будь-якої аналітичної функції. За наявності (або відсутності) конструкції ORDER BY аналітичні функції обчислюються по-різному. Розглянемо приклади для обчислення середнього значення AVG із конструкцією ORDER BY і без неї.

За відсутності конструкції // ORDER BY середнє значення обчислюється по цілій групі, і те саме значення видається для кожного рядка (функція використовується як підсумкова).

Приклад 2.31

SELECT ename, sal, AVG(sal) OVER () FROM emp

	ENAME	SAL	ROUND(AVG(SAL)OVER())
1	SMITH	800	2073
2	ALLEN	1600	2073
3	WARD	1250	2073
4	JONES	2975	2073
5	MARTIN	1250	2073
6	BLAKE	2850	2073
7	CLARK	2450	2073
8	SCOTT	3000	2073
9	KING	5000	2073
10	TURNER	1500	2073
11	ADAMS	1100	2073
12	JAMES	950	2073
13	FORD	3000	2073
14	MILLER	1300	2073

Рис. 2.4. Результат виконання запиту, приклад 2.31

Коли функція AVG() використовується з конструкцією ORDER BY, середнє значення в кожному рядку є середнім по поточному і всіх попередніх рядках (функція використовується як віконна):

Приклад 2.32

```
SELECT ename, sal, AVG(sal) OVER (ORDER BY ENAME))
FROM emp
ORDER BY ename
```

	ENAME	SAL	ROUND(AVG(SAL)OVER(ORDERBYENAME))
1	ADAMS	1100	1100
2	ALLEN	1600	1350
3	BLAKE	2850	1850
4	CLARK	2450	2000
5	FORD	3000	2200
6	JAMES	950	1992
7	JONES	2975	2132
8	KING	5000	2491
9	MARTIN	1250	2353
10	MILLER	1300	2248
11	SCOTT	3000	2316
12	SMITH	800	2190
13	TURNER	1500	2137
14	WARD	1250	2073

Рис. 2.5. Результат виконання запиту, приклад 2.32

Наприклад, середня зарплата для користувача ALLEN у результатах виконання запиту з конструкцією ORDER BY = 1350 (середнє для значень 1100 і 1600).

Формування запитів за допомогою конструкції «вікна»

Синтаксис цієї конструкції на перший погляд видається складним через ключові слова, що в ньому використовуються. Конструкція виду RANGE BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW**,** що задає стандартне вікно під час використання конструкції ORDER BY, не схожа на ті, що постійно використовуються розробниками. Синтаксис конструкції вікна досить складний для опису. Перерахуємо всі варіанти конструкції вікна й опишемо, який набір даних у межах групи задає відповідний варіант.

Конструкція вікна дозволяє задати вікно, що переміщується або жорстко прив'язане до даних у межах групи, з яким буде працювати аналітична функція.

Наприклад, конструкція діапазону RANGE UNBOUNDED PRECEDING означає: «застосовувати аналітичну функцію до кожного рядка даної групи, з першої до поточної».

Приклад 2.33

Стандартним є жорстко прив'язане вікно, що починається з першого рядка групи і закінчується поточним. Якщо використовується така аналітична функція:

```
SUM(sal) OVER
(PARTITION BY deptno
ORDER BY ename
ROWS 2 PRECEDING) department_total2,
```

тоді буде створене вікно, що переміщається, у групі, і сума зарплат буде обчислюватися по стовпцю SAL поточному й двох попередніх рядках у цій групі. Якщо необхідно створити звіт, що показує суму зарплат поточного й двох попередніх співробітників відділу, сценарій може виглядати так:

Приклад 2.34

```
BREAK ON deptno
SELECT deptno, ename, sal,
SUM(sal) OVER (PARTITION BY deptno
ORDER BY ename rows 2 preceding) sliding_total
FROM emp
ORDER by deptno, ename
Нас цікавить ця частина запиту:
SUM(sal) OVER (PARTITION BY deptno ORDER BY ename
ROWS 2 PRECEDING) sliding_total
```

Конструкція, що визначає фрагментацію, приводить до обчислення SUM(SAL) по відділах, незалежно від інших груп (значення SUM(SAL) скидається при зміні номера відділу). Конструкція ORDER BY ENAME приводить до сортування даних у кожному відділі по стовпцю ENAME; це дозволяє за допомогою конструкції вікна, **rows 2 preceding**, при підсумовуванні зарплат звертатися до двох попередніх рядків відповідно до заданого порядку сортування. Наприклад, значення в стовпці SLIDING_TOTAL для співробітника SMITH = 6775, що дорівнює сумі значень 800, 3000 і 2975. Це сума зарплат у рядку для SMITH і двох попередніх рядках вікна.

	DEPTNO	ENAME	SAL	SLIDING_TOTAL
1	10	CLARK	2450	2450
2	10	KING	5000	7450
3	10	MILLER	1300	8750
4	20	ADAMS	1100	1100
5	20	FORD	3000	4100
6	20	JONES	2975	7075
7	20	SCOTT	3000	8975
8	20	SMITH	800	6775
9	30	ALLEN	1600	1600
10	30	BLAKE	2850	4450
11	30	JAMES	950	5400
12	30	MARTIN	1250	5050
13	30	TURNER	1500	3700
14	30	WARD	1250	4000

Рис. 2.6. Результат виконання запиту, приклад 2.34

Можна створювати вікна за двома критеріями: за діапазоном (RANGE) значень даних або по зсуву (ROWS) щодо поточного рядка. Конструкція RANGE уже зустрічалася раніше, RANGE UNBOUNDED PRECEDING. Вона вимагає брати всі рядки аж до поточного, відповідно до порядку, що задається конструкцією ORDER BY. Потрібно пам’ятати, що для використання вікон необхідно задавати конструкцію ORDER BY. Зараз розглянемо завдання вікон за допомогою конструкцій ROWS і RANGE та інші способи завдання вікон.

Формування запитів у форматі «Вікна діапазону»

Вікна діапазону об'єднують рядки відповідно до заданого порядку. Якщо умова запиту, наприклад, «range 5 preceding», то буде сгенероване вікно, що переміщається, включає попередні рядки групи, що відстають від поточного рядка не більше ніж на 5 рядків. Діапазон можна задавати у вигляді числового виразу або виразу, значенням якого є дата. Застосовувати конструкцію RANGE з іншими типами даних не можна.

Якщо є таблиця EMP зі стовпцем HIREDATE типу дати й задана аналітична функція

COUNT(*) OVER (ORDER BY HIREDATE ASC RANGE 100 PRECEDING),

вона знайде всі попередні рядки фрагмента, значення яких у стовпці HIREDATE лежить у межах 100 днів від значення HIREDATE поточного рядка. У цьому випадку, оскільки дані сортуються за зростанням (ASC), значення у вікні будуть включати всі рядки поточної групи, у яких значення в стовпці HIREDATE менше значення HIREDATE поточного рядка, але не більше ніж на 100 днів. Якщо використовувати функцію

COUNT(*) OVER (ORDER BY HIREDATE DESC RANGE 100 PRECEDING)

і сортувати фрагмент за спаданням (DESC), базова логіка роботи залишиться тією ж, але, оскільки група відсортована інакше, у вікно потрапить інший набір рядків. У розглянутому випадку функція знайде всі рядки, що передують поточному, де значення в полі *HIREDATE* більше значення *HIREDATE* у поточному рядку, але не більше ніж на 100 днів. Наприклад, під час використання, запиту з аналітичною функцією *FIRST_VALUE*. Ця функція повертає обчислене значення для першого рядка вікна. Так можна зрозуміти, де починається вікно:

Приклад 2.35

```
SELECT ename, sal, HIREDATE, HIREDATE-100
WINDOWTOP,
FIRST_VALUE (ename)
OVER (ORDER BY HIREDATE ASC
RANGE 100 PRECEDING) ename_prec,
FIRST_VALUE(hiredate)
OVER (ORDER BY HIREDATE ASC
RANGE 100 PRECEDING) hiredate_prec
FROM emp
ORDER BY HIREDATE ASC
```

	ENAME	SAL	HIREDATE	WINDOWTOP	ENAME_...	HIREDATE_PREC
1	SMITH	800	17.12.80	08.09.80	SMITH	17.12.80
2	ALLEN	1600	20.02.81	12.11.80	SMITH	17.12.80
3	WARD	1250	22.02.81	14.11.80	SMITH	17.12.80
4	JONES	2975	02.04.81	23.12.80	ALLEN	20.02.81
5	BLAKE	2850	01.05.81	21.01.81	ALLEN	20.02.81
6	CLARK	2450	09.06.81	01.03.81	JONES	02.04.81
7	TURNER	1500	08.09.81	31.05.81	CLARK	09.06.81
8	MARTIN	1250	28.09.81	20.06.81	TURNER	08.09.81
9	KING	5000	17.11.81	09.08.81	TURNER	08.09.81
10	FORD	3000	03.12.81	25.08.81	TURNER	08.09.81
11	JAMES	950	03.12.81	25.08.81	TURNER	08.09.81
12	MILLER	1300	23.01.82	15.10.81	KING	17.11.81
13	SCOTT	3000	19.04.87	09.01.87	SCOTT	19.04.87
14	ADAMS	1100	23.05.87	12.02.87	SCOTT	19.04.87

Рис. 2.7. Результат виконання запиту, приклад 2.35

Впорядкували один фрагмент за критерієм HIREDATE ASC. При цьому використовувалася аналітична функція FIRST_VALUE для пошуку першого значення ENAME і першого значення HIREDATE у відповідному вікні. Подивившись на рядок даних для співробітника CLARK, можна виявити, що для нього значення в стовпці HIREDATE = 9 червня 1981 року, 09-JUN-81, а дата за 100 днів до цієї відповідає 1 березня 1981 року, 01-MAR-81. Для зручності ця дата поміщена в стовпець WINDOWTOP. Аналітична функція потім обчислюється для всіх рядків відсортованого

фрагмента, що передує рядку для співробітника CLARK, і що має значення в стовпці HIREDATE у діапазоні з 01-MAR-81 по 09-JUN-81. Перше значення ENAME для цього вікна – JONES. Це ім'я й видає аналітична функція в стовпці ENAME_PREC.

Упорядкувавши дані за критерієм HIREDATE DESC, одержимо:

Приклад 2.36

```
SELECT ename, sal, HIREDATE, HIREDATE+100
WINDOWTOP,
FIRST_VALUE(ename)
OVER (ORDER BY HIREDATE DESC
RANGE 100 PRECEDING) ename_prec,
FIRST_VALUE(hiredate)
OVER (ORDER BY HIREDATE DESC
RANGE 100 PRECEDING) hiredate_prec
FROM emp
ORDER BY HIREDATE DESC
```

	ENAME	SAL	HIREDATE	WINDOWTOP	ENAME_PREC	HIREDATE_PREC
1	ADAMS	1100	23.05.87	31.08.87	ADAMS	23.05.87
2	SCOTT	3000	19.04.87	28.07.87	ADAMS	23.05.87
3	MILLER	1300	23.01.82	03.05.82	MILLER	23.01.82
4	FORD	3000	03.12.81	13.03.82	MILLER	23.01.82
5	JAMES	950	03.12.81	13.03.82	MILLER	23.01.82
6	KING	5000	17.11.81	25.02.82	MILLER	23.01.82
7	MARTIN	1250	28.09.81	06.01.82	FORD	03.12.81
8	TURNER	1500	08.09.81	17.12.81	FORD	03.12.81
9	CLARK	2450	09.06.81	17.09.81	TURNER	08.09.81
10	BLAKE	2850	01.05.81	09.08.81	CLARK	09.06.81
11	JONES	2975	02.04.81	11.07.81	CLARK	09.06.81
12	WARD	1250	22.02.81	02.06.81	BLAKE	01.05.81
13	ALLEN	1600	20.02.81	31.05.81	BLAKE	01.05.81
14	SMITH	800	17.12.80	27.03.81	WARD	22.02.81

Рис. 2.8. Результат виконання запиту, приклад 2.36

Якщо проаналізувати рядок співробітника CLARK, з'ясуємо, що обрано інше вікно, оскільки дані фрагмента розподілені по-іншому. Вікно для рядка CLARK за умовою RANGE 100

PRECEDING тепер доходить до рядка TURNER, оскільки значення HIREDATE для TURNER – остання дата серед значень HIREDATE у рядках, що передують рядку CLARK, що відрізняється не більше ніж на 100 днів. Іноді досить складно зрозуміти, які значення будуть входити в діапазон.

Використання функції FIRST_VALUE є зручним методом, що допомагає побачити діапазони вікна й перевірити, чи коректно встановлені параметри. Тепер, представивши діапазони вікон, використовуємо їх для обчислення чогось більш істотного.

Нехай необхідно вибрати зарплату кожного співробітника й середню зарплату всіх прийнятих на роботу протягом 100 попередніх днів, а також середню зарплату всіх прийнятих на роботу протягом 100 наступних днів. Відповідний запит буде мати такий вигляд:

Приклад 2.37

```
SELECT ename, hiredate, sal,
AVG(sal)
OVER (ORDER BY HIREDATE ASC RANGE 10 0
PRECEDING)
avg_sal_l00_days_before,
AVG(sal)
OVER (ORDER BY HIREDATE DESC RANGE 100
PRECEDING)
avg_sal_100_days_after
FROM emp
```

Якщо знову звернутися до рядка для співробітника CLARK, тоді, оскільки вже розуміємо, яке вікно в групі буде з нею позв'язане, легко переконатися, що середня зарплата (2758,33) становить (2450+2850+2975)/3. Це середня зарплата для рядка CLARK і рядків, що передують CLARK (це рядки для співробітників JONES і BLAKE) під час впорядкування даних за зростанням.

З іншого боку, середня зарплата 1975,00 становить (2450+1500)/2. Це середня зарплата для рядка CLARK і рядків, що передують CLARK при впорядкуванні даних за спаданням. За

допомогою цього запиту можна одночасно обчислити середню зарплату для співробітників, прийнятих на роботу за 100 днів до й за 100 днів після співробітника CLARK.

	ENAME	HIREDATE	SAL	AVG_SAL_L00_DAYS_BEFORE	AVG_SAL_100_DAYS_AFTER
1	ADAMS	23.05.87	1100	2050	1100
2	SCOTT	19.04.87	3000	3000	2050
3	MILLER	23.01.82	1300	2563	1300
4	FORD	03.12.81	3000	2340	1750
5	JAMES	03.12.81	950	2340	1750
6	KING	17.11.81	5000	2583	2563
7	MARTIN	28.09.81	1250	1375	2550
8	TURNER	08.09.81	1500	1975	2340
9	CLARK	09.06.81	2450	2758	1975
10	BLAKE	01.05.81	2850	2169	2650
11	JONES	02.04.81	2975	1942	2758
12	WARD	22.02.81	1250	1217	2358
13	ALLEN	20.02.81	1600	1200	2169
14	SMITH	17.12.80	800	800	1217

Рис. 2.9. Результат виконання запиту, приклад 2.37

Вікна діапазону можна задавати тільки за даними типу NUMBER або DATE, оскільки не можна додати або відняти N одиниць зі значення типу VARCHAR2. Ще одне обмеження для таких вікон полягає у тому, що в конструкції ORDER BY може бути тільки один стовпець – діапазони за своєю природою одномірні. Не можна задати діапазон в N-вимірному просторі.

Вікна рядків

Вікна рядків задаються у фізичних одиницях, рядках. Розглянемо такий приклад, задавши вікно рядків:

COUNT (*) OVER (ORDER BY x ROWS 5 PRECEDING)

Це вікно буде включати до 6 рядків: поточний і п’ять попередніх (порядок визначається конструкцією ORDER BY). Для вікон по рядках немає обмежень, властивих вікнам по діапазону; дані можуть бути будь-якого типу й упорядковувати можна за будь-якою кількістю стовпців (приклад 2.38).

Приклад 2.38

```
SELECT ename, sal, hiredate,
FIRST_VALUE(ename)
OVER (ORDER BY HIREDATE ASC
ROWS 5 PRECEDING) ename_prec,
FIRST_VALUE(hiredate)
OVER (ORDER BY HIREDATE ASC
ROWS 5 PRECEDING) HIREDATE_PREC
FROM emp
ORDER BY HIREDATE ASC
```

	ENAME	SAL	HIREDATE	ENAME_PREC	HIREDATE_PREC
1	SMITH	800	17.12.80	SMITH	17.12.80
2	ALLEN	1600	20.02.81	SMITH	17.12.80
3	WARD	1250	22.02.81	SMITH	17.12.80
4	JONES	2975	02.04.81	SMITH	17.12.80
5	BLAKE	2850	01.05.81	SMITH	17.12.80
6	CLARK	2450	09.06.81	SMITH	17.12.80
7	TURNER	1500	08.09.81	ALLEN	20.02.81
8	MARTIN	1250	28.09.81	WARD	22.02.81
9	KING	5000	17.11.81	JONES	02.04.81
10	JAMES	950	03.12.81	BLAKE	01.05.81
11	FORD	3000	03.12.81	CLARK	09.06.81
12	MILLER	1300	23.01.82	TURNER	08.09.81
13	SCOTT	3000	19.04.87	MARTIN	28.09.81
14	ADAMS	1100	23.05.87	KING	17.11.81

Рис. 2.10. Результат виконання запиту, приклад 2.38

Розглянувши рядок для співробітника CLARK (рис. 2.10), можна побачити, що першим у вікні ROWS 5 PRECEDING йде рядок для співробітника SMITH – він просто п'ятий перед рядком для CLARK відповідно до заданого порядку. Рядок для співробітника SMITH буде першим у вікні й для всіх попередніх рядків (для BLAKE, JONES тощо). Рядок для SMITH – перший у цій групі (він буде перший і для самого себе). Під час сортування групи за зростанням вікна змінюються (приклад 2.39).

Приклад 2.39

```
SELECT ename, sal, hiredate,
FIRST_VALUE(ename)
OVER (ORDER BY HIREDATE DESC
ROWS 5 PRECEDING) ename_prec,
FIRST_VALUE(hiredate)
OVER (ORDER BY HIREDATE DESC
ROWS 5 PRECEDING) hiredate_prec
FROM emp
ORDER BY HIREDATE DESC
```

	ENAME	SAL	HIREDATE	ENAME_PREC	HIREDATE_PREC
1	ADAMS	1100	23.05.87	ADAMS	23.05.87
2	SCOTT	3000	19.04.87	ADAMS	23.05.87
3	MILLER	1300	23.01.82	ADAMS	23.05.87
4	FORD	3000	03.12.81	ADAMS	23.05.87
5	JAMES	950	03.12.81	ADAMS	23.05.87
6	KING	5000	17.11.81	ADAMS	23.05.87
7	MARTIN	1250	28.09.81	SCOTT	19.04.87
8	TURNER	1500	08.09.81	MILLER	23.01.82
9	CLARK	2450	09.06.81	FORD	03.12.81
10	BLAKE	2850	01.05.81	JAMES	03.12.81
11	JONES	2975	02.04.81	KING	17.11.81
12	WARD	1250	22.02.81	MARTIN	28.09.81
13	ALLEN	1600	20.02.81	TURNER	08.09.81
14	SMITH	800	17.12.80	CLARK	09.06.81

Рис. 2.11. Результат виконання запиту, приклад 2.39

Перше значення для набору з 5 рядків, що передують у групі рядка для співробітника CLARK, – рядок для співробітника JAMES. Тепер можна обчислити середню зарплату для зазначеного співробітника й п’яти прийнятих на роботу до нього й після нього.

Зверніть увагу, вибрано також значення COUNT(*). Це дозволяє зрозуміти, за якою кількістю рядків було обчислено середнє значення. Можна чітко побачити, що для обчислення середньої зарплати співробітників, прийнятих до співробітника ALLEN, використовувалося тільки 2 записи, а для обчислення середньої зарплати співробітників, найнятих після нього, – 6. У тому місці

групи, де перебуває рядок для співробітника ALLEN, є тільки 1 попередній запис, а під час обчислення аналітичної функції використовуються усі наявні рядки.

```
SELECT ename, hiredate, sal, AVG(sal)
OVER (ORDER BY HIREDATE ASC ROWS 5 PRECEDING)
avg_5_before, COUNT(*)
OVER (ORDER BY HIREDATE ASC ROWS 5 PRECEDING)
obs_before, AVG(sal)
OVER (ORDER BY HIREDATE DESC ROWS 5 PRECEDING)
avg_5_after, COUNT(*)
OVER (ORDER BY HIREDATE DESC ROWS 5
 PRECEDING) obs_after
FROM emp
ORDER BY HIREDATE
```

	ENAME	HIREDATE	SAL	AVG_5_BEFORE	OBS_BEFORE	AVG_5_AFTER	OBS_AFTER
1	SMITH	17.12.80	800	800	1	1988	6
2	ALLEN	20.02.81	1600	1200	2	2104	6
3	WARD	22.02.81	1250	1217	3	2046	6
4	JONES	02.04.81	2975	1656	4	2671	6
5	BLAKE	01.05.81	2850	1895	5	2333	6
6	CLARK	09.06.81	2450	1988	6	2358	6
7	TURNER	08.09.81	1500	2104	6	2167	6
8	MARTIN	28.09.81	1250	2046	6	2417	6
9	KING	17.11.81	5000	2671	6	2392	6
10	JAMES	03.12.81	950	2333	6	1870	5
11	FORD	03.12.81	3000	2358	6	2100	4
12	MILLER	23.01.82	1300	2167	6	1800	3
13	SCOTT	19.04.87	3000	2417	6	2050	2
14	ADAMS	23.05.87	1100	2392	6	1100	1

Рис. 2.12. Результат виконання запиту

2.3.2. Формування запитів у конструкції «вікна»

Розуміючи відмінність між вікнами діапазонів і вікнами рядків, можна вивчати способи їх задавання. У найпростішому випадку, вікно задається за допомогою однієї із трьох наступних взаємовиключних конструкцій.

UNBOUNDED PRECEDING – вікно починається з першого рядка поточної групи й закінчується поточним рядком, що обробляється.

CURRENT ROW – вікно починається (і закінчується) поточним рядком.

Числовий_вираз PRECEDING – вікно починається з рядка за числовим_виразом рядків до поточного, якщо він задається по рядках, або з рядка, меншого за значенням стовпця, згаданого в конструкції ORDER BY, але не більше ніж на числовий вираз, якщо він задається по діапазону.

Вікно CURRENT ROW у найпростішому вигляді ніколи не використовується, оскільки обмежує застосування аналітичної функції одним поточним рядком, а для цього аналітичні функції не потрібні. У більш складному випадку для вікна задається також конструкція BETWEEN. У ній CURRENT ROW можна вказувати як початковий або кінцевий рядок вікна. Початковий і кінцевий рядок вікна в конструкції BETWEEN можна задавати з використанням кожної з перерахованих вище конструкцій і ще однієї додаткової.

Числовий_вираз FOLLOWING. Вікно закінчується (або починається) з рядка, через числовий_вираз рядків після поточного, якщо він задасться по рядках, або з рядка, більшого за значенням стовпця, згаданого в конструкції ORDER BY, але не більше ніж на числовий_вираз, якщо він задається по діапазону.

Як видно, вікно RANGE 100 PRECEDING містить тільки рядки поточного фрагмента, що передує поточному рядку, і ті, значення яких HIREDATE знаходяться в діапазоні HIREDATE-100 і HIREDATE щодо поточного. У цьому випадку таких рядків завжди 1 або 2, тобто інтервал між прийманням людей на роботу звичайно перевищує 100 днів (за винятком двох випадків). Вікно ROWS 2 PRECEDING містить від 1 до 3 рядків (це визначається тим як далеко поточний рядок знаходиться від початку групи). Для першого рядка групи маємо значення 1 (попередніх рядків немає).

Для наступного рядка в групі таких рядків 2. Нарешті, для третього і наступних рядків значення COUNT(*) залишається постійним, оскільки ми вважаємо тільки поточний рядок і два попередніх.

Тепер потрібно розглянути використання конструкції BETWEEN. Усі задані дотепер вікна закінчувалися поточним рядком і верталися по результуючій безлічі в пошуках додаткової інформації.

Звернувшись знову до рядка для співробітника CLARK, можна переконатися, що вікно починається з рядка для JONES і триває до рядка для співробітника TURNER. Тепер у вікно входять рядки для тих, хто прийнятий на роботу за 100 днів до й (а не або, як у попередніх прикладах) після поточного співробітника.

Отже, синтаксис чотирьох компонентів виклику аналітичних функцій – це:

– ім’я функції;

– конструкція фрагментації, що використовується для розподілу результуючої множини на незалежні групи;

– конструкція ORDER BY, що сортує дані в групі для віконних функцій;

– конструкція вікна, що задає набір рядків, до яких застосовується аналітична функція.

2.3.3. Використання метасимволів для підготовки запитів Select

Використання логічного оператора LIKE

Усі попередні оператори, що були розглянуті, здійснювали фільтрацію за відомими значеннями. Вони шукали збігу за одним або декількома значеннями. При цьому скрізь шукали відоме значення. Однак фільтрація даних таким способом не завжди працює.

Наприклад, якби ви шукали інформацію, у назві яких містяться слова «сільська рада»? Цього не можна зробити за допомогою простих операцій порівняння. У такому випадку на допомогу приходить пошук з використанням метасимволів. За допомогою метасимволів можна створювати умови пошуку даних.

У цьому прикладі для того, щоб знайти рядки, у назві яких містяться слова «сільська рада», необхідно скласти шаблон пошуку, що дозволяє знайти даний текст у будь-якому місці стовпця таблиці.

Метасимволи самі по собі є символами, які мають в умові WHERE спеціальне значення. У SQL підтримуються метасимволи декількох типів. Для того, щоб застосовувати метасимволи в

умовах пошуку, необхідно використовувати ключове слово LIKE. Воно повідомляє СУБД, що наступний шаблон для пошуку необхідно порівнювати з використанням метасимволів, а не шукати точний збіг.

Пошук з використанням метасимволів може здійснюватися тільки в текстових полях (рядках), не можна використовувати метасимволи при пошуку полів з нетекстовим типом даних.

Метасимвол знак відсотка (%)

Найчастіше використовується метасимвол – знак відсотка (%). У рядку пошуку «%» означає знайти всі входження будь-якого символу.

Приклад 2.40

Щоб знайти всі сільські ради, назви яких закінчуються "сільська рада", можна виконати такий запит:

```
SELECT name, kod
FROM aisr.accounts
WHERE name LIKE '%сільська рада';
```

Результат

```
Name        kod
-----------------------------------------------------
Обухівська сільська рада 2367958314
Маркушівська сільська рада     1593456135
```

У цьому прикладі використовується шаблон пошуку '%сільська рада'. При виконанні цієї умови повертаються всі значення, які закінчуються символами «сільська рада». Знак % вказує СУБД приймати всі символи перед словами "сільська рада" незалежно від їхньої кількості. Метасимволи можна використовувати в будь-якому місці шаблону пошуку, причому в необмеженій кількості.

Метасимвол можна також використовувати всередині шаблону пошуку, хоча це рідко буває корисним.

Важливо пам’ятати!

Крім пошуку одного або кількох символів, знак % також означає і відсутність символів у вказаному місці шаблону пошуку.

Метасимвол «символ підкреслення (_)»

Ще одним корисним метасимволом є символ підкреслення (_). Символ підкреслення використовується так само, як і %, але при цьому враховується не багато символів, а тільки один.

Приклад 2.41

Щоб знайти кількість платників податків, які здали декларацію з податку на додану вартість (юридичні особи) у травні 2015 року, (код форми J02001_ останній символ є змінним, оскільки змінюється форма декларації з ПДВ), можна виконати такий запит:

```
SELECT COUNT(distinct tin) AS pdvfdd
FROM bestzvit.zregdoc
WHERE period_year=2009 AND period_month=05 AND
c_doc||c_doc_sub||c_doc_ver like 'J02001_'
AND c_doc_stat NOT IN (5,6,7) AND c_doc_src IN (1,2,3)
AND period_type IN (1,2);
```

На відміну від знаку %, що має на увазі також відсутність символів, знак (_) завжди означає один символ – не більше і не менше.

Метасимвол «квадратні дужки» ([ ])

Метасимвол «квадратні дужки» ([ ]) використовується для зазначення набору символів, кожен з яких має збігатися зі значенням, причому точно у вказаному місці (в місцезнаходженні метасимвола).

Приклад 2.42

Щоб знайти всі назви підприємств, назви яких починаються на букву Т або І, необхідно зробити таке:

```
SELECT NAME, KOD FROM aisr.Accounts
WHERE NAME LIKE '[TI]%>
ORDER BY NAME;
```

Результат

NAME KOD

Туринське РАЙСТ 589632596

Іріпнське відділення 00292988

Транспортне агенство 17529638

Умова WHERE в цьому вираженні виглядає як '[TI]%'. У цьому шаблоні пошуку використовуються два різних метасимволи. За метасимволами [TI] здійснюється пошук усіх підприємств, назви яких починаються на одну із зазначених у дужках літер, але при цьому враховується тільки один символ. Тому всі імена більше одного символу не будуть вилучені. Застосував метасимвол %, наступний після [TI], проводимо пошук будь-якої кількості символів після першої літери, що і приводить до необхідного результату.

2.3.4. Використання статистичних аналітичних функцій

Часто буває необхідно просумувати дані без їхньої вибірки, і в SQL передбачені для цього спеціальні функції. SQL-запити з цими функціями часто використовуються з метою вибірки даних для аналізу та створення звітів. Прикладами таких вибірок можуть бути:

– визначення числа рядків у таблиці (або числа рядків, які задовольняють умову або містять певне значення);

– отримання суми по набору рядків у таблиці;

– пошук найбільшого, найменшого та середнього значень зі стовпця таблиці (з усіх або з якихось конкретних рядків).

У кожному з цих прикладів необхідні якісь підсумкові дані з таблиці, а не самі дані. Тому повернення реальних даних таблиці було б марною тратою часу та ресурсів. Отже, все, що потрібно, – це тільки підсумкова інформація.

Щоб полегшити такий спосіб отримання інформації, в SQL передбачений набір з п’яти статистичних функцій, які перераховані

в таблиці. Ці функції дозволяють виконувати всі варіанти вибірки, які були перераховані вище. На відміну від функцій маніпулювання даними з попереднього уроку, статистичні SQL функції підтримуються без особливих змін у більшості реалізацій SQL.

Статистичні (підсумкові) функції – функції, що обробляють набір рядків для підрахунку та повернення одного значення

Функція	Опис
AVG ()	Повертає середнє значення стовпця
COUNT ()	Повертає кількість рядків у стовпці
MAX ()	Повертає найбільше значення у стовпці
MIN ()	Повертає найменше значення у стовпці
SUM ()	Повертає суму значень стовпця

Функція AVG ()

Функцію **AVG()** використовують для повернення середнього значення певного стовпця шляхом підрахунку кількості рядків у таблиці і додавання їхніх значень. Цю функцію можна використовувати для повернення середнього значення всіх стовпців або певних стовпців або рядків.

Приклад 2.43

Функцію AVG () використовують для повернення середньої суми нарахованої пені для підприємства з кодом 00292988 таблиці T21PENOTTI:

```
SELECT AVG(ZN) AS avg_penia
FROM tax_I.T21Pennoti
WHERE TIN=00292988;
```

Результат

```
-----------------------------------
avg_penia 6.823333
```

Вираз SELECT, наведений вище, повертає одне значення, avg_price, в якому міститься середня ціна всіх продуктів таблиці Products. Тут avg_price – це псевдонім. Функцію AVG () можна також використовувати для знаходження середнього значення певних стовпців або рядків.

Приклад 2.44

У якому повертається середня нарахована заробітна плата по району згідно з даними **1ДФ:**

```
SELECT AVG(I_S_NAR) AS "Середня нарахована
зарплата"
FROM bestzvit.card8dr
WHERE PER_DATE = '01.03.2009';
```

Результат

```
------------------------------------------------------------------------
"Середня нарахована зарплата" 110085.128337838
```

Цей оператор SELECT відрізняється від попереднього лише тим, що в ньому міститься пропозиція WHERE. Відповідно до пропозиції WHERE вибирають тільки той період, значення PER_DATE для якого дорівнює '01.03.2009' (тобто 1 квартал 2009 року), тому значення, повернення в стовпці з псевдонімом "Середня нарахована зарплата" є середнім тільки для цього періоду.

Функцію AVG () можна використовувати тільки для обчислення середнього значення певного числового стовпця, ім’я цього стовпця повинно бути вказано як параметр функції. Щоб отримати середнє значення декількох стовпців, необхідно використовувати кілька функцій AVG ().

Важливо пам’ятати! Рядки стовпця, що містять значення NULL, ігноруються функцією AVG ().

Функція COUNT ()

Функція COUNT () підраховує кількість рядків. За допомогою функції COUNT () можна дізнатися загальне число рядків у таблиці або число рядків, які задовольняють певні критерії. Цю функцію можна використовувати двома способами:

– як COUNT (*) для підрахунку кількості рядків у таблиці незалежно від того, містять стовпці значення NULL чи ні;

– як COUNT (column) для підрахунку кількості рядків, які мають значення у вказаних стовпцях, причому значення NULL ігноруються.

Приклад 2.45

Повернути загальну кількість поданих декларацій з ПДВ юридичними особами за формою J0200108 у 2013 році, що містяться у таблиці ZREGDOC:

```
SELECT COUNT(*) AS PDVJJR
FROM ZREGDOC
WHERE C DOC = 'JO2'
AND C_DOC_SUB ='001'
AND C_DOC_VER =8
AND PERIOD_YEAR= 2 013;
```

Результат

```
---------------------------------------
PDV_UR 261
```

У цьому прикладі функцію COUNT (*) використовують для підрахунку всіх рядків незалежно від їхніх значень.

Важливо пам’ятати! Рядки стовпців зі значеннями NULL ігноруються функцією COUNT (), якщо вказано ім’я стовпчика, і враховуються, якщо використовується зірочка (*).

Функції MAX ( ) і MIN ()

Функція **MAX** () повертає найбільше значення з зазначеного стовпця. Для цієї функції необхідно вказувати ім’я стовпця:

Приклад 2.46

```
SELECT MAX(ZN) AS max_penia FROM tax_l.T21Pennoti
WHERE IDROW=209130009784867;
```

Результат

```
-----------------------------------------
 max_penia       3.2
```

Тут функція MAX () повертає максимальну нараховану заробітну плату в таблиці. Рядки стовпців зі значеннями NULL ігноруються функцією MAX (). Функція **MIN** () протилежна відносно дії функції MAX () – вона повертає найменше значення у зазначеному стовпці.

Так само, як і для функції MAX (), для MIN () потрібно вказати ім’я стовпця. Рядки стовпців зі значеннями NULL ігноруються функцією MIN ().

Приклад 2.47

SELECT min(ZN) AS min_penia FROM tax_l.T21Pennoti
WHERE IDROW=209130009788430;

Результат

---

min_penia 3.01

Функція SUMQ

Функція SUM () повертає суму (загальну) значень у певному стовпці.

Приклад 2.48

SELECT sum (ZN) AS sum_penia FROM tax_l.T21Pennoti;

Результат

---

sum_penia17760.01

Усі статистичні функції можна використовувати для виконання обчислень над кількома стовпцями за допомогою стандартних математичних операторів. Рядки стовпців зі значеннями NULL ігноруються функцією SUM ().

Статистичні розрахунки для окремих значень

Усі п’ять статистичних функцій можуть бути використані двома способами:

1) для обчислення у всіх рядках за умов згадування аргументу ALL або без зазначення будь-якого аргументу (оскільки ALL є аргументом за замовчуванням);

2) для вказівки окремих значень за допомогою аргументу DISTINCT. Якщо не вказано аргумент DISTINCT, то мається на увазі аргумент ALL.

Приклад 2.49

Використаємо функцію AVG () для повернення середньої ціни продуктів, пропонованих певним постачальником. Це такий оператор SELECT, як і попередній, але з використанням ключового слова DISTINCT – під час обчислення середнього значення враховують тільки певні ціни.

```
SELECT AVG (DISTINCT prod_price) AS avg_price
FROM Products
WHERE vend id = 'LLOll;
```

Результат

```
---------------------------------------------
avg_price 4.24 0 0
```

У цьому прикладі внаслідок використання ключового слова DISTINCT значення avg_price виходить більш високим, оскільки в таблиці є кілька предметів з однаково низькою ціною. Не враховуючи їх, отримуємо більш високу середню вартість.

Ключове слово DISTINCT можна використовувати з функцією COUNT () тільки в тому випадку, якщо вказано ім’я стовпця. Його не можна використовувати з функцією COUNT (*). Аналогічно, DISTINCT потрібно використовувати з ім’ям стовпця. Незважаючи на те, що ключове слово DISTINCT технічно можна використовувати з функціями MIN () і MAX (), реальної потреби в цьому немає. Мінімальні і максимальні значення в стовпці будуть тими, незалежно від того, зазначені певні значення чи ні.

Комбінування статистичних функцій

У всіх прикладах застосування статистичних функцій, приведених до сьогодні, вказувалася тільки одна функція. Насправді ж оператори SELECT можуть містити стільки статистичних функцій, скільки потрібно.

Приклад 2.50

```
SELECT COUNT (*) AS num_items,
MIN (prod_price) AS price_min,
MAX (prod_price) AS price_max,
AVG (prod_price) AS proce_avg
FROM Products;
```

Результат

```
-------------------------------------------------------------------------
num_items price_min price_max price_avg
 9       3.4900 11.9900 6.823333
```

В одному операторі SELECT використовують одразу чотири статистичні функції і повертають чотири значення (число елементів у таблиці Products, найвища ціна та найнижча ціна товару, середня їхня вартість).

2.3.5. Використання функцій перетворення типів даних

Числові функції, пов'язані із заокругленням

Функція **ROUND** (числовий_аргумент [позиція]) заокруглює значення параметра числовий аргумент з точністю, що визначається параметром позиція. Параметр позиція визначає кількість десяткових знаків після коми. Якщо параметр позиції вказаний, то аргумент заокруглюється до цілих чисел відповідного масштабу (для значення -1 до десятих, від -2 – до сотих тощо). Значення параметра позиція за замовчуванням – 0.

Приклад 2.51

```
SELECT ROUND(-.65,1) AS "Функція ROUND"
2 FROM dual;
```

Результат

```
----------------------------------
Функция ROUND -7
```

Функція **TRUNC** (числовий_аргумент [позиція]) «обрізає» значення параметра числовий_аргумент з точністю, що визначається параметром позиція. Параметр позиція визначає кількість десяткових знаків після коми. Якщо параметр позиція заперечна, то аргумент заокруглюється до цілих чисел відповідного масштабу (для значення -1 до десятків, від -2 до сотень тощо). Значення параметра позиція за замовчуванням – 0.

Приклад 2.52

```
SELECT TRUNC(-.65,1)
2 FROM dual;
----------------------------------------
```

Результат
- .6

Функції, що оперують датами

Функція **SYSDATE** повертає дату і час, що надаються засобами операційної системи сервера бази даних.

Функція **ROUND** (*дата [формат])* – округляється значення параметра дата за шаблоном, який визначається параметром *формат*. Якщо параметр *формат* опущений, то аргумент *дата* округлюється до днів (час встановлюється – 00:00).

Приклад 2.53

```
SELECT TO_CHAR (ROUND (SYSDATE, 'DD'),
'DD-MM-YYYY HH24:MI' )
FROM dual;
---------------------------------------------
```

Результат
15-04-2002 00:00

Функція **TRUNC** (*дата [формат])* «обрізає» значення параметра *дата* за шаблоном, який визначається параметром *формат*. Якщо параметр формат опущений, то аргумент *дата* «обрізає» до найближчого дня (час встановлюється 00:00).

Таблиця 2.4

Найбільш вживані значення параметра формат для дат

Параметр	Тип заокруглення
YYYY YEAR YY	Округлення до року Округлення в більшу сторону відбувається з 1 липня
MONTH MON MM	Округлення до місяця Округлення у більшу сторону відбувається з 16 числа.
W	Округлення до того ж дня тижня, що і перший день місяця.
DDD DD J	Округлення до дня
DAY DY D	Округлення до початкового дня тижня
HH HH12 HH24	Округлення до години
MI	Округлення до хвилин

Приклад 2.54

SELECT TO_CHAR(TRUNC(SYSDATE,'DD'), 'DD-MM-YYYY HH24:MI') FROM dual;

Результат

15-04-2002 00:00

Функції перетворення типів даних

Функція **TO_CHAR** (*аргумент [формат*]) повертає результат перетворення значення параметра аргумент типу NUMBER або DATE в символьний рядок. Для чисел, якщо параметр *формат* опущений, аргумент перетворюється у рядок з довжиною, достатньою для зберігання всіх значущих цифр. Деякі значення параметра *формат* для перетворення числових значень представлені в таблиці 2.5.

Таблиця 2.5

Формат	Вигляд виведеного результату
9	Виводиться цифра. Лідируючий нуль замінюється пробілом.
0	Виводиться цифра. Лідируючий 0 виводиться
EEEE	Результат виводиться в експоненціальній нотації
G	Виводиться символ-розділювач (зазвичай, кома)

Приклад 2.55

```
SELECT TO_CHAR (1000000,'9G999G999') AS "Функція
TO_CHAR'
FROM dual;
```

Результат

```
-----------------------------------------------
Функція TO CHAR
1,000,000
```

Функція **TO_DATE** (*символьний_аргумент, формат])* повертає результат перетворення значення параметра *символьний аргумент* символьного типу у тип DATE. Якщо параметр *формат* опущений, символьний аргумент повинен відповідати формату дати, прийнятому в системі за замовчуванням.

Приклад 2.56

```
SELECT TO_DATE('01.04.2002','DD.MM.YYYY[1] AS "Функція
TO_DATE"
FROM dual;
```

Результат

```
-------------------------------------------
Функція TO_DATE
01-04-2002 00:00:00
```

Функція **TO_NUMBER** (*символьний_аргумент*) повертає результат перетворення значення параметра *символьний_аргумент* символьного типу в значення типу NUMBER.

Приклад 2.57

```
SELECT TO_NUMBER('3.14') * TO_NUMBER('2.7 8E+01')
AS "Функція TO_NUMBER"
FROM dual;
----------------------------------------------------
```

Результат

Функція TO NUMBER 87.292

Питання для самоконтролю до теми 2.3

1. Назвіть основні види аналітичних функцій Oracle.

2. Опишіть основні технічні особливості застосування аналітичних функцій Oracle.

3. Наведіть приклад виклику аналітичної функції.

4. Наведіть приклад конструкції фрагментації.

5. Наведіть приклад запиту за допомогою конструкції «вікна».

6. Наведіть приклад формування запиту форматі «Вікна діапазону».

7. Напишіть запит для вибірки зарплати кожного співробітника й середньої зарплати всіх прийнятих на роботу протягом 300 попередніх днів, а також середньої зарплати всіх прийнятих на роботу протягом 200 наступних днів.

8. Особливості застосування числового виразу PRECEDING.

9. Особливості застосування числового виразу FOLLOWING.

10. Опишіть особливості використання метасимволів для підготовки запитів Select.

11. Особливості використання метасимволу знака відсотка (%).

12. Особливості використання метасимволу «символ підкреслення (_)».

13. Особливості використання метасимволу «квадратні дужки» ([ ]).

14. Особливості використання статистичних аналітичних функцій.

15. Особливості використання функції AVG ().

16. Особливості використання функції COUNT ().

17. Особливості використання функції TRUNC().

18. Особливості використання функції Функція TO_CHAR (аргумент *[формат*]).

2.4. АІС «Галузь» як джерело податкової інформації для аналізу і прогнозування

2.4.1. Основні відомості про АІС «Галузь»

АІС «Галузь» – автоматизована інформаційна система *«Галузь»* податкового відомства України є основним джерелом інформації щодо розрахунків платників з бюджетами усіх рівнів та для формування звітності за встановленими формами і є основним компонентом ІС «Облік платежів».

АІС «Галузь» дозволяє своєчасно отримувати достовірну та повну інформацію для виконання працівниками відомства своїх функціональних обов'язків і забезпечує:

– підвищення оперативності роботи та продуктивності праці податкових інспекторів;

– скорочення обсягу паперового документообігу;

– підвищення достовірності даних щодо обліку платників податків і ефективності контролю за додержанням податкового законодавства;

– оперативне отримання даних про надходження податків (за кожним платником податків або їхньою групою, за кожним видом податку або групою податків) на будь-якому вертикальному рівні системи органів податкового відомства;

– поглиблений аналіз динаміки надходження сум податків і можливість прогнозування цієї динаміки;

– підвищення оперативності та якості управлінських рішень з питань оподаткування на будь-якому рівні.

Системи, побудовані на основі бази даних Oracle та WEB-технологій, містять достатньо широкий перелік можливостей за різними напрямами роботи структурних підрозділів. Разом з тим не завжди є можливість у користувача отримувати необхідну інформацію у вигляді, зручному для використання.

Тому виникає необхідність у виборі даних безпосередньо з баз даних Oracle, що вимагає від користувачів знання місцезнаходження інформації та структури відповідних об’єктів бази даних, а також навичок у побудові запитів з використанням мови SQL.

Якісна та оперативна вибірка даних із баз даних можлива за наявності у користувача чітких знань про взаємодію та структуру об'єктів прикладної системи.

Одним із джерел інформації щодо розрахунків платників з бюджетами є зведені показники АІС «Галузь».

Зведені показники АІС «Галузь»

Облік, контроль і складання звітності щодо платежів, що надходять до бюджету, здійснюється органами податкового відомства України районного рівня з використанням ІС «Податковий блок», що забезпечує єдиний технологічний процес уведення, обробки документів та контролю інформації.

Зведені показники АІС «Галузь» – підготовлена спеціальним способом інформація, яка на сьогодні в податкових органах регіонального (районного) і центрального рівнів є основним джерелом інформації щодо розрахунків платників з бюджетами усіх рівнів та для формування звітності за встановленими формами.

На час впровадження системи дані АІС «Галузь» мали фіксований набір полів, в яких містилися дані про нараховані, сплачені суми, недоїмку та переплату. Такий формат даних не дозволяв додавати нові показники без змін структури файлу, об'єктів баз даних (додавання нових стовбців) та програмного забезпечення.

Впровадження відокремленого обліку платежів з урахуванням розмежування сум по окремих платежах між державним та місцевими бюджетами, вимоги до створення нових звітів (з більшою кількістю показників), перехід на формування даних за КВЕД (а не по галузях), необхідність щоденного отримання звітів та аналітичних інформацій призвело до суттєвого перепроектування об'єктів АІС «Галузь». В основу перепроектування було покладено принцип отримання зведених показників у вертикальній структурі, що дозволяло не змінювати об'єкти при додаванні нових зведених показників.

На сьогодні існує багато видів операцій для проведення їх у картках особових рахунків платників (КОР), і ця кількість постійно збільшується у зв'язку з реалізацією вимог нормативних та законодавчих актів. Але для формування звітності та для аналізу розрахунків платника з бюджетами не потрібно детально знати алгоритми включення операцій (які можуть змінюватися з часом),

а необхідно оперувати більш статичними розрахованими показниками, що можуть порівнюватися за різні періоди. Тому було створено ***єдиний перелік зведених показників*** щодо розрахунків платників з бюджетами та довідник операцій КОР з їхніми алгоритмами включення до цих зведених показників.

На районному та обласному рівнях зведені показники АІС «Галузь» використовують для формування звітності за встановленими формами та отримання інформації про стан розрахунків платників з бюджетами. Крім того, зведені показники АІС «Галузь» використовують під час формування запитів з баз даних за довільними формами тими спеціалістами, які мають доступ до об'єктів БД Oracle. Тому виникає питання побудови запиту до баз даних АІС «Галузь» та виконання їхнім користувачем Oracle, який має прямий доступ до бази даних.

Для формування запитів до баз даних АІС «Галузь» необхідно знати структуру даних та логічну схему прикладної системи. Якісна та оперативна вибірка даних із баз даних можлива за наявності у користувача чітких знань про взаємодію та структуру об'єктів прикладної системи.

Довідники схеми «Еталон», які використовуються АІС «Галузь»

Важливою подією, що вплинула на розвиток АІС «Галузь», було створення еталонної схеми довідників. Перехід на використання довідників дозволив скоротити час на супроводження довідників різних схем, зменшити обсяги пам'яті для збереження довідників, виключити дублювання довідників для різних систем, мати однакові довідники на кожному рівні та забезпечувати достовірність інформації, яка вводиться та отримується за допомогою довідників.

Довідники за функціональним призначенням розподіляються за такими групами:

– група А – довідники державні та затверджені іншими міністерствами і відомствами;

– група В – довідники, створені на основі відомчих довідників (нормативних документів);

– група С – галузеві довідники ДПС України;

– група D – локальні довідники ДПС України.

Окремі довідники схеми «Еталон», які використовуються АІС «Галузь»

У цьому розділі розглянемо довідники схеми «Еталон», які найчастіше використовують під час формування запитів до даних АІС «Галузь»:

– основні E_SPR_LS, E_BUDJTKLS, E_KVED, E_11REGION, E_RAJ_00, E_OPF;

– похідні D_BUDJT_H, D_BUDJT_V, D_KVED_GROUP, D_BDROZM.

Отримати інформацію про назви довідників можна за допомогою скрипта:

```
SELECT b.table_name "таблиця",
UPPER (b.comments) "назва довідника"
FROM all_TAB_COMMENTS b
WHERE b.owner='ETALON'
AND substr (b.table_name,1,2) IN ('E_', 'D_')
ORDER BY b.table_name ;
```

Результатом даного запиту є відібрані основні довідники, які використовуються АІС «Галузь» (табл. 2.6).

Таблиця 2.6

Основні довідники, які використовуються АІС «Галузь»

Таблиця	Назва довідника, призначення довідника
D_BDROZM	Відсотки розмежування сум для відокремленого обліку платежів
D_BUDJT_H	Довідник бюджетної класифікації
D_BUDJT_V	Довідник бюджетної класифікації (вертикальний)
D_KVED_GROUP	Довідник ієрархії (входження) кодів КВЕД
E_BUDJTKLS	Коди бюджетної класифікації
E_DB_FOUND	Довідник розподілу між загальним та спеціальним фондом
E_KVED	Класифікація видів економічної діяльності
E_OPF	Класифікація організаційно-правових форм
E_RAJ_00	Коди регіональних підрозділів ДПС
E_11REGION	Коди областей України

Для побудови запитів важливо мати інформацію про структуру довідників, які оперативно поновлюються на центральному рівні і є свого роду репозитарієм. Вони забезпечують цілісність інформації і ефективність виконання запитів до баз даних. Отримати інформацію про структуру довідників можна за допомогою наступного скрипта:

```
SELECT  c.table_name "таблиця",
        c.column_name "стовпець",
        b.comments "зміст стовпцю",
            c.data_type||'('||
            decode(c.data_length,22,
            to_char (c.data_precision) ||
              decode(c.data_scale, NULL,'', ','||
to_char(c.data_scale)), c.data_length) ||') ' "тип даних"

    FROM all_TAB_COLS c, all_COL_COMMENTS b

WHERE b.table_name = c.table_name
        AND c.owner='ETALON'
        AND substr(b.table_name,1,2) IN ('E_', 'D_')
        AND b.column_name = c.column_name
ORDER BY c.table_name, c.column_name
```

Опис роботи з похідними довідниками

При формуванні запитів використовуються такі похідні довідники: D_BUDJT_H, D_BUDJT_V, D_KVED_GROUP, D_BDROZM.

Наведемо опис деяких важливих показників.

Довідник ієрархії (входження) кодів КВЕД **D_KVED_GROUP** містить реквізити:

KVED – код КВЕД (7 символів).

MKVED – повний код КВЕД (батьківський код) складається із секції, підсекції та коду КВЕД. До MKVED=’Z’ входять усі коди КВЕД.

Таблиця 2.7

Структура таблиць довідників E_SPR_LS, D_BUDJT_V

Таблиця	Стовбець	Зміст стовпця	Тип даних
E_SPR_LS	CODE	Символьний код показника (унікальне значення)	VARCHAR2 (10)
	D_BEGIN	Дата початку дії запису	DATE(7)
	D_DICT	Дата модифікації довідника	DATE(7)
	D_END	Дата кінця дії запису	DATE(7)
	IDPARM	Числовий код показника (унікальне значення)	NUMBER(5,0)
	ID_REC	Модифікатор запису	NUMBER(8,0)
	NAME	Економічний зміст показника	VARCHAR2 (59)
	OPR	Ознака того, що показник відображається в КОР наростаючим підсумком	NUMBER(2,0)
	RULEASM	Номер правила для групування згідно з бюджетною класифікацією (0-для звіту 1ПО, 5-для звіту 1П5, 6-для звіту 1П6, 8-для звіту 1П8)	NUMBER(2,0)
	SHORTNAME	Скорочена назва показника	VARCHAR2 (32)
D_BUDJT_V	D_BUDJTKLS_B	Дата початку дії класифікатора	DATE(7)
	D_BUDJTKLS_E	Дата кінця дії класифікатора	DATE(7)
	EXCL	Ознака виключення	VARCHAR2(1)
	RULEASM	Тип агрегації даних за бюджетною класифікацією	NUMBER(1,0)
	SHOT	Шифр бюджетної класифікації	NUMBER(8,0)
	SHOTTO	Батьківський шифр бюджетної класифікації	NUMBER(8,0)
	SIGN	Знак	NUMBER(1,0)
	SMPL	Ознака первинного шифру бюджетної класифікації	VARCHAR2 (1)

Тобто для отримання даних у розрізі КВЕДів потрібно проводити групування по MKVED за умови порівняння реквізиту KVED з відповідним реквізитом іншої таблиці (наприклад, VG306).

Довідник бюджетної класифікації ***D_BUDJT_H*** містить дані довідника E_BUDJTKLS на кожну дату змін довідника. Наприклад, для отримання назви платежу за кодом необхідно враховувати період дії запису: початок d_budjtkls_b і кінець d_budjtkls_e. У цей проміжок повинна потрапляти звітна дата (RDATA).

Довідник бюджетної класифікації (вертикальний) ***D_BUDJT_V*** містить інформацію про ієрархію кодів бюджетної класифікації (входження кодів до агрегованих) на кожну дату змін довідника. Реквізит SHOTTO означає агрегований код платежу, SHOT – код платежу, який входить у SHOTTO. Обов'язково слід враховувати період дії запису: початок d_budjtkls_b і кінець d_budjtkls_e. У цей проміжок повинна потрапляти звітна дата (RDATA).

Крім того, потрібно враховувати правила входження платежів до розділів звіту 1-П (нараховано – 5, сплачено – 0, борг – 6, переплата – 8) – реквізит RULEASM. Для кожного зведеного показника IDPARM встановлено це правило в довіднику E_SPR_LS. Тому вам необхідно в запиті додатково вказувати умову:

```
E_SPR_LS.RULEASM = D_BUDJT_V.RULEASM
AND VG306.IDPARM = E_SPR_LS.IDPARM.
```

2.4.2. Особливості роботи із довідниками схеми «Еталон»

У схемі ETALON також можуть розміщуватися функції, які використовують дані довідників (наприклад, отримання назви за кодом – первинним ключем). Ці функції розробники в обов'язковому порядку розміщують у схемі ETALON та надають привілеї на виконання grant execute on <назва функції > to public.

Уже розроблені функції, які також розміщені у схемі «Еталон», можна використовувати для уникнення у запиті зайвого з'єднання основних таблиць та довідників. Перелік функцій схеми «Еталон», які можуть використовуватися при формуванні запитів, подано у табл. 2.8.

Таблиця 2.8

Перелік функцій схеми «Еталон», які можуть використовуватися під час формування запитів

Функція	Призначення
KVED_NAME (prm1, prm2)	Параметри: {prm1} - ознака (=1 або =2) {prm2} - код КВЕД (KOD для ознаки 1) або повний код КВЕД (KVED для ознаки 2) Функція, яка відповідно до ознаки та коду КВЕД повертає назву NU з довідника ETALON.E_KVED
OPF_NAME (prm1)	Функція, яка за кодом ОРГАНІЗАЦІЙНО-ПРАВОВОЇ ФОРМИ {prm1} повертає назву з довідника ETALON.E_KFV
RAJ_DST(prm1, prm2)	Функція, яка за кодом регіону {prm1} та кодом району {prm2} повертає код ДПІ з довідника ETALON.E_RAJ_00
RAJ_NAME (prm1, prm2)	Функція, яка за кодом регіону {prm1} та кодом адміністративно-територіального району {prm2} повертає назву району з довідника ETALON.E_RAJ_00
RAJ_STINAME (prm1, prm2)	Функція, яка за кодом регіону {prm1} та кодом адміністративно-територіального району {prm2} повертає назву ДПІ з довідника ETALON.E_RAJ_00
REGION_NAME (prm1)	Функція, яка за кодом регіону {prm1} повертає назву регіону з довідника ETALON.E_11REGION
TAX_SHOT_NAME (prm1)	Функція, яка за кодом бюджетної класифікації {prm1} повертає назву податку (платежу) на останню дату з похідного довідника ETALON.D_BUDJT_H
STAN_NAME (prm1)	Функція, яка за кодом СТАНУ ПЛАТНИКА {prm1} повертає назву з довідника ETALON.E_S_STAN

Враховуючи зазначене вище щодо довідників та функцій схеми «Еталон», можна виділити основне, що обов’язково потрібно знати:

– створення запитів до АІС «Галузь» практично завжди вимагає використання існуючих довідників схеми «Еталон»;

– довідники поділяються на основні та похідні. Назви основних довідників починаються з префіксу E_, похідних довідників з префіксу D_. Похідні довідники, як правило, створюються з основних довідників за допомогою процедур з метою організації ієрархії даних;

– використання похідних довідників схеми «Еталон» найчастіше потребує встановлення додаткових умов для періоду дії запису;

– для уникнення зайвого з’єднання таблиць можна використовувати функції схеми «Еталон».

Нижче подано приклад запиту з використанням функцій:

```
SELECT
        FLOOR(k.c_distr/100) "регіон",
        SUBSTR(etalon.region_name(floor(k.c_distr/100)),1,20
0) "назва регіону",
        MOD(k.c_distr,100) "ДПІ",
        SUBSTR( etalon.dst_name( floor(k.c_distr/100),
        etalon.raj_dst(
 FLOOR (k.c_distr/100),mod(k.c_distr,100) ) ),1,200) "назва ДПІ",
kod KOD "платник", k.name "назва платника", k.kved "КВЕД",
     SUBSTR(etalon.kved_name(1,k.kved),1,200) "назва КВЕД",
          g1,g2,g3,g4
FROM
  (SELECT kod, sum (NADHOD)/1000 G1,
          SUM ( decode(shot,11010100,
          NADHOD,0))/1000 G2,
          SUM (NADHOD_m)/1000 G3,
          SUM ( decode(shot,11010100,
          NADHOD_m,0))/1000 G4
          FROM
....
GROUP BY kod ) z, r21tp k
WHERE z.kod = k.tin(+)
AND z.facemode = k.face_mode(+)
ORDER BY g1 desc
```

2.4.3. Структура таблиць зведених показників АІС «Галузь»

Файл зведених показників АІС «Галузь» (типів VG,VP, VD) – це структурована сукупність даних у форматі *.dbf та *.txt, яка являє собою згруповану за певними критеріями сукупність показників з карток особових рахунків платників податків станом на звітну дату. Структура файлу зведених показників АІС «Галузь» подана у таблиці 2.9.

Таблиця 2.9

Структура файлу зведених показників АІС «Галузь»

Зміст реквізиту	Назва реквізиту	Тип та розмірність
Звітна дата	RDATA	D
Код області	C_REG	N(2)
Код ДШ	C_DST	N(2)
Код адміністративного району	C_RAJ	N(2)
Код платника	KOD	N(10)
Код ДПІ реєстрації згідно з splat.c_sti	C_STI	N(4)
Код органу управління	KOD_U	N(5)
Код форми власності	KOD_SOB	N(2)
Код виду діяльності	VD	N(5)
КВЕД	KVED	C(7)
Код організаційно-правової форми	SS	N(5)
Код бюджету	VB	N(2)
Код бюджетної класифікації	SHOT	N(8)
Стаття	ST	N(2)
Код показника згідно з spr_ls	IDPARM	N(5)
Значення показника	ZN	N(16,2)
Ознака типу даних (1 – юридичні особи, 2 – фізичні особи-СГД, 4 – земля фіз.осіб, 8 – сумісники)	DATA_TYPE	N(l)
Контрольна сума запису	CONTROLSUM	N(6)
Код форми	KF	N(2)
Код стану платника	C_STAN	N(2)

Значення основних полів файлу зведених показників АІС «Галузь»:

Звітна дата – дата, станом на яку сформовано файл зведених показників (для звітних файлів типу VG(VP) – це завжди 1-ше число місяця, для файлів змін до попереднього дня типу VD – це поточні робочі дати).

Код області, код ДПІ, код адміністративного району – коди територіальних одиниць податкових органів, де ведеться оперативний облік платежів (відкриті КОР платника). Відповідні назви територіальних одиниць для комбінації цих полів можна знайти в довіднику «Коди регіональних підрозділів ДФС» (ETALON.E_RAJ_00).

Код області та код ДПІ (або код адміністративного району) можуть відрізнятися від значення поля "Код ДПІ реєстрації", оскільки СПД, зареєстрований як основний платник в одній ДПІ, одночасно може бути зареєстрований як платник окремих податків в одній або декількох інших ДПІ.

КВЕД – код виду економічної діяльності платника податків. Наявність цього реквізиту дозволяє проводити аналіз зведених показників у розрізі галузей (секторів) економіки та видів економічної діяльності. Значення поля можна знайти в довіднику "Коди видів економічної діяльності" (ETALONE_KVED)

Код бюджету, код бюджетної класифікації, стаття – визначають реквізити окремого платежу (відповідають кількості відкритих КОР для кожного окремого платника в межах однієї інспекції).

Назву коду бюджетної класифікації можна знайти в довіднику "Коди бюджетної класифікації" (ETALONE_BUDJTKLS). Для платежів, які обліковуються за кодом бюджету "85" (Рахунок розмежування казначейства), необхідно враховувати відсоток розмежування між бюджетами відповідно до даних довідника "Напрями розмежування" (ETALON.EJTREND).

Код показника згідно з SPR_IS – саме цей реквізит дозволяє вибирати із загального масиву даних окремі показники (або їхню комбінацію), які характеризують стан розрахунків платників з бюджетом. Загальний перелік зведених показників та їхніх назв міститься в довіднику **ETALONE_SPR_LS.**

Ознака типу даних (DATA_TYPE) – визначає зріз даних відповідно до типу платника: 1 – юридичні особи, 2 – фізичні особи-СГД, 4 – земля фізичних осіб, 8 – сумісники. Цей реквізит дозволяє вибирати із загального масиву даних показники за окремими категоріями платників, наприклад «юридичні особи» або «фізичні особи».

2.4.4. Рекомендації щодо формування та оптимізації аналітичних запитів

Метою цього параграфа є надання рекомендацій щодо побудови коректних запитів до БД АІС «Галузь» для отримання аналітичної інформації.

Увесь попередній матеріал за зазначеною темою стосувався базових знань щодо еталонних довідників, які використовуються при формуванні запитів та знань про фізичну та логічну структуру об’єктів бази даних.

Але при написанні запитів також необхідно застосовувати знання щодо коректної побудови та оптимізації запитів з метою підвищення швидкодії виконання. Метою оптимізації оператора SQL є прискорення доступу до бази даних в її поточному вигляді, використовуючи особливості структури самої бази даних, системних ресурсів та індексів.

Розглянемо ці питання детальніше.

Для створення ефективних запитів необхідно знати джерела даних; обсяг даних і їхній розподіл у базі даних. Необхідно мати повне розуміння моделі даних (особливо зв’язків між різними об’єктами), що використовується до написання необхідних операторів SQL.

Зовнішній вигляд оператора SQL ніяк не впливає на швидкість його виконання, але його форматування для зручності читання допомагає швидше побачити і внести зміни при налагодженні довгих і складних операторів. У операторі, написаному за правилами, легко побачити, які стовпці і з яких таблиць відібрані, як обрані таблиці пов’язані і які умови накладені на дані, що вибираються запитом.

Оптимізація оператора SELECT означає вибір такої форми оператора, при якій він працює максимально швидко й ефективно. Оптимізація оператора SELECT починається з вибору найкращого порядку розміщення елементів, з яких оператор складається (головним чином, у виборі правильної форми виразів ключових слів FROM і WHERE). Саме залежно від форми цих двох виразів сервер бази даних вирішує, як і в якому порядку слід виконати запит.

Під час вибору оптимальної форми оператора SELECT слід враховувати й інші моменти:

– використання ключового слова LIKE і знаків підстановки;
– заміна операцій OR виразом з ключовим словом IN;
– недоліки використання виразу з ключовим словом HAVING;
– необхідно уникати довгих операцій сортування;
– використання готових процедур.

Для фільтрування записів у більшості випадків доцільніше використовувати WHERE, а не HAVING. Слід уникати використання оператора HAVING разом з GROUP BY на індексованих стовпцях (у цьому випадку індекс не використовується).

Одним з основних моментів у справі оптимізації операторів SQL є вміле використання індексів.

Оператори SQL можуть бути виконані з використанням оптимізатора, керованого правилами, або оптимізатора, заснованого на оцінках. Компанія Oracle удосконалює оптимізатор, заснований на оцінках, в кожному випуску, роблячи його більш стабільним і надійним. Для таблиць АІС «Галузь» постійно проводиться збір оціночної статистики (ANALIZE ESTIMATE STATISTICS). Не рекомендується в тексті скрипта вказувати підказки щодо підключення індексів (наприклад, /*+ INDEX(a) */ або /*+ FULL(a) */ і т.д.). При розборі тексту запиту оптимізатор самостійно визначає, які індекси потрібно підключати. Але якщо ви володієте навичками щодо створення планів розбору запитів, то зможете проаналізувати варіанти швидкодії при різних варіантах підключення індексів.

Питання для самоконтролю до теми 2.4

1. Опишіть основні завдання, виконання яких забезпечує АІС «Галузь».

2. Дайте визначення поняття Зведені показники АІС «Галузь».

3. Опишіть призначення і функції Картки особових рахунків платника податків (КОР).

4. Для чого використовується відокремлений облік платежів?

5. Для чого використовуються в АІС «Галузь» використовуються довідники схеми «Еталон»?

6. Опишіть основні групи Довідників за функціональним призначенням.

7. Наведіть та опишіть економічний зміст основних стовпців таблиць довідників E_SPR_LS, D_BUDJT_V.

8. Опишіть основний зміст та призначення Довідника бюджетної класифікації D_BUDJT_V.

9. Опишіть як враховується входження платежів до розділів звіту 1-П.

10. Опишіть структуру таблиць зведених показників АІС «Галузь». Опишіть основні реквізити та їхній економічний зміст.

11. Наведіть коди бюджетів, бюджетної класифікації, статті декількох основних податків (податку на додану вартість, податку на прибуток підприємств тощо).

Лабораторний практикум 2

Лабораторна робота 2.1
Доступ до баз даних Oracle на основі технології ODBC. Відбір даних та побудова зведених таблиць.

Мета: отримання навичок у використанні технології ODBC для аналітичних запитів із середовища Excel до АІС «Галузь» БД Oracle. Придбання навичок у використанні MS Query для доступу до БД Oracle, аналізу даних типу .dbf та їх аналізу в MS Excel, формування навичок використання платформи ODBC та OLEDB для доступу до даних.

Завдання роботи

1) створити ODBC джерело даних для під'єднання до БД Oracle із середовища Excel.

2) вибрати за допомогою MsQuery необхідні дані для аналізу із таблиць АІС «Галузь».

3) побудувати зведену таблицю та зведений графік відповідно до варіанта завдання. Наприклад, результуюча зведена таблиця повинна:

– відобразити дані про суми недоїмки всіх платників податків за основним місцем обліку (C_STAN=0) (по всіх податках або по вказаному, наприклад, з податку на прибуток (SHOT=110100XX)) по актах перевірок (IDPARM-119), по основному платежу – (IDPARM-119), по фінсанкціях (IDPARM-123);

– вибирати (фільтрувати) для аналізу такі показники: вид операції (IADPARM); вид податку (SHOT); вид діяльності (KVED); стан платника податків (C_STAN);

– відображати дані у вигляді стовпчикової діаграми (по осі x – показники недоїмки, по осі y – суми недоїмки;

– таблиця повинна оновлюватися даними з БД Oracle кожні 10 хвилин;

4) сформувати додатковий обчислюваний об'єкт (поле) таблиці «Коефіцієнт заборгованості» як відношення суми боргу до суми «Нараховано всього (+)» - (IDPARM =7);

5) підготувати звіт про виконання роботи, у якому відобразити хід роботи, екрани отриманих результатів із поясненнями, висновки до отриманих результатів.

Вихідні дані для виконання лабораторної роботи:

1) IP адреса сервера БД Oracle, де розміщуються доступні користувачу об'єкти АІС «Галузь». У наведеному прикладі це адреса 192.168.150.5;

2) унікальне ім'я екземпляра Oracle: SID (SERVICE_ NAME) = ORCL;

3) Аліас (ім'я) з'єднання з БД Oracle, прописаний у файлі *tnsnames.ora*. (У прикладі – ORCL5 (пп. 2.1.3);

1) структура таблиць АІС «Галузь» та опис полів таблиці (Детальний опис таблиць дано у пп. 2.4.3):

– таблиця VG – файл зведених показників АІС «Галузь»;

– таблиця SPR_LS – довідник показників IDPARM карток особових рахунків платників податків;

– таблиця S_STAN – довідник станів платників податків.

2) основні зведені показники та їхній зміст, які використовуються у завданнях (табл. 2.10) .

Таблиця 2.10

Код зведеного показника (IDPARM)	**Назва зведеного показника згідно з SPR_LS**	**Примітка (де і як може використовуватися)**
119 120 123 76	Недоїмка з основного платежу Недоїмка за актами Залишок фінансових санкцій, які підлягають списанню Недоїмка з пені	Формула розрахунку загальної суми податкового боргу (без урахування сум податкового боргу платників "банкрутів"): = 119+ 120+ 123 + 76. Формула розрахунку загальної суми податкового боргу за податковими зобов'язаннями платників без урахування сум податкового боргу платників "банкрутів": 119+ 120+ 123
7 8 54	Нараховано всього (+) Нараховано всього (-) Нараховано або зменшено	

Порядок виконання лабораторної роботи

1. Формування джерела даних з'єднання з БД Oracle.

Для формування під'єднання до БД Oracle на основі ODBC виконується така послідовність операцій:

1.1. В Адміністраторі ODBC («Пуск» / «Панель управления» / «Система и безопасность» / «Администрирование») створюється джерело даних для з'єднання з БД Oracle (DSN – Data Source Name). Вибираємо «Источники данных ODBC» (рис. 2.13).

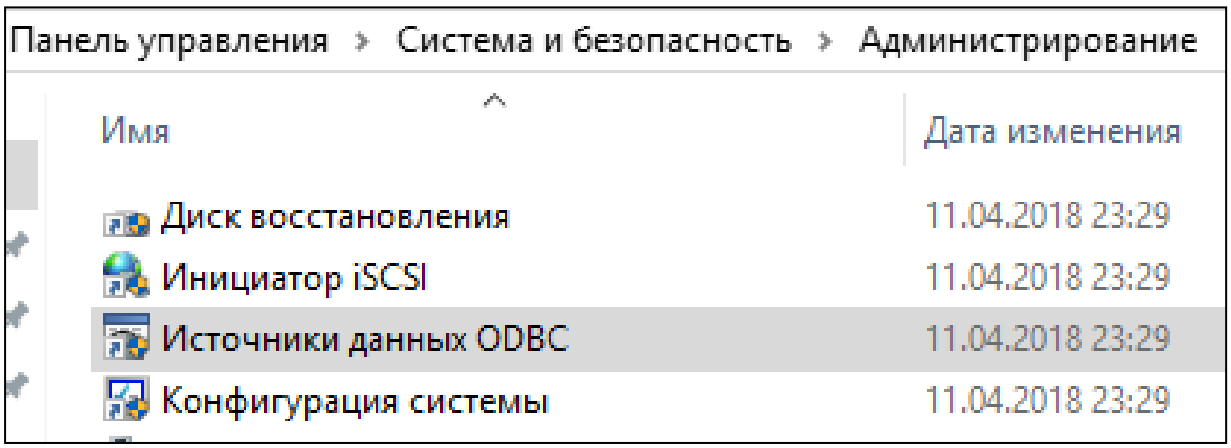


Рис. 2.13. Вікно вибору режиму «Адміністрування ODBC»

1.2. Вибирається тип джерела даних (DSN): User DSN (Користувацький DSN – доступний поточному користувачу), System DSN (Системний DSN – доступний всім користувачам), File DSN (Файловий DSN – налаштування зберігаються в окремому файлі), рис. 2.14.

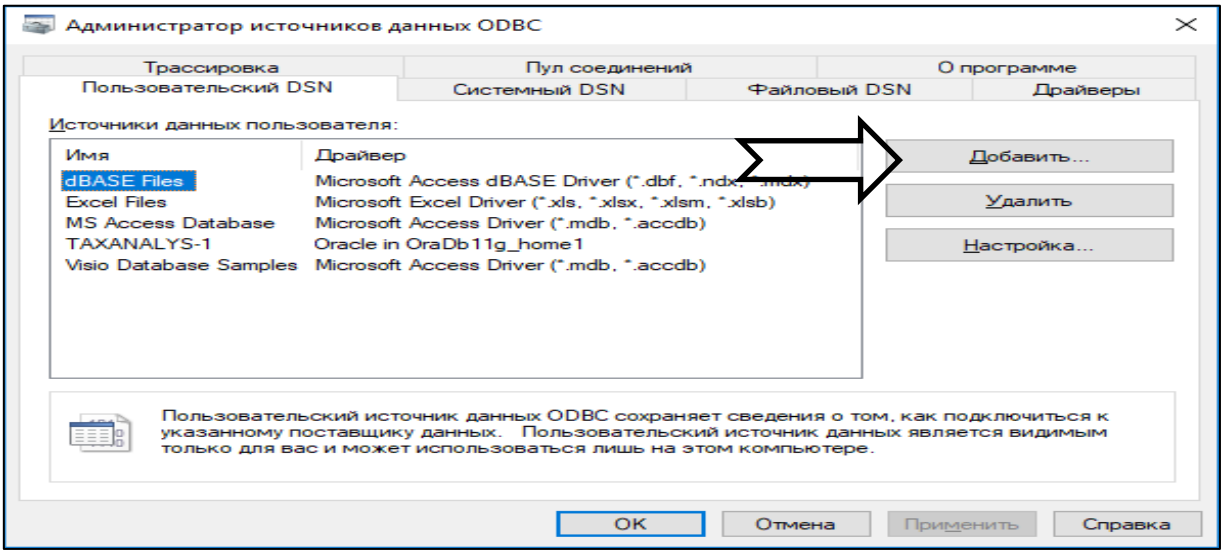


Рис. 2.14. Вибір типу створюваного DSN: створення користувацького DSN

1.3. Вибирається відповідний драйвер ODBC. При інсталяції на сервері лише сервера ДБ відображається: Oracle in OraDB11g_Home1 (рис. 2.15.). Можна використовувати такі драйвери: *Microsoft ODBC for Oracl;i Oracle in OraDB11g_home 1; OraClent11g_home 1*.

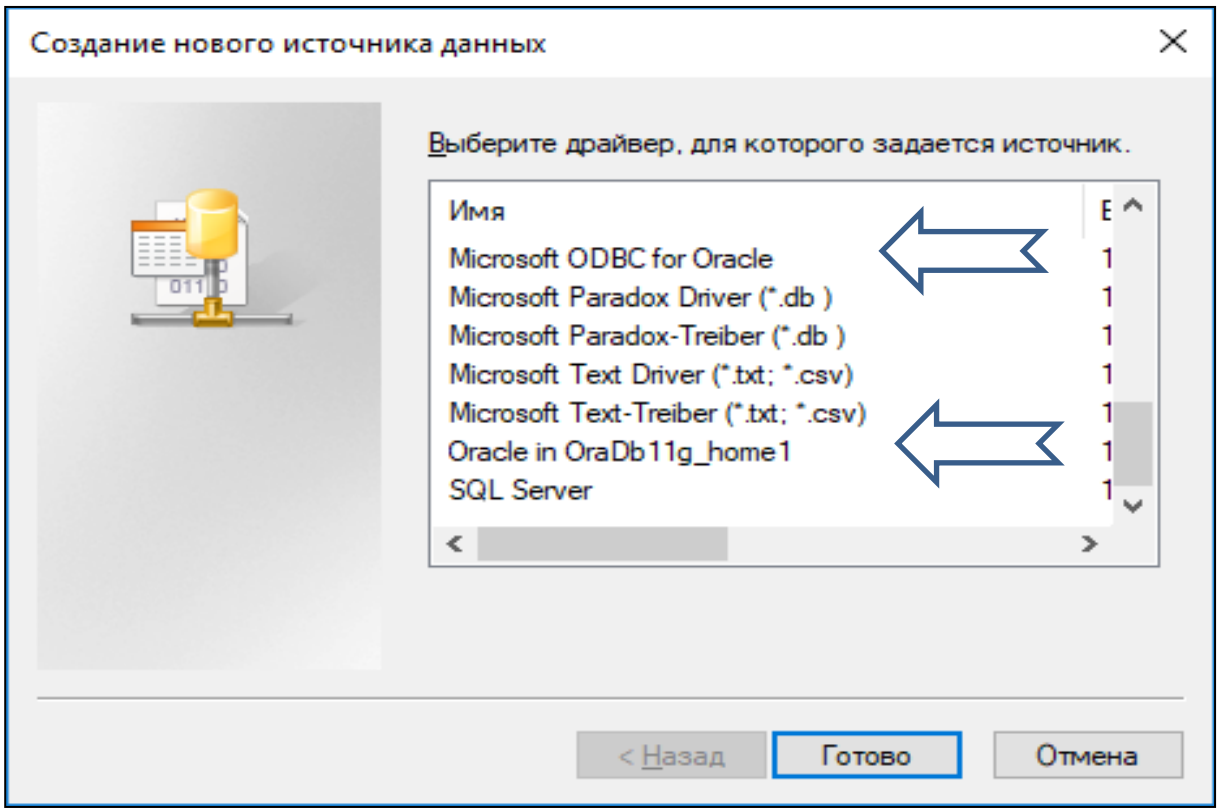


Рис. 2.15. Вибір драйвера ODBC

1.4. Здійснюється конфігурація драйвера ODBC, для чого:

– Data Source Name – вказується DSN – довільна назва з'єднання із конкретною БД. Вводимо для прикладу ORCL5-TAXANALYS;

– Description – довільне пояснення щодо цього з'єднання;

– TNS Service Name – вказується SID екземпляра БД Oracle (береться із файлу tnsnames.oraf або надається адміністратором БД. У нашому випадку це ORCL5;

– Usrer ID – вказується ім'я користувача із відповідними правами доступу до БД (користувач створюється адміністратором БД). МИ будемо під'єднуватися із правами користувача TAXABALYS;

– Test Connection – перевіряється з'єднання з БД Oracle (рис. 2.16).

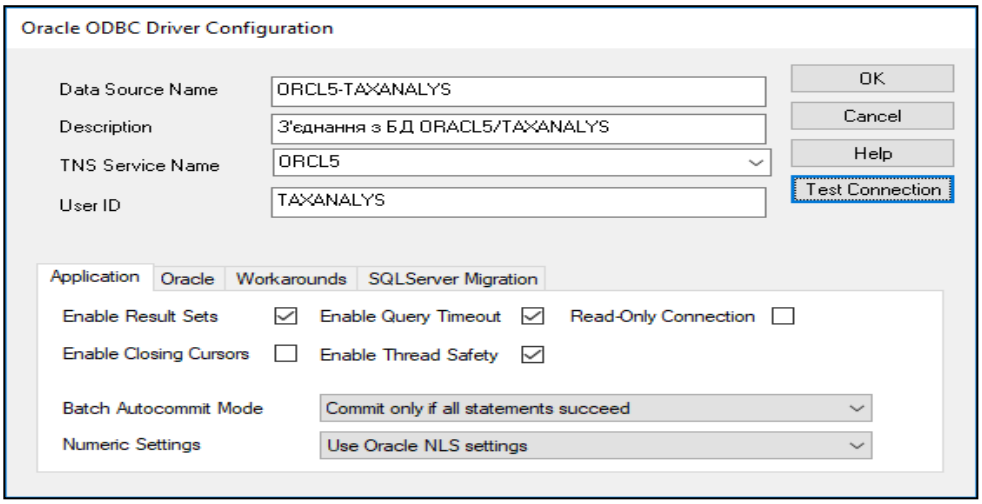


Рис. 2.16. Конфігурація драйвера ODBC

2. ODBC з'єднання з БД Oracle через MS Query та побудова зведеної таблиці відповідно.

Рішення (відповідно до вихідних даних для виконання роботи):

2.1. Для створення DSN викликаємо Адміністратор ODBC: «Панель управления»/ «Администрирование»/ «Источники данных ODBC». Для надання доступу до джрела всім користувачам ПЕОМ на вкладці «Системные DSN» створюємо DSN з іменем ORCL5-TAXANALYS. Налаштовуємо DSN як показано на рис. 2.16.

2.2. У Microsoft Excel створюємо зв'язок з БД Oracle, використовуючи майстер зі створення SQL запитів Microsoft Query. Для версії Office 10 потрібно виконати такі дії:

2.2.1. Вибрати опцію «Из Microsoft Query» («Данные»/»Из других источников»/«Из Microsoft Query») (рис. 2.17).

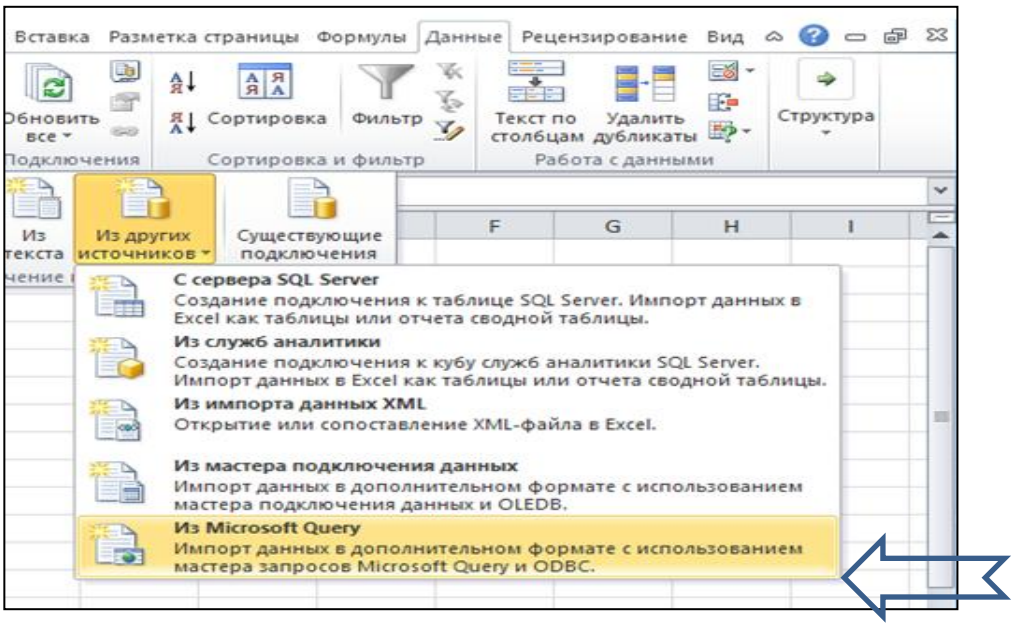


Рис. 2.17. Вибір опції роботи із Microsoft Query

2.2. Вибрати у вікні «Выбор источника данных» попередньо створене джерело даних. Вибираємо попередньо створене джерело даних ORCL5-TAXANALYS, далі – «ОК» (рис. 2.18). Важливо знати, що на цьому етапі можна створити потрібний DNS, вибравши у вікні «Выбор источника данных» режим <Новый источник данных>;

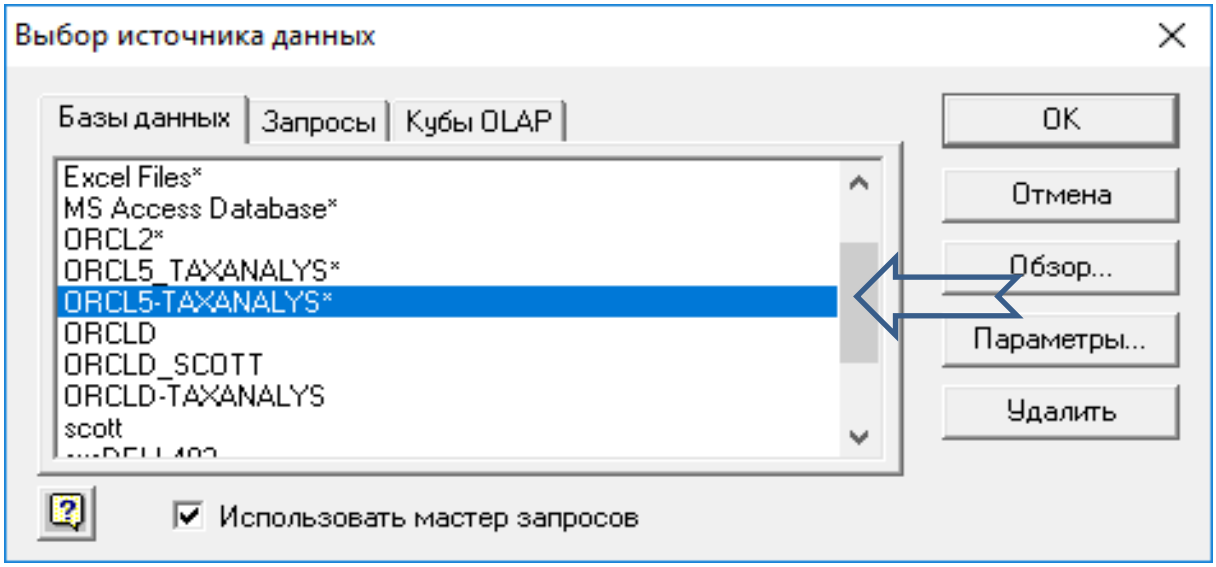


Рис. 2.18. Вибір джерела даних

2.2.3. Після введення пароля доступу у вікні, що з'являється, потрібно вибрати схему БД Oracle, до об'єктів якої має доступ користувач (таблиць, уявлень тощо), рис. 2.19.

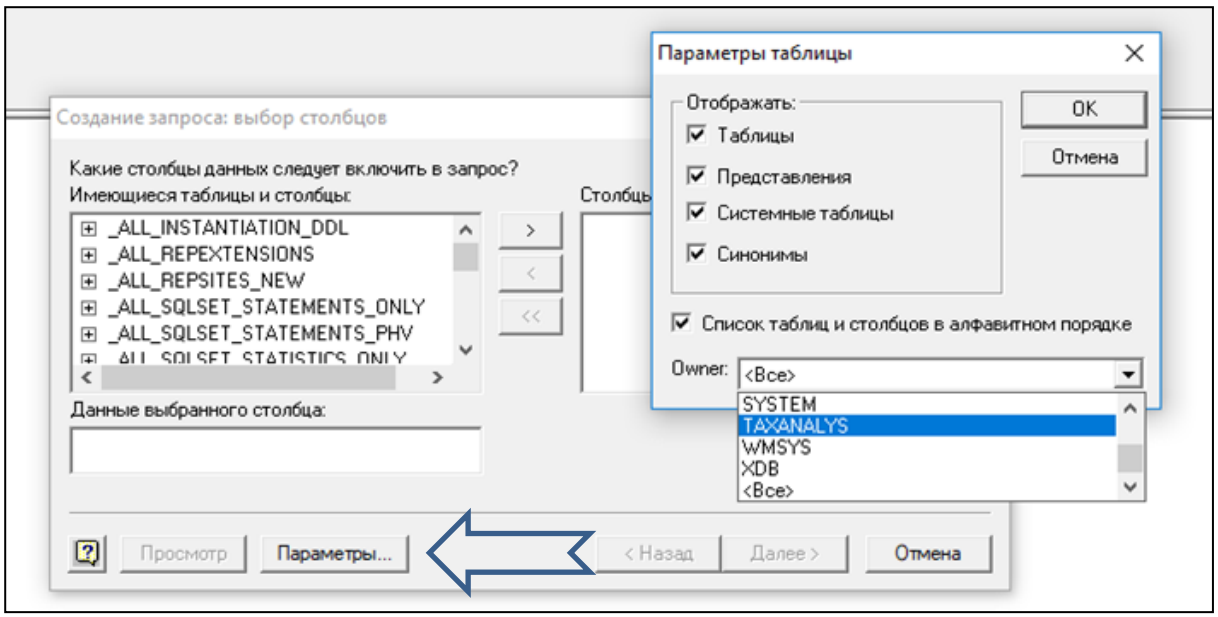


Рис. 2.19. Вибір схеми БД Oracle з об'єктами користувача TAXANALYS

2.2.4. Далі вибираються потрібні поля із таблиць БД Oracle. Користувач (фахівець підрозділу звітності) повинен знати зміст показників таблиць, до яких звертається.

Для прикладу показано вибір трьох таблиць автоматизованої системи «Облік платежів» GALUZ: зведені показники (таблиця VG), довідник операцій IDPARM (таблиця SPR_LS), довідник станів платників податків (таблиця S_STAN), рис. 2.20.

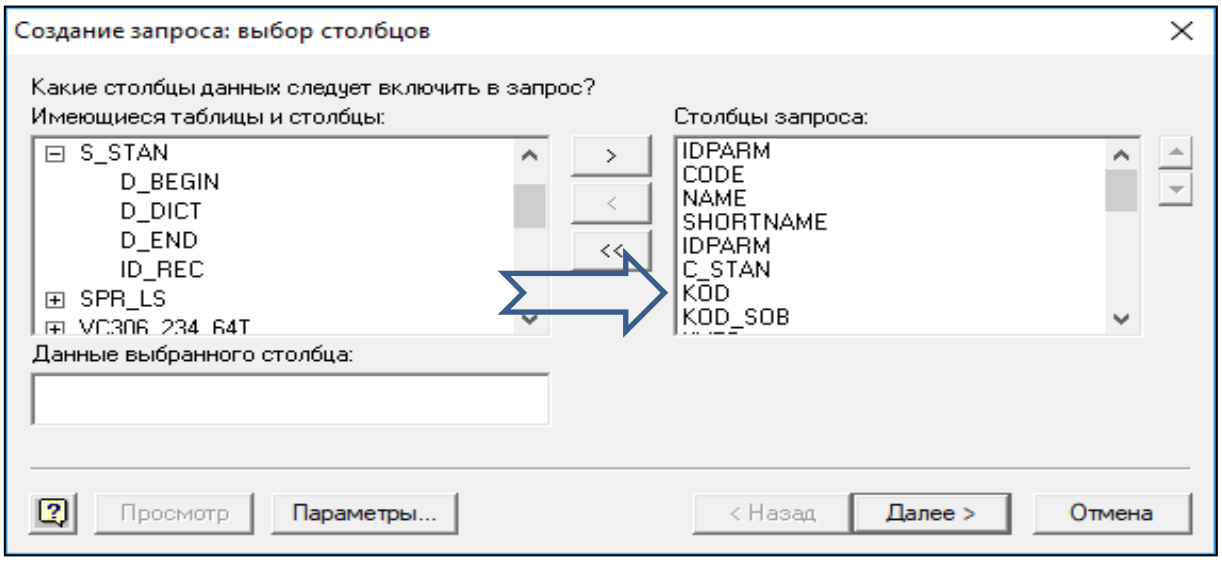


Рис. 2.20. Вибір полів таблиць схеми GALUZ AC «Облік платежів»

2.2.5. Після вибору таблиць майстер запитів може попросити об’єднати таблиці в Microsoft Query, якщо у них не вказані первинні та зовнішні ключі. У цьому прикладі таблиці з’єднані таким чином:

– поле IDPARM таблиці VG з полем IDPARM таблиці SPR_LS;

– поле C_STAN таблиці S_STAN із полем C_STAN таблиці VG.

3. Побудова зведеної таблиці та графіка.

3.1. Після об’єднання таблиць дані можна передати у лист Excel, причому одночасно із запуском майстра зведених таблиць та зведеної діаграми (рис. 2.21–2.22).

3.2. У вікні «Список полей сводной таблицы» вибрані поля розміщуються відповідно до поставленого у варіанті завдання формування звітності. Результат формування зведеної таблиці відповідно до завдання прикладу показано на рис. 2.23–2.24.

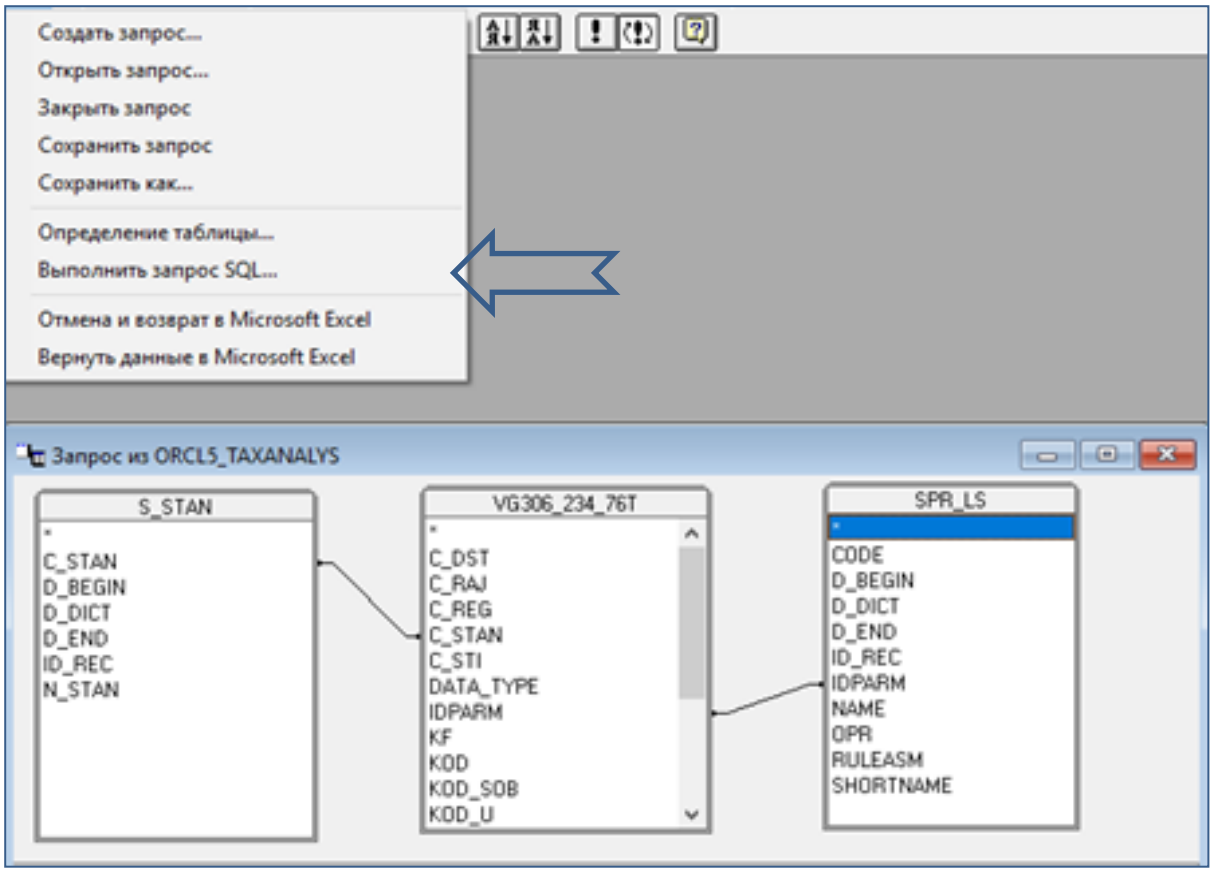


Рис. 2.21. Об'єднання таблиць за ключовими полями в Microft Query

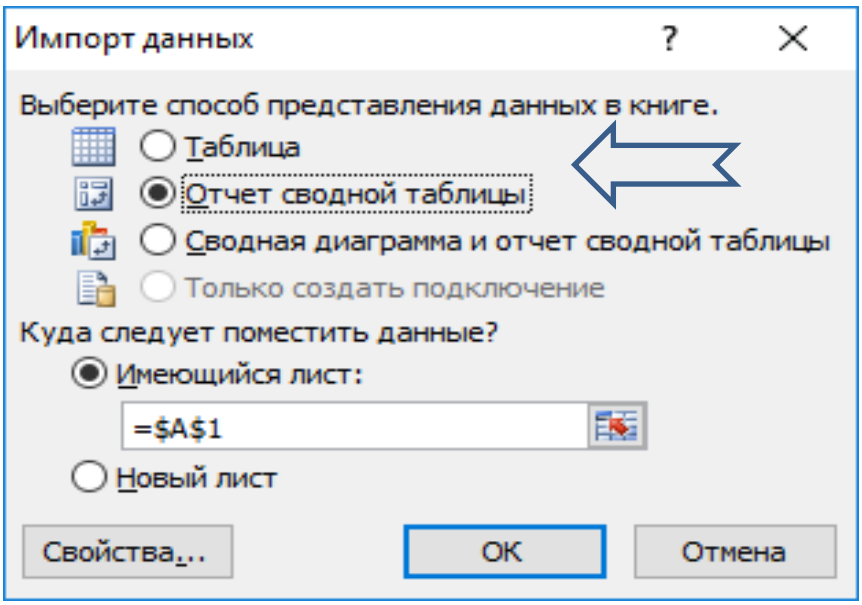


Рис. 2.22. Вибір опції запуску майстра зведених таблиць після імпорту даних

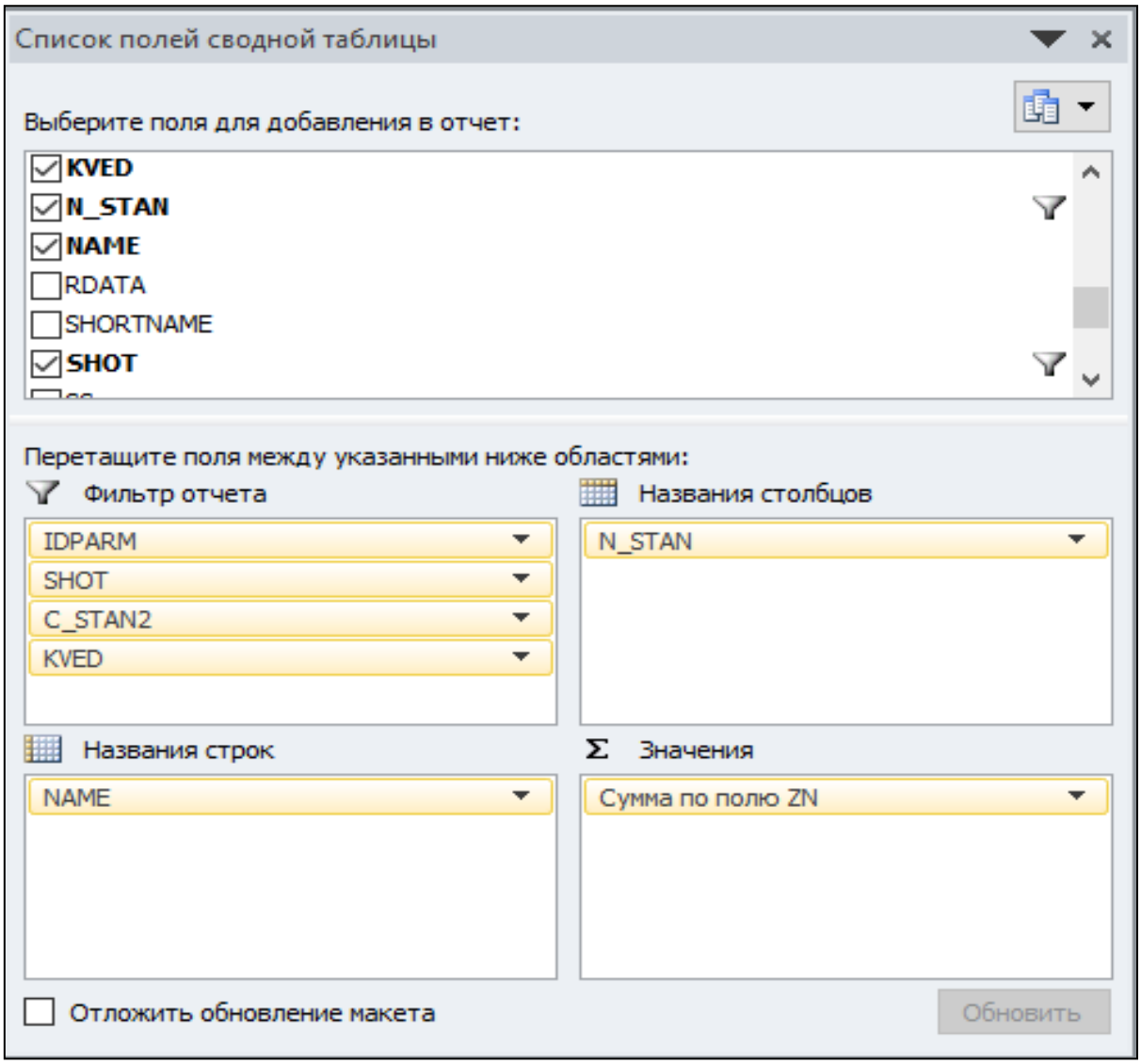


Рис. 2.23. Результат розміщення полів вибраних таблиць у конструкторі зведеної таблиці

3.3. Для автоматичного періодичного відновлення даних із БД Oracle в контекстному меню «Свойства подключения» встановлюємо поновлення даних кожні 10 хвилин та поновлення при відкритті файлу, рис. 2.25.

Таким чином сформована таблиця та графік будуть оновлюватися актуальними даними з БД Oracle кожні 10 хв, до того ж користувачі будуть мати можливість відбирати потрібні дані за допомогою кнопок фільтрації (рис. 2.24).

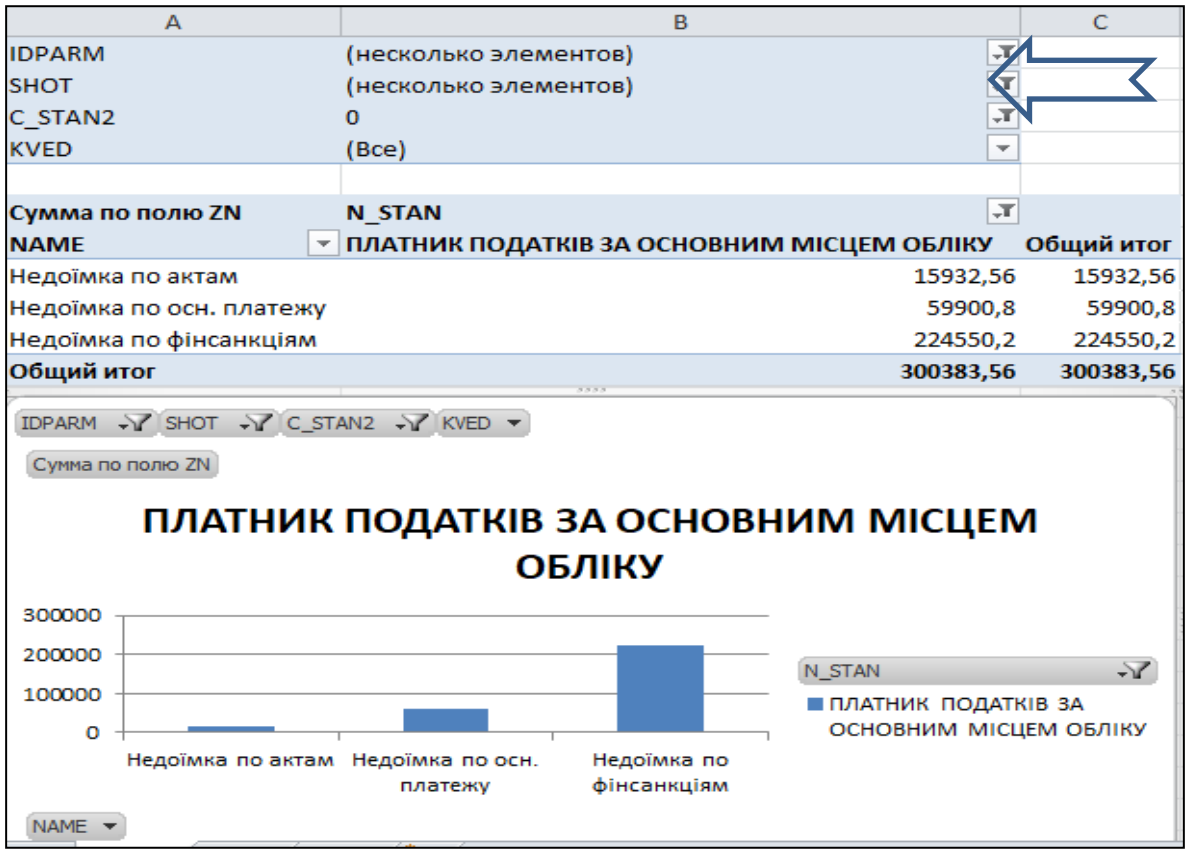


Рис. 2.24. Результат побудови зведеної таблиці та зведеної діаграми

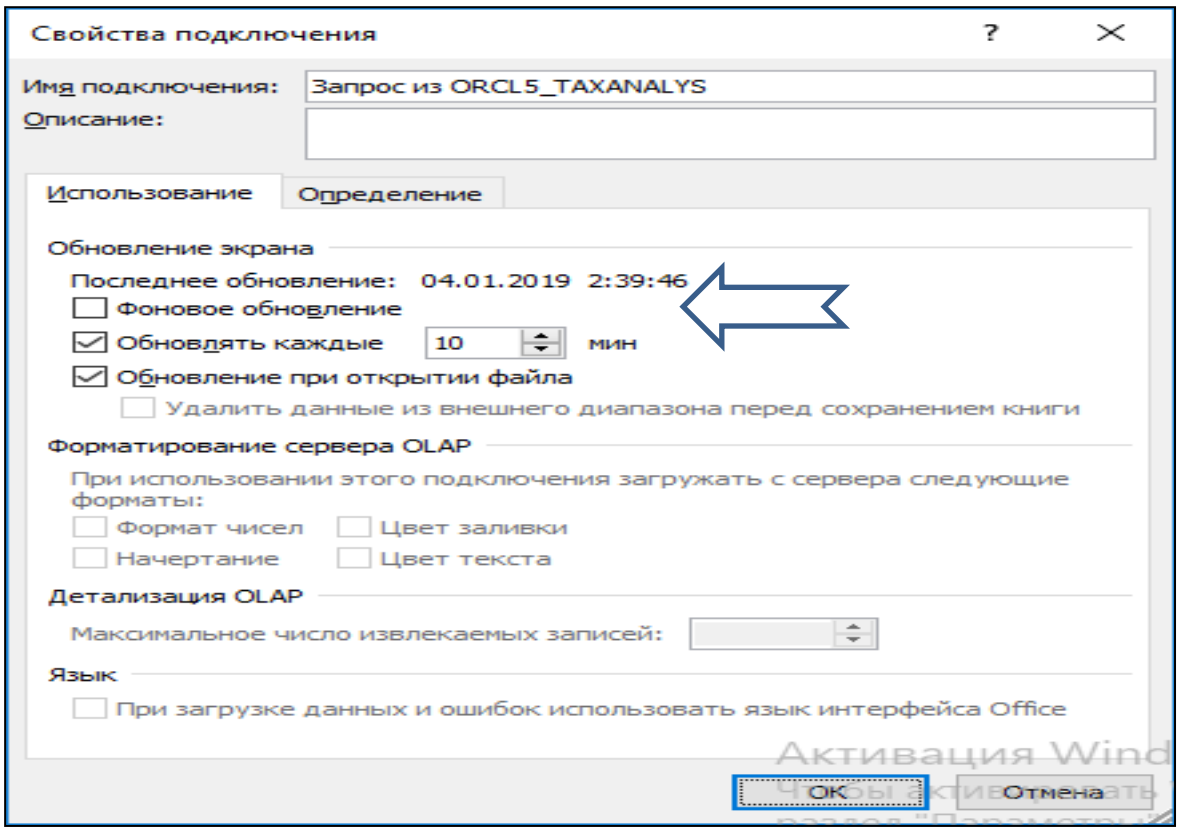


Рис. 2.25. Налаштування оновлення даних з БД Oracle кожні 10 хвилин

4. Формування додаткових обчислювальних об’єктів зведеної таблиці (додаткових зведених показників).

При аналізі потрібно формувати додаткові звітні показники на основі даних таблиць баз даних. Зазвичай це виконується написанням спеціального SQL запиту. Проте за наявності сформованого джерела даних Oracle і зведеної таблиці, що формується на основі даного джерела, можна формувати додаткові поля (додаткові обчислювані об’єкти), що розраховуються на основі наявних полів у зведеній таблиці.

Наприклад, потрібно розрахувати питому вагу заборгованості кожного платника в сумі нарахованих платежів, як «Коефіцієнт заборгованості» (скорочено «Коеф. заб»). Тобто потрібно підрахувати суму показників IDPARM [119] («Недоїмка з основного платежу»), IDPARM [120] («Недоїмка за актами перевірок»), IDPARM [123] («Недоїмка з фін. санкцій») та частку цієї суми у значенні IDPARM [7] («Нарахованого всього (+)»).

4.1. У форматі показників операцій IDPARM це можна записати так:

$$\text{Коеф. заб.}_i = \sum \text{IDPARM}_i\,[119+120+122+123] / \text{IDPARM}_i.$$

де Коеф. заб.$_i$ – коефіцієнт заборгованості і-го платника;

$\sum$ IDPARM$_i$ [119+120+122+123] – сума недоїмки і-го платника;

IDPARM$_i$ [7] – значення «нараховано платежів всього» і-го платника;

i – (1-N), N – кількість платників.

4.2. Розмістити показники у конструкторі та сформувати зведену таблицю, як показано на рис. 2.26;

4.3. Для формування додаткового обчислюваного елементу зведеної таблиці «Коеф. заб» – виконуємо такі дії:

4.3.1. Відмічаємо комірку з назвою будь-якого стовпця у меню Excel;

4.3.2. Вибираємо пункт «Параметры»> «Вычисления»> «Поля, элементы, наборы»> «Вычисляемый объект» (рис. 2.26);

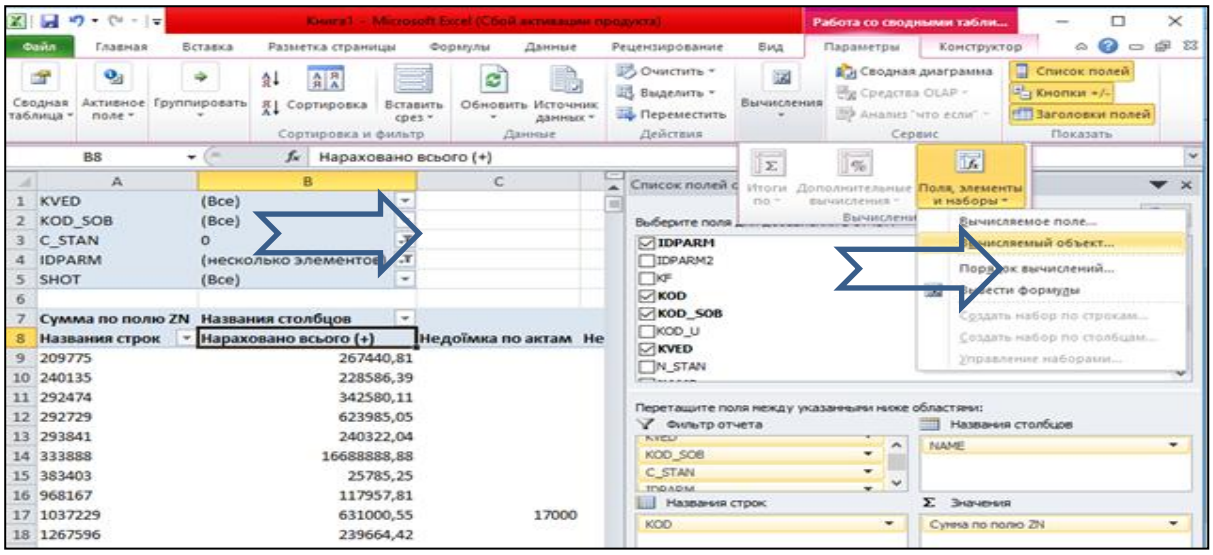


Рис. 2.26. Формування додаткових обчислюваних елементів зведеної таблиці

4.3.4. У вікні «Вычисляемый объект» вводимо назву додаткового обчислюваного об’єкта і в полі «Формула» вводимо формулу у такому форматі (рис. 2.27):

= ('Недоїмка з осн. платежу'+ 'Недоїмка за актами' + 'Недоїмка за фінсанкціями')/ 'Нарахованого всього+'.

Якщо у поля потрібно вивести коди операцій, вводиться така формула:

=IDPARM['119']+ IDPARM['120']+ IDPARM['122']+ IDPARM['123'])/ IDPARM['7']

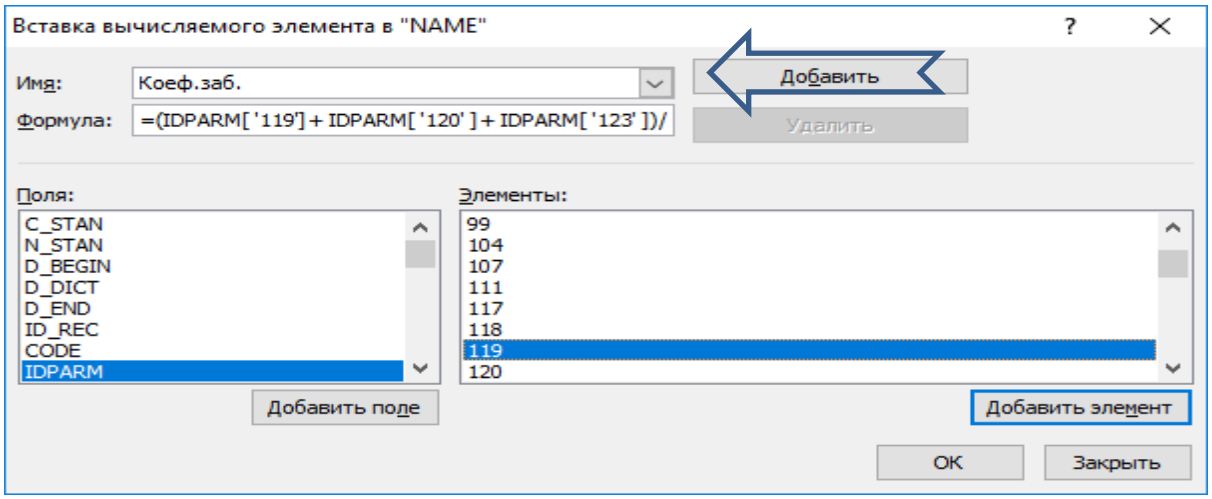


Рис. 2.27. Введення формули для обчислення показника

4.3.4. Результат обчислення додаткового об’єкта «Коеф. заб.» показано на рис. 2.28.

KOD_SOB	(Все)					
C_STAN	0					
IDPARM	(несколько элементов)					
SHOT	(Все)					
KVED	(Все)					
Сумма по полю ZN	Названия столбцов					
Названия строк	Нараховано всього (+)	Недоїмка по актам	Недоїмка по осн. платежу	Недоїмка по фінсанкціям	Коеф.заб.	Общий итог
137:	888559,66	311125,00	1757428,49	838858,92	3,27	3795975,34
868	650248,19		1745246,33	111899,25	2,86	2507396,63
868	247521,56		926198,60	226292,82	4,66	1400017,64
235	296356,72		887534,16	321347,28	4,08	1505242,24
143	143150,73		549083,08	206100,71	5,28	898339,80
322	856622,60		296706,78	127101,97	0,49	1280431,84

Рис. 2.28. Зведена таблиця із додатково розрахованим полем – обчислюваним об’єктом – «Коефіцієнт заборгованості» («Коеф. заб.») за введеною формулою

Для аналізу показників заборгованості у розрізі видів діяльності, коди яких відображаються у полі «KVED», можна це поле перенести в область «Названия строк», а «KOD» відповідно в область «Фильтр отчета». Результуючу таблицю із відсортованими значеннями недоїмки з основного платежу показано на рис. 2.29.

	A	B	C	D	E	F	G
1							
2	KOD_SOB	(Все)					
3	C_STAN	0					
4	IDPARM	(несколько элементов)					
5	SHOT	(Все)					
6	KOD	(Все)					
7							
8	Сумма по полю ZN	Названия столбцов					
9	Названия строк	Нараховано всього (+)	Недоїмка по актам	Недоїмка по осн. платежу	Недоїмка по фінсанкціям	Коеф.заб.	Общий итог
10	40.30.0	1915682,99	311125,00	2645628,58	1160206,20	2,15	6032644,92
11	20.51.0	654673,04		1749121,97	112595,60	2,84	2516393,45
12	31.20.1	249552,76		926198,60	226292,82	4,62	1402048,80
13	45.21.1	3267812,07	97869,14	881879,99	381380,71	0,42	4628942,33
14	70.20.0	3379233,43		298017,69	127280,53	0,13	3804531,78
15	90.03.0	687665,64		213917,36	7689,40	0,32	909272,72
16	01.25.0	918899,36	6767,00	93984,00	3383,50	0,11	1023033,97

Рис. 2.29. Зведена таблиця показників боргу у розрізі видів діяльності

Висновки до отриманих результатів:

1. Отримані результати дають можливість сформувати такі висновки: найбільшу суму боргу з податку на прибуток (SHOT=) серед платників за основним місцем реєстрації (C_STAN=0) з найбільшими сумами нарахувань мають такі п’ять платників: 137, 868, 235, 143, 322 з відповідними коефіцієнтами заборгованості (коди наведено умовні).

2. Серед видів діяльності найбільшу заборгованість з податку на прибуток серед підприємств за основним місцем реєстрації мають такі види діяльності: 40.30.0; 20.51.0; 31.20.1.

4.3.5. Підготуйте звіт про виконання роботи і розмістіть його для оцінювання в СДН Moodle.

Питання для самоконтролю

1. Поясніть суть технології ODBC.
2. Що означає створити «Джерело даних» з БД Oracle?
3. Який порядок формування джерела даних у пакеті Excel?
4. Чому потрібно зв’язувати таблиці з конкретним полем перед аналізом?
5. Що означає поняття «Первинний ключ», «Зовнішній ключ»?
6. Що означає поняття «Зведена таблиця»?
7. Як формуються сторінки вимірів при формуванні зведеної таблиці?
8. Який зміст має IDPARM із кодом 119? У чому його суть?
9. Який порядок внесення формул у зведеній таблиці?
10. Як налаштовується періодичне оновлення даних з БД Oracle у зведеній таблиці Excel?
11. Що використовується як серверна платформа ДФС України? У чому її основні переваги?
12. У чому суть загальної схеми обліку, накопичення, пошуку та надання інформації в інформаційній системі податкового відомства України?
13. Дайте визначення СУБД Oracle.
14. У чому відмінність мови SQL від PL/SQL?

15. Які вимоги пред’являють до баз даних, а також до СУБД?

16. Що означає технологія клієнт-сервер? Поясніть її переваги.

17. Суть поняття СХЕМА в БД Oracle.

18. У чому відмінність ПРОЦЕДУРИ (PROCEDURE) та ФУНКЦІЇ (FUNCTION) як об’єктів БД Oracle?

19. Із чого складається ПАКЕТИ (PACKAGE) як програма для БД Oracle?

20. У чому відмінність ПРЕДСТАВЛЕННЯ (VIEW) та ПОСЛІДОВНІСТІ (SEQUENCE)?

21. Для чого використовується СИНОНІМ (SYNONYM)?

22. Як пояснити, що таке ТРАНЗАКЦІЯ (TRANSACTION)?

23. Охарктеризуйте Oracle Application Express (APEX). Його переваги, недоліки.

Лабораторна робота 2.2
Організація OLEDB доступу до БД Oracle із пакета Statistica та проведення дисперсійного аналізу даних

Мета роботи: отримання навичок у використанні OLEDB інтерфейсу для доступу до БД Oracle із середовища пакета Statistica та проведення дисперсійного аналізу економічної інформації, придбання студентами навичок роботи щодо виявлення впливу на статистичну ознаку за допомогою модуля ANOVA в пакеті Statistica.

Завдання роботи

1. Сформувати OLEDB з'єднання з БД Oracle із середовища Statistica.
2. Вибрати із БД Orcle дані, вказані у варіанті.
3. Провести однофакторний дисперсійний аналіз даних обсягу реалізованої продукції (бази оподаткування) в середовищі Statistica.
4. Провести багатофакторний дисперсійний аналіз.
5. Підготувати звіт і розмістити у середовищі СДН Moodle.

Вихідні дані для виконання роботи

1) IP адреса серверу БД Oracle, де розміщуються доступні користувачу об'єкти 192.168.4.5;

3) унікальне ім'я екземпляра Oracle: SID (SERVICE_NAME) = ORCL;

4) Аліас (ім'я) з'єднання з БД Oracle, прописаний у файлі *tnsnames.ora* -ORCL5;

5) структуру таблиці подано у підрозділі 2.4).

Основні відомості про дисперсійний аналіз

Дисперсійний аналіз використовується у статистичних дослідженнях з метою виявлення впливу на статистичну ознаку, що вивчається та яка не має кількісного виразу. Суть дисперсійного аналізу полягає в розкладанні загальної варіації результуючого показника на компоненти, які відображають вплив чинника (чинників), що вивчається (що вивчаються), і неврахованих чинників.

Залежно від кількості досліджуваних чинників моделі дисперсійного аналізу прийнято підрозділяти на одно-, дво- та багатофакторні.

За допомогою однофакторного дисперсійного аналізу досліджується вплив одного чинника. Дані для проведення аналізу отримують у процесі проведення спеціальним чином організованих досліджень. Початковими даними є значення результуючої ознаки, зафіксовані при ***m*** рівнях чинника. Кількість спостережень при кожному рівні чинника дорівнює **n.**

За допомогою багатофакторного дисперсійного аналізу досліджується вплив на результуючий показник двох або більше чинників. Чинники встановлюють, проводячи процедуру багатофакторного дисперсійного аналізу, при цьому варіація результуючої ознаки розкладається на компоненти, що відображають вплив на неї досліджуваних і неврахованих чинників. Початкові дані можна фіксувати на m рівнях чинника A і на n рівнях чинника B. Загальна кількість груп спостережень дорівнює mn. Для спрощення в кожній групі фіксується по одному спостереженню, тому загальне число спостережень буде дорівнювати N = mn. При проведенні однофакторного дисперсійного аналізу перевіряється статистична гіпотеза:

$$H_0: a_1 = a_2 = \ldots = a_m \text{ або } H_1: a_1 \neq a_2 \neq \ldots \neq a_m,$$

де a_1, a_2, …, a_m – генеральні середні значення результуючої ознаки, зафіксовані при першому, другому, *m*-ому рівні фактора.

Порядок виконання лабораторної роботи

Завдання 1. Формування ODBC з’єднання з БД Oracle.

1.1. Для створення з’єднання з БД Oracle із програми Statistica потрібно виконати такі процедури:

– вибрати пункт «Data» у головному меню програми Statistica та закладку «External Data»;

– вибрати постачальника даних, у нашому випадку це Microsoft OLE DB Provider for Oracle (рис. 2.30).

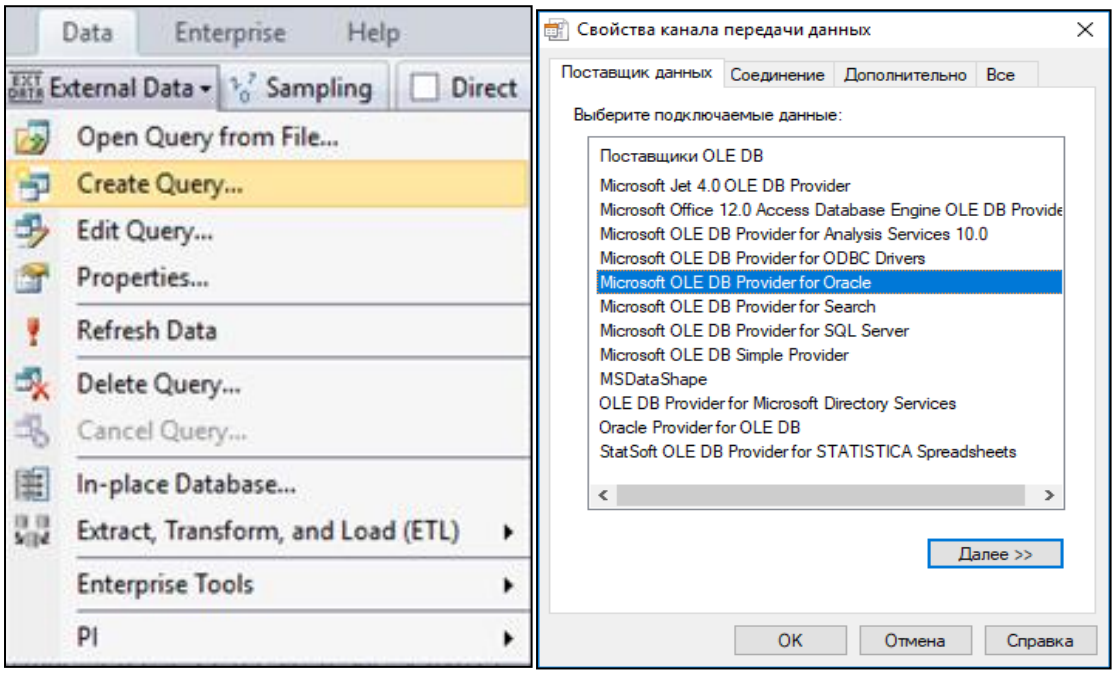


Рис. 2.30. Вибір постачальника даних OLE DB для доступу до БД Oracle

1.2. У вікні «Свойства канала передачи данных» вказати ім’я сервера, у ролі якого можна вказати аліас, прописаний у файлі *tnsname.ora*, та ввести довільну назву з’єднання з БД (п. 2.1.4). Введемо *ORCL5* як ім’я сервера і *ORCL5-TAXANALYS* як назву з’єднання з БД та користувача із відповідними правами на об’єкти БД (рис. 2.31). Важливо пам’ятати, що користувач має привілеї на конкретні об’єкти БД (таблиці, уявлення, процедури тощо, визначені адміністратором БД).

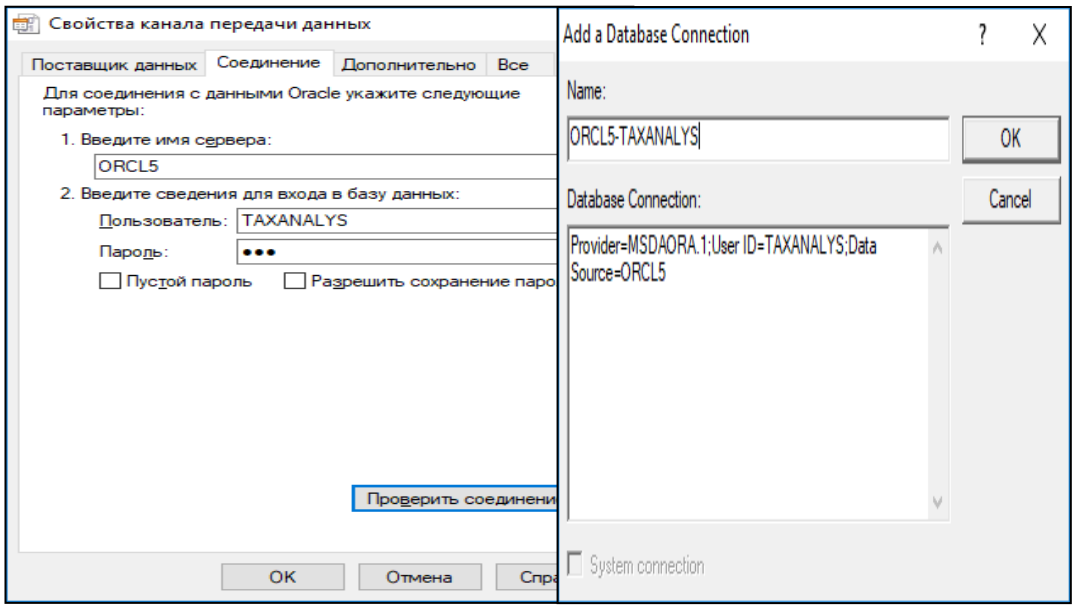


Рис. 2.31. Введення параметрів OLEDB з’єднання з БД

Завдання 2. Вибір даних із БД Oracle.

Відібрати потрібні (вказані у варіанті) для роботи поля таблиці у вікні конструктора запитів, об'єднати їх за відповідними ключовими полями та відмітити потрібні поля, які відображаються у закладці «Preview Data» (на рис. 2.32 показано приклад під'єднання до схеми АІС «Галузь»).

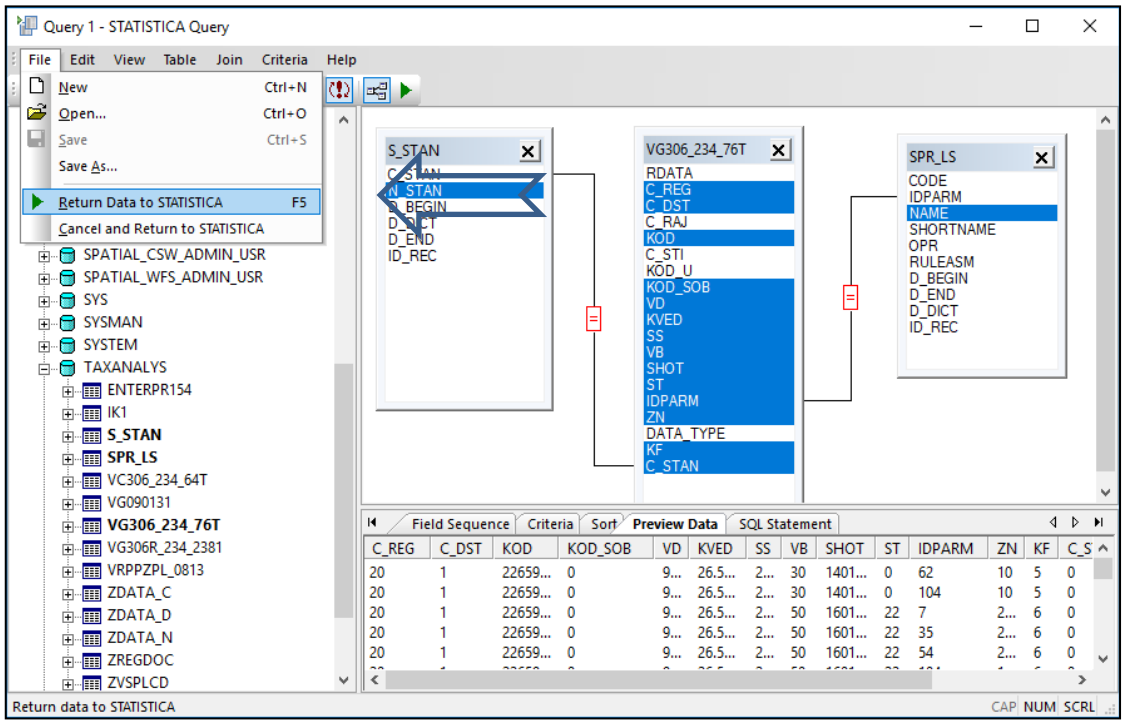


Рис. 2.32. Вікно вибору-зв'язування таблиць та вибору потрібних показників для аналізу у середовищі Statistica

Передати дані у робочий лист клавішею F5 або вибрати пункт «Return Data to Statistica» у головному меню «File», (рис. 2.32).

Таким чином, аналітику стануть доступні всі можливості модулів статистичного пакета Statistica, зокрема: Статистика, *Data Mining, Machine Learning, Neural Networks та ін.*

Завдання 3. Однофакторний дисперсійний аналіз

3.1. Виберіть із БД Oracle дані обсягу реалізованої продукції, отриманої в чотирьох торгових місцях та розташованих в чотирьох районах міста за чотири дні тижня (назву схеми, варіант та назву таблиць дає викладач).

3.2. Для проведення дисперсійного аналізу використовується модуль ANOVA ППП Statistica. Вихідні дані для проведення аналізу подано в табл. 2.11.

Таблиця 2.11

Початкові дані для проведення дисперсійного аналізу

Район	Обсяг реалізації продукції, у.о.			
	Пн.	**Вт.**	**Ср.**	**Чт.**
"А"	14	15	14,8	15
"Б"	14,4	14,9	14,9	15,5
"В"	14,2	15,2	14,6	15,4
"Г"	14,5	15	14,7	15,2

Даними є значення обсягу реалізованої продукції, отриманої в чотирьох торгових місцях та розташованих у чотирьох районах міста за чотири дні тижня. Умовно райони міста назвемо "А", "Б", "В", "Г". Перевіряється вплив розташування торгового місяця на обсяг реалізованої продукції.

3.3. При неможливості отримання даних з БД Oracle створіть файл розмірності 3x16. Перша змінна файла буде призначена для фіксації району розташування торгового місяця, друга – для фіксації дня тижня (знадобиться надалі для проведення двофакторного дисперсійного аналізу), третя змінна – для фіксації обсягу реалізованої продукції (рис. 2.33).

3.4. У пункті меню «Статистика» виберіть пункт ANOVA (від analyze of variance), який призначений для проведення процедури як однофакторного, так і багатофакторного ТАК. У результаті буде відкрито стартове вікно модуля (рис. 2.34).

	1 Розміщення торгового місця	2 День тижня	3 Обсяг РП
1	А	Пн.	14
2	А	Вт.	15
3	А	Ср.	14,8
4	А	Чт.	15
5	Б	Пн.	14,4
6	Б	Вт.	14,9
7	Б	Ср.	14,9
8	Б	Чт.	15,5
9	В	Пн.	14,2
10	В	Вт.	15,2
11	В	Ср.	14,6
12	В	Чт.	15,4
13	Г	Пн.	14,5
14	Г	Вт.	15
15	Г	Ср.	14,7
16	Г	Чт.	15,2

Рис. 2.33. Початкові дані для проведення дисперсійного аналізу

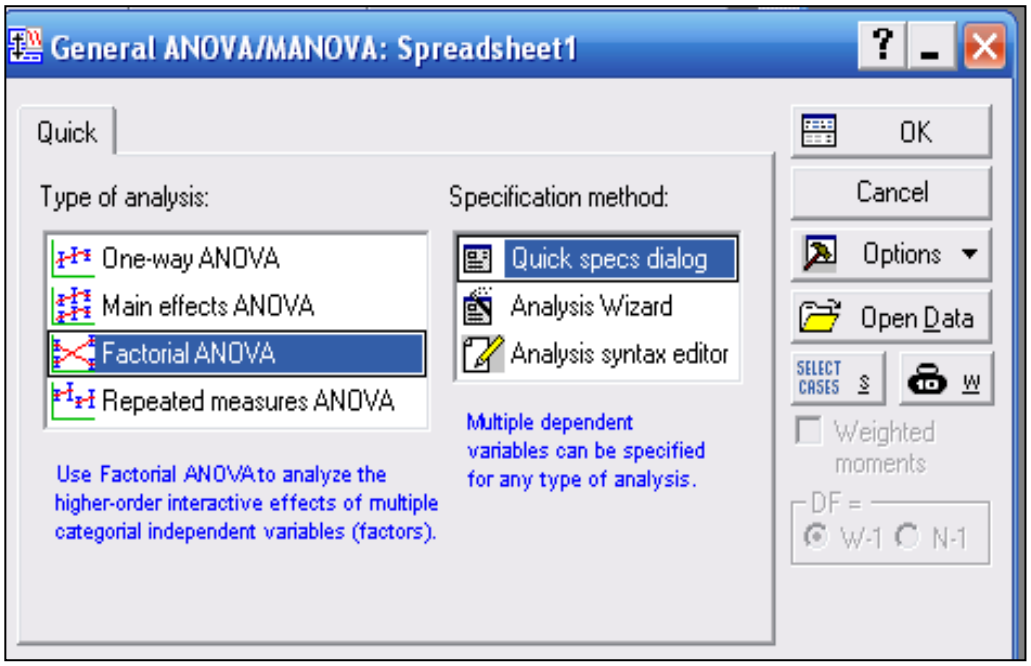


Рис. 2.34. Стартове вікно модуля ANOVA

3.5. Спочатку проведемо процедуру однофакторного дисперсійного аналізу. Для цього встановимо тип аналізу One-way ANOVA і метод Quick specs dialog. Відкриється вікно установки змінних та опцій аналізу (рис. 2.35).

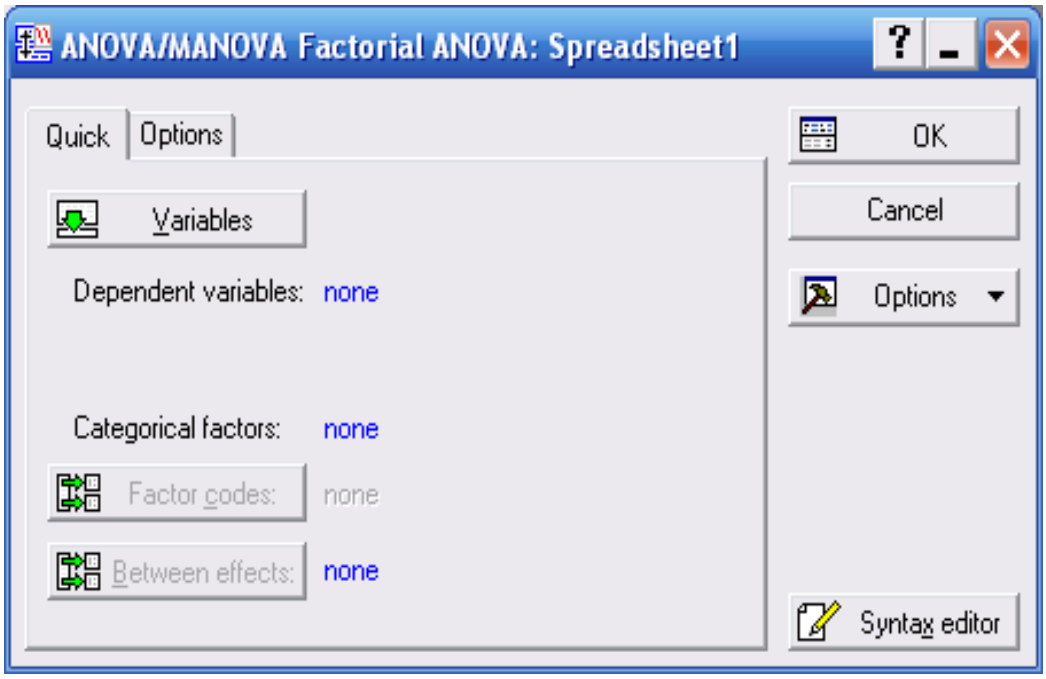


Рис. 2.35. Вікно установки змінних та опцій аналізу

3.6. У діалоговому вікні виберемо факторну змінну "Розміщення торгового місця", а результуюча ознака – "Обсяг реалізованої продукції" (рис. 2.36).

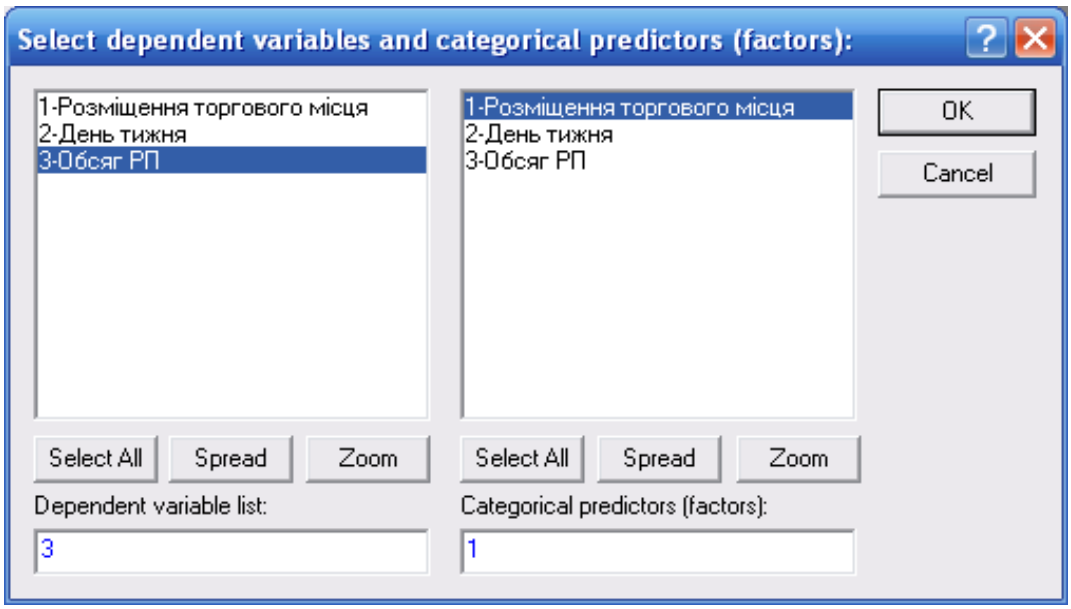


Рис. 2.36. Вікно вибору змінних для аналізу

3.7. Після вибору досліджуваних ознак відкривається вікно результатів аналізу (рис. 2.37).

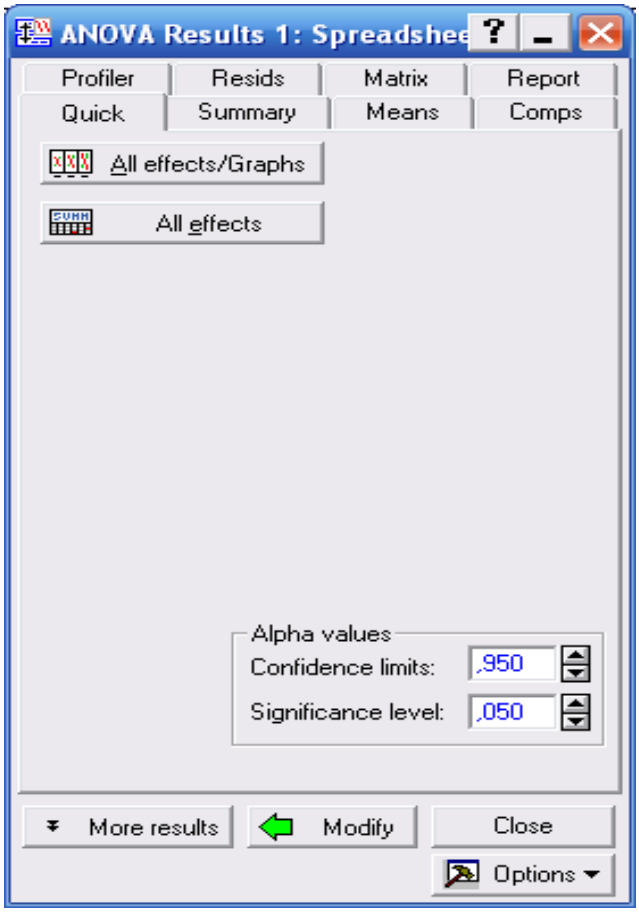


Рис. 2.37. Вікно результатів аналізу

3.8. Залишаємо рівень значущості "Significance level" рівним α=0,05 для розрахунку компонент варіації і перевірки гіпотези з використанням статистики Фішера (рис. 2.38).

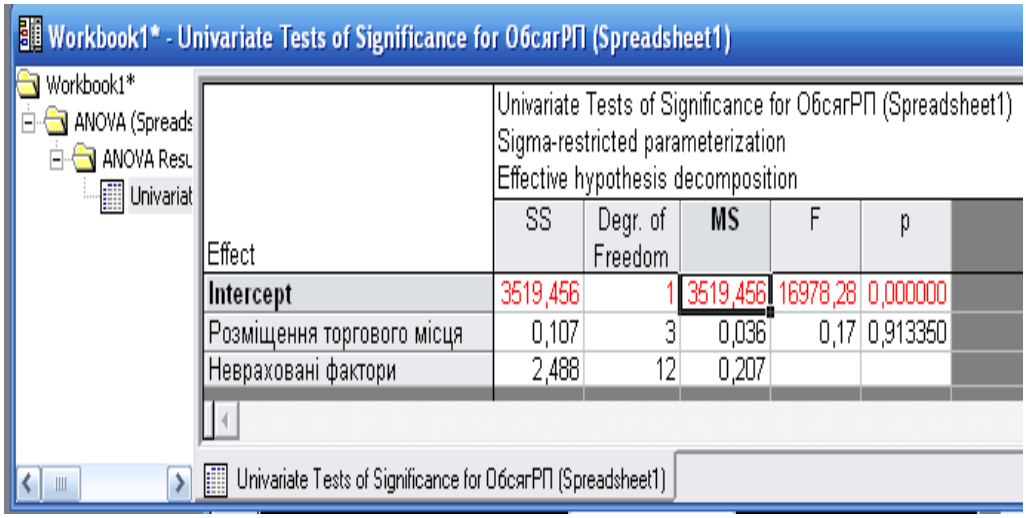


Рис. 2.38. Вікно результатів однофакторного дисперсійного аналізу

3.9. Проаналізуємо отримані результати:

1. Системою розраховано компоненти "Сума квадратів" (SS) Q_1^2 =0,107 та Q_2^2 =2,488 загальної варіації Q^2, які обумовлені чинником "Розташування торгових місць" та неврахованих чинників.

2. Визначено число мір свободи (Degr. of Freedom) кожної компоненти. Зокрема, для компоненти варіації, обумовленої факторною змінною, число мір свободи дорівнює m–1=3, а для компоненти, обумовленою неврахованими чинниками, число мір свободи дорівнює m(n–1)=12.

3. Розраховано величини "Середній квадрат" (MS) $Q_1^2/(m-1) = 0{,}036$ та $Q_2^2/m(n-1) = 0{,}207$.

4. Розраховано величину статистики Фішера, $F = \dfrac{Q_1^2/(m-1)}{Q_2^2/m(n-1)} = 0{,}17$ та визначила вірогідність p=0,913350. При правдивій гіпотезі H_0 величина F може прийняти такі ж або більші значення. Оскільки ця вірогідність істотно більше рівня значущості α = 0,05, то гіпотезу про значущість впливу чинника "Розташування торгових місць" на обсяг реалізованої продукції можна не приймати.

Завдання 4. Багатофакторний дисперсійний аналіз (на прикладі двофакторного).

4.1. У стартовому вікні модуля ANOVA встановимо тип аналізу Main effects ANOVA. У відкриваючому діалоговому вікні виберемо результуючу ознаку і змінні чинники (рис. 2.39).

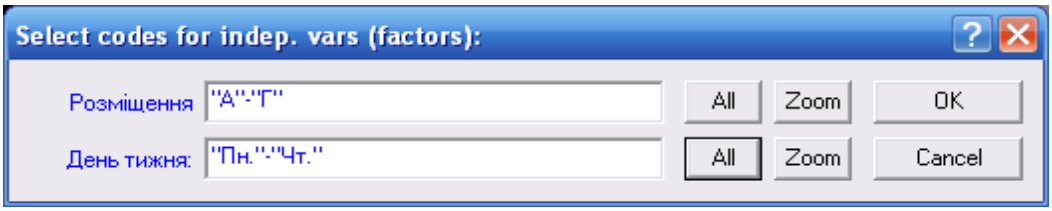


Рис. 2.39. Вікно результатів двофакторного дисперсійного аналізу

На рис. 2.40 подані установки рівнів змінних чинників.

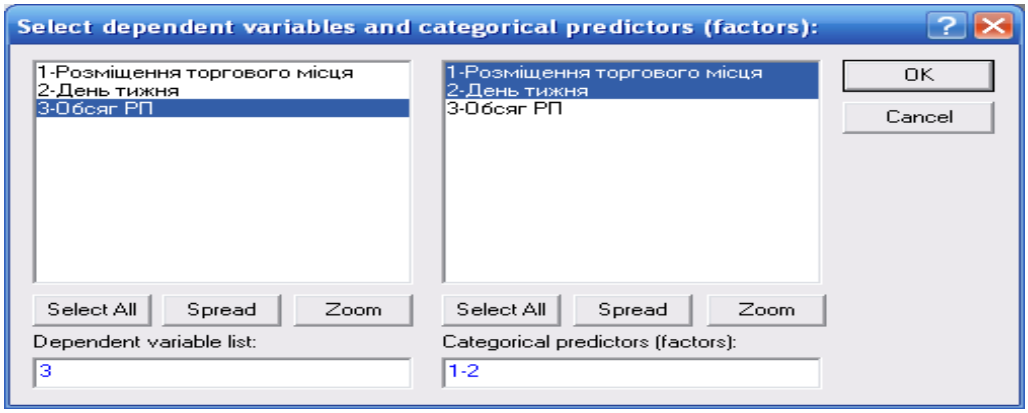


Рис. 2.40. Вікно установки рівнів змінних чинників

У вікні, що відкрилось, результат аналізу залишаємо рівним значущості α = 0,05, після чого відкривається вікно із результатами розрахунку компонентів варіації та перевірки гіпотез з використанням статистик Фішера (рис. 2.41).

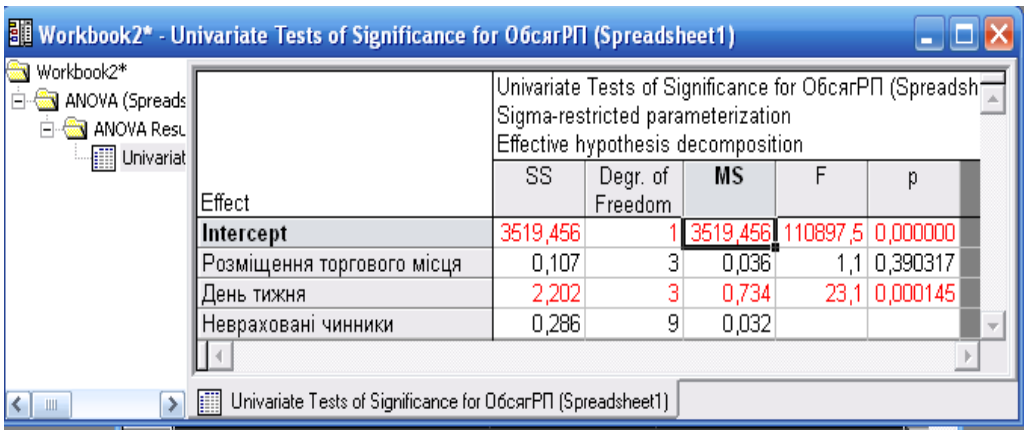


Рис. 2.41. Вікно результатів двофакторного дисперсійного аналізу

4.2. Проаналізуємо результати:

1. Система розрахувала компоненти "Сума квадратів" Q_1^2 = 2,202 та Q_2^2 = 0,286 загальній варіації Q^2, обумовлені вибраними і неврахованими чинниками.

2. Визначила число мір свободи кожної компоненти. Зокрема для компоненти варіації, обумовленої змінною "Розташування торгових місць", число мір свободи дорівнює m–1=3, для компоненти, обумовленої змінною "День тижня", число мір свободи дорівнює n–1=3, а для компоненти, обумовленої неврахованими чинниками, число мір свободи дорівнює (m–1)(n–1)=9.

3. Розрахуємо величини "Середній квадрат":

$Q_1^2/(m-1) = 0,036$;

$Q_2^2/(n-1) = 0,734$ та

$Q_3^2/((m-1)(n-1)) = 0,032$ відповідно.

4. Розрахуємо значення статистик Фішера:

$$F_A = \frac{Q_1^2/(m-1)}{Q_3^2/(m-1)(n-1)} = 1,1;$$

$$F_B = \frac{Q_2^2/(n-1)}{Q_3^2/(m-1)(n-1)} = 23,1$$

та визначимо відповідні значення F_A і F_B при вірогідності p_A=0,390317 і p_B=0,000145 того, що при справедливості гіпотез H_0 величини F_A і F_B можуть прийняти такі ж або більші значення.

Оскільки вірогідність p_A істотно більше, а вірогідність p_B істотно менше рівня значущості α=0,05, то гіпотезу про значущість впливу чинника "Розташування торгових місць" на обсяги реалізованої продукції можна не приймати, а гіпотезу про значущість впливу чинника «День тижня» на обсяги реалізованої продукції необхідно прийняти.

Завдання 5. Дайте відповідь на питання для самоконтролю. Підготуйте звіт про виконання лабораторної роботи і розмістіть у СДН Moodle для оцінювання. У звіті поясніть результати дисперсійного аналізу та дайте свої висновки.

Питання для самоконтролю:

1. У чому відмінність технологій доступу до БД ODBC І OLEDB?

2. Поясніть порядок під'єднання до БД Oracle із пакета Statistica.

3. У чому відмінність технологій обробки даних Oracle Data Miner і Statistica?

4. Що означає дисперсійний аналіз. Що оцінюється при його проведенні?

5. Що може бути цікавим для податкового відомства в отриманій інформації щодо результатів дисперсійного аналізу денного виторгу підприємством?

Лабораторна робота 2.3
SQL запити до БД АІС «Галузь»
(Oracle SQL Developer)

Мета: отримання практичних навичок роботи з БД Oracle в середовищі Oracle SQL Developer, з'єднуватися з екземпляром бази даних Oracle, здійснювати вибірки даних за допомогою мови запитів SQL. Ознайомлення із фізичною та логічною структурою АІС «Галузь», «Бест звіт» ДФС України, надання студентам навичок у формуванні аналітичних запитів до баз даних.

Завдання

1. Під'єднатися до навчальної БД АІС «Бест звіт» із вказаними викладачем параметрами. Виконати приклади 2.10–2.25 розділу 2.4.

2. Ознайомитися із наданими нижче скриптами, виконати їх. За наявності помилок відкоригувати, пояснити свої дії; підготувати запити до АІС «Галузь» за наданими у варіанті завданнями.

3. Підготувати звіт про виконання роботи і розмістити в середовищі ДН за адресою http://moodle.nusta.edu.ua

Порядок виконання лабораторної роботи

Завдання 1. Під'єднання до БД Oracle із Oracle SQL Developer

Нове з'єднання з БД Oracle можна зробити з меню Oracle SQL Developer File | New або із вікна Connection (потрібно натиснути правою кнопкою миші по контейнеру Connections (+) і в контекстному меню вибрати кнопку New). Відкриється вікно створення нового з'єднання (рис. 2.42).

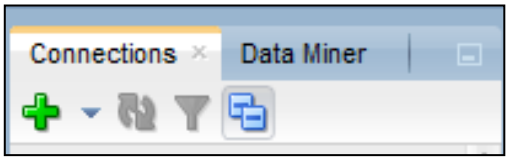


Рис. 2.42. Вікно створення нового з'єднання

Створіть нове з'єднання з БД Oracle, яка інстальована на сервері, наприклад, із IP адресою 192.168.4.123 з назвою Taxanalys від імені користувача stud_01, який має визначені адміністратором привілеї і доступ до об'єктів БД . У вікні New вводимо такі дані (рис. 2.43):

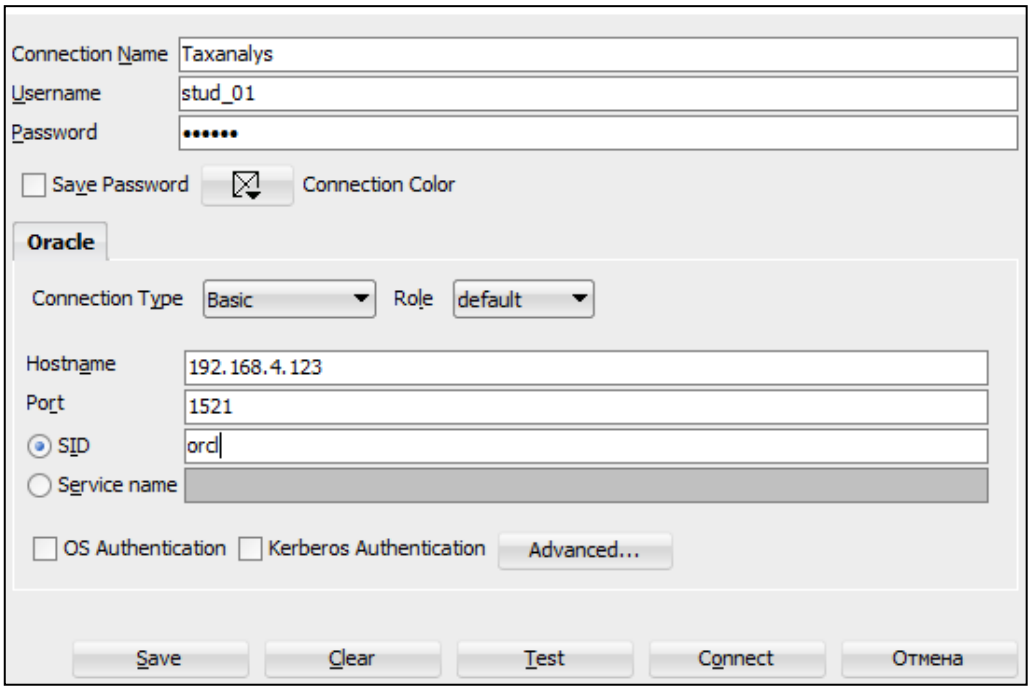


Рис. 2.43. Вікно введення параметрів з'єднання з БД Oracle

Connection Name (Назва з'єднання). У нашому прикладі Taxanalys – псевдонім для підключення до бази даних з використанням інформації, яку ви вводите. (Ім'я з'єднання не зберігається в базі даних, і з'єднання не є об'єктом бази даних).

Username (Ім’я користувача): у нашому прикладі stud_01 – ім’я існуючого користувача бази даних з правами доступу до об’єктів БД. Створюється адміністратором БД.

Password (Пароль): у нашому прикладі може бути, наприклад u01- пароль, встановлений початково адміністратором БД, який може бути змінений даним користувачем.

Save Password (Зберегти пароль): якщо ця опція включена, то пароль буде збережено в інформації про з’єднання, і вам не буде потреби повторно вводити пароль при наступних спробах підключення за допомогою цього з’єднання.

Host Name (Ім’я хоста): у нашому прикладі 192.168.4.123 – IP адреса бази даних Oracle.

Port (Порт): 1521 – порт очікування за умовчанням.

SID (Унікальне ім’я екземпляра БДOracle): у нашому прикладі це **orcl** – надається адміністратором БД).

Після введення параметрів нового з’єднання в цьому вікні можна скористатися кнопкою **Test,** щоб переконатися про успішність з’єднання. При успішному тесті (і при неуспішному) параметри підключення можна зберегти для подальшого використання або налагодження (кнопка **Save**). Для з’єднання із сервером натискаємо кнопку **Connect**.

При з’єднанні з БД відкриється вікно **SQL Worksheet**, у якому можна вводити і виконувати команди SQL, програмні компоненти PL/SQL). Орієнтовний вигляд вікна SQL Worksheet показано на рис. 2.44.

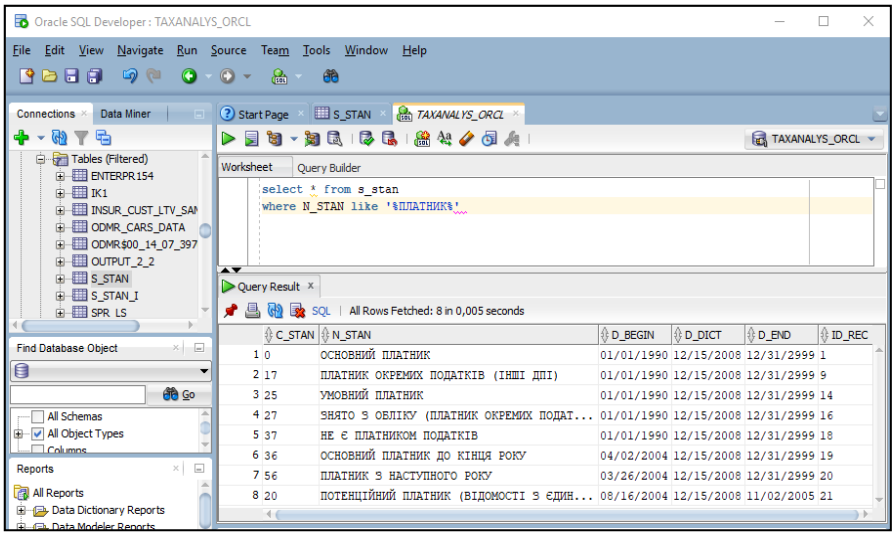


Рис. 2.44. Вікно SQL Worksheet із виконаним запитом SQL до таблиці довідника станів платників s_stan

Метою оптимізації оператора SQL є прискорення доступу до бази даних у її поточному вигляді, використовуючи особливості структури самої бази даних, системних ресурсів та індексів. Основою формування запитів до баз даних АІС «Галузь» є використання відповідних зведених показників з довідника E_SPR_LS «Показники картки особового рахунку платника» (табл. 2.12).

Таблиця 2.12

Основні зведені показники (та їхній зміст) для використання у прикладах

Код зведеного показника (IDPARM)	**Назва зведеного показника згідно з SPR_LS**	**Примітка (де і як може використовуватися)**
119	Недоїмка з основного платежу	Для розрахунку загальної суми податкового боргу (без урахування сум податкового боргу платників «банкрутів»): 119+ 120+ 122+123 + 76. Для розрахунку загальної суми податкового боргу за податковими зобов'язаннями платників без урахування сум податкового боргу платників «банкрутів»: 119+ 120+ 122+123
120	Недоїмка за актами	
122	Недоїмка з відстрочок/ розстрочок	
123	Залишок фінансових санкцій, які підлягають списанню	
76	Недоїмка з пені	
7	Нараховано всього (+)	
8	Нараховано всього (-)	
54	Нараховано або зменшено	
104	Сплачено до бюджету	Для розрахунку загальної суми надходжень: 104 – 58
58	Повернуто з бюджету	
53	Переплата	
62	Переплата – зайво внесена сума податку	

Завдання 1.

1.1.Під'єднатись до схеми з таблицями податкової звітності «Бест звіт». У навчальній БД всі таблиці розміщені у схемі «Taxanalys». Тому при виконанні запитів у скриптах потрібно змінювати аліаси запитів «bestzvit» на «taxanalys».

1.2. Виконайте приклади 2.10.-2.25 підрозділу 2.4.

1.3.Модифікуйте надані приклади і дайте до них своє пояснення.

Завдання 2.

2.1.Під'єднайтесь за допомогою SQL до СКБД Oracle із вказаними викладачем параметрами з'єднання.

2.2. Ознайомтесь із наданими скриптами. Виконайте запити. Поясніть результат. Зафіксуйте свій висновок у звіті.

Задача 1. Вибрати станом на звітну дату в розрізі районів інформацію про нараховані та сплачені суми (по платежу 99999999 «Всього») до Державного бюджету.

Приклад 2.3.1. Рішення. Дані будуть розраховуватися до Державного бюджету та враховувати розмежування платежів (тобто необхідно використовувати довідник D_BDROZM).

Джерело даних – основна таблиця VG306.

Дані на звітну дату складаються з щомісячних (станом на 01 число) та оперативних (на декадні та звітну дату) даних.

Сума надходжень розраховується з таблиці VG306 як сума значень ZN для IDPARM= {104, 58}, при цьому для IDPARM= 104 береться значення ZN, для IDPARM = 58 береться – ZN.

Сума нарахувань розраховується як сума значень ZN для IDPARM=54.

Для включення умови "по платежу 99999999 «Всього»" необхідно враховувати еталонні довідники D_BUDJT_H, E_SPR_LS.

Для отримання назви району можна використовувати функцію RAJ_NAME (c_reg, c_raj).

Запит для обласного (районного) рівня буде мати такий вигляд (вхідний параметр – звітна дата, наприклад, станом на 25.05.2011):

```
SELECT c_reg "регіон", c_raj "район",
                etalon.RAJ_NAME (c_reg, c_raj) "назва району",
                sum (decode( x.idparm,54,zn,0)) "нараховано",
                sum (decode( x.idparm,104,zn,58,-zn,0)) "сплачено"
 FROM
(SELECT b.c_reg, b.c_raj, b.rdata, b.vb, b.shot, b.idparm,
 b.data_type, b.zn
 FROM galuz.VG306 b
        WHERE rdata = ANY (
                                to_date('01052011','ddmmyyyy'),
                                to_date('11052011','ddmmyyyy'),
                                to_date('21052011','ddmmyyyy'),
                                to_date('25052011','ddmmyyyy'))
 AND idparm IN (54,58,104)
 AND vb = 30
 UNION ALL
 SELECT b.c_reg, b.c_raj, b.rdata,
                        c.VB_ROZM AS vb, c.SHOT_ROZM AS shot,
                        b.idparm, b.data_type, b.zn*c.percent_ AS zn
 FROM galuz.VG306 b, etalon.d_bdrozm c
 WHERE rdata = ANY(
                          to_date('01052011','ddmmyyyy'),
                          to_date('11052011','ddmmyyyy'),
                          to_date('21052011','ddmmyyyy'),
                          to_date('25052011','ddmmyyyy'))
        AND idparm IN (54,58,104)
        AND b.vb = 85
        AND C.VB_ROZM = 30
        AND b.c_reg = c.c_reg
        AND b.vb = c.vb
        AND b.shot=c.shot
        AND b.data_type = c.data_type
        AND b.RDATA BETWEEN c.d_begin AND
                                  nvl(c.d_end,trunc(sysdate))
 ) x, ETALON.D_BUDJT_V b, ETALON.E_SPR_LS c
WHERE x.shot=b.shot AND x.idparm=c.idparm
AND b.ruleasm=c.ruleasm
AND B.EXCL=0
AND SHOTTO= 99999999
AND RDATA BETWEEN b.d_budjtkls_b AND b.d_budjtkls_e
GROUP BY C_REG, c_raj
ORDER by C_REG, c_raj
```

Якщо б у завданні була умова – станом на 1 число місяця (з щомісячних даних), наприклад, станом на 01.05.2011, запит мав би вигляд:

```
SELECT c_reg "регіон", c_raj "район",
            etalon.RAJ_NAME (c_reg, c_raj) "назва району",
            sum ( decode( x.idparm,54,zn,0)) "нараховано",
            sum ( decode( x.idparm,104,zn,58,-zn,0)) "сплачено"
 FROM
(SELEC b.c_reg, b.c_raj, b.rdata, b.vb, b.shot, b.idparm,
      b.data_type, b.zn
 FROM galuz.VG306 b
      WHEREe rdata = to_date('01052011','ddmmyyyy')
      AND idparm IN (54,58,104)
      AND vb = 30
 UNION ALL
 SELECT b.c_reg, b.c_raj, b.rdata,
                  c.VB_ROZM AS vb, c.SHOTROZM AS shot,
                  b.idparm, b.data_type, b.zn*c.percent_ AS zn
 FROM galuz.VG306 b, etalon.d_bdrozm c
 WHERE rdata = to_date('01052011','ddmmyyyy')
AND idparm IN (54,58,104)
 AND b.vb = 85
 AND C.VB_ROZM = 30
 AND b.c_reg = c.c_reg
 AND b.vb = c.vb
 AND b.shot=c.shot
 AND b.data_type = c.data_type
 AND b.RDATA BETWEEN c.d_begin AND
                        nvl(c.d_end,trunc(sysdate))
 ) x, ETALON.D_BUDJT_V b, ETALON.E_SPR_LS c
WHERE x.shot=b.shot AND x.idparm=c.idparm
AND b.ruleasm=c.ruleasm
AND B.EXCL=0
AND SHOTTO= 99999999
AND RDATA BETWEEN b.d_budjtkls_b AND b.d_budjtkls_e
group by C_REG, c_raj
ORDER by C_REG, c_raj
```

Задача 2. Вибрати станом на звітну дату в розрізі видів діяльності КВЕД (з ієрархією) інформацію про нараховані і сплачені суми та переплату з податку на прибуток (11020000) до Зведеного бюджету.

Приклад 2.3.2. Рішення. Дані будуть розраховуватися до Зведеного бюджету (сума Державного та місцевих бюджетів) та враховувати розмежування платежів (тобто необхідно використовувати довідник D_BDROZM).

Джерело даних – основна таблиця VG306.

Дані на звітну дату складаються з щомісячних (станом на 01 число) та оперативних (на декадні та звітну дату) даних.

Сума надходжень розраховується з таблиці VG306 як сума значень ZN для IDPARM= {104, 58}, при цьому для IDPARM=104 береться значення ZN, для IDPARM = 58 береться – ZN.

Сума нарахувань розраховується як сума значень ZN для IDPARM=54.

Сума переплат розраховується як сума значень ZN для IDPARM=53.

Для включення умови "по податку на прибуток 11020000" необхідно враховувати еталонні довідники D_BUDJT_H, E_SPR_LS.

Формування даних буде проведено в розрізі КВЕД (з ієрархією), тому обов’язково підключати еталонний похідний довідник D_KVED_GROUP, а назву виду діяльності для повного коду КВЕД отримувати за допомогою функції KVED_NAME(2, KVED).

Запит буде мати такий вигляд (вхідний параметр – звітна дата, наприклад, станом на 25.05.2011):

```
SELECT MKVED "Код КВЕД",
                etalon.KVED_NAME(2,MKVED) "Назва КВЕД",
                sum ( decode( x.idparm,54,zn,0)) "нараховано",
                sum ( decode( x.idparm,104,zn,58,-zn,0)) "сплачено",
                sum ( decode( x.idparm,53,zn,0)) "переплата"
 FROM
(SELECT b.c_reg, b.c_raj, b.rdata, b.vb, b.shot, b.idparm,
          b.data_type, b.zn, b.kved
 FROM galuz.VG306 b
 WHERE rdata = ANY(
                        to_date('01052011','ddmmyyyy'),
                        to_date('11052011','ddmmyyyy'),
                        to_date('21052011','ddmmyyyy'),
                        to_date('25052011','ddmmyyyy'))
              AND idparm in (53, 54,58,104)
              AND vb BETWEEN 30 AND 59
 UNION ALL
 SELECT b.c_reg, b.c_raj, b.rdata,
                  VB_ROZM AS vb, SHOT_ROZM AS shot,
                  b.idparm, b.data_type, b.zn*c.percent_ as zn, b.kved
 FROM galuz.VG306 b, etalon.d_bdrozm c
 WHERE rdata = ANY (
                        to_date('01052011','ddmmyyyy'),
                        to_date('11052011','ddmmyyyy'),
                        to_date('21052011','ddmmyyyy'),
                        to_date('25052011','ddmmyyyy'))
 AND idparm IN (53, 54,58,104)
 AND b.vb = 85
 AND b.c_reg = c.c_reg
 AND b.vb = c.vb
 AND b.shot=c.shot
 AND C.VB_ROZM BETWEEN 30 AND 59
 AND b.data_type = c.data_type
 AND b.RDATA BETWEEN c.d_begin AND
                          nvl(c.d_end,trunc(sysdate))
 ) x, ETALON.D_BUDJT_V b, ETALON.E_SPR_LS c,
etalon.D_KVED_GROUP d
WHERE x.shot=b.shot AND x.idparm=c.idparm
AND b.ruleasm=c.ruleasm
AND B.EXCL=0
AND SHOTTO= 11020000
AND RDATA BETWEEN b.d_budjtkls_b AND b.d_budjtkls_e
AND TRIM( x.KVED) = d.KVED
GROUP BY MKVED
ORDER BY MKVED
```

Основою формування запитів до баз даних АІС «Галузь» є використання відповідних зведених показників з довідника E_SPR_LS «Показники картки особового рахунку платника». У таблиці наведено основні зведені показники (та їхній зміст), які будуть використані у прикладах типових завдань (табл. 2.13):

Таблиця 2.13

Основні зведені показники (та їхній зміст для використання у прикладах типових розрахункових завдань

Код зведеного показника (IDPARM)	**Назва зведеного показника згідно з SPR_LS**	**Примітка (де і як може використовуватися)**
119	Недоїмка з основного платежу	Для розрахунку загальної суми податкового боргу: 119+120+122+123+76+369+370+371+372+ 373+374+375+376+377+379+380+381+382+383+384+385
120	Недоїмка за актами	
122	Недоїмка з відстрочок/розстрочках	
123	Залишок фінансових санкцій, які підлягають списанню	
76	Недоїмка з пені	Для розрахунку загальної суми податкового боргу за податковими зобов'язаннями платників: 119+120+122+123+369+370+371+372+ 377+379+380+381
369	Виділ. борг за актом з осн. пл. (Гос. суд)	
370	Виділ. боргу (ГС) за ШС	
371	Виділ. боргу (ГС) за платежем	
372	Виділ. боргу (ГС) за відст/розстр.	Для розрахунку суми податкового боргу без урахування банкрутів: 119+120+122+123+76
373	Виділ. боргу (ГС) фінсанкій	
374	Виділ. боргу (ГС) пені з КОР	Для розрахунку загальної суми податкового боргу за податковими зобов'язаннями платників без урахування банкрутів: 119+120+122+123
375	Виділ. боргу (ГС) пені за актом	
376	Виділ. боргу (ГС) % за кор.відстр.	
377	Виділ. борг за актом з осн. пл. (моратор)	
379	Виділ. боргу (морат) за ШС	

380 381 382 383 384 385	Виділ. боргу (морат) за платежем Виділ. боргу (морат) за відст/розс. Виділ. боргу (морат) фінсанкій Виділ. боргу (морат) пені з КОР Виділ. боргу (морат) пені за актом Виділ. боргу (морат) % за кор.відс.	**Важливо!** Показник ZN=0 для IDPARM=122 і даний idparm не використовується при розрахунку податкового боргу, але для сумісності з попередніми періодами його потрібно включати у запит
7 8 54	Нараховано всього (+) Нараховано всього (–) Нараховано або зменшено платежів (сальдо)	54 = 7 – 8
104 58 947	Сплачено до бюджету Повернуто з бюджету Поверн.внесених коштів у межах одного бюдж. рахунку (вн.висн)	Для розрахунку загальної суми надходжень: 104 – 58 Для сум, що сплачені: 104 – 947 Для сум, що повернуті: 58 – 947 **Важливо!** При формуванні звітності найчастіше використовується показник «збір» замість «надходжень». При формуванні Збору дані за кодом бюджетної класифікації 14010200 не враховуються, при розрахунку Надходжень – враховуються
53 62 64 693	Переплата Переплата – зайво внесена сума податку Переплата–за рахунок податкових векселів – W Переплата А0	Переплата всього: 53 Переплата, що утворилась за рахунок надміру сплачених сум: 62–693 Переплата, що утворилася за рахунок сум платежів, що сплачені та будуть нараховані в наступному періоді: 693+64

Питання для самоконтролю

1. Що означає визначення станів суб’єктів обліку на підставі документів, які надходять до податкових органів?

2. Що означає аналіз повноти документів, задіяних у визначенні стану платника?

3. Назвіть складові інформаційного забезпечення діяльності органів ДПС України.

4. Які ви можете назвати зведені показники АІС «Галузь».

5. Структура файлу зведених показників АІС «Галузь» та відповідні об’єкти бази даних.

6. Призначення АІС «Галузь».

7. Еталонна схема довідників.

8. Які додаткові об’єкти баз даних можуть бути створеними за даними зведених показників АІС «Галузь»?

9. Які довідники схеми «Еталон» використовуються АІС «Галузь»?

10. Основні правила побудови довідників схеми ETALON СКБД Oracle.

11. Особливості організація реєстрації фізичних осіб у ДРФО. Особливості автоматизації.

Лабораторна робота 2.4
Аналітичні SQL запити до БД АС «Податкова звітність» (Oracle SQL Developer)

Мета: формування навичок у використанні аналітичних функцій Oracle для побудови ефективних запитів до БД ДФС України.

Завдання роботи

1. Ознайомтеся із теоретичним матеріалом підрозділу 2.3. «Особливості застосування аналітичних функцій Oracle».

2. Ознайомтеся із наданими прикладами запитів з використанням аналітичних функцій та виконайте їх у середовищі SQL Developer.

3. Запропонуйте скрипти з використанням аналітичних функцій для аналізу даних АІС «Галузь».

4. Поясніть результати запитів, підготуйте звіт. Розмістіть звіт для оцінювання в СДН Moodle.

План виконання лабораторної роботи

Завдання 2

Приклади, що наводяться, не демонструють усі можливості, а лише дають початкове уявлення. Розглянемо декілька прикладів використання аналітичних функцій за допомогою функції SUM. У процесі виконання своїх завдань потрібно мати статистичні дані щодо всіх касових апаратів із сумами виторгу.

Приклад 2.4.1. Завдання: необхідно підрахувати суму виторгу щодо кожного касового апарату за раніше відомий період. Формується такий запит до БД:

```
SELECT DISTINCT zregdoc.tin,zdata_c.zn fiskal,
SUM(zdata_n.zn) over (PARTITION BY zdata_c.zn) summa
FROM bestzvit.zregdoc, bestzvit.zdata_n, bestzvit.zdata_c
WHERE (zdata_c.cod_regdoc=zdata_n.cod_regdoc
AND zdata_c.cod_regdoc=zregdoc.cod_regdoc
AND zdata_c.line=zdata_n.line
AND zdata_n.c_doc_rowc LIKE'^R000%G4'
AND zdata_c.zn LIKE'0106%')
ORDER BY zregdoc.tin,zdata_c.zn
```

Опис роботи програми

Для використовування в роботі візьмемо зі схеми bestzvit таблиці zdata_c, zdata_n, zdata_d, zregdoc з вивантаженими даними по ЕККА за певний період. У даних таблицях, зокрема в таблиці zdata_n у стовпці c_doc_rowc, розміщені необхідні нам дані. Але оскільки в цій таблиці дані розташовуються в рядках, необхідне застосування одночасно декількох умов:

zdata_c.cod_regdoc=zdata_n.cod_regdoc – для того, щоб зв'язати таблиці zdata_c і zdata_n по стовпцю cod_regdoc, в якому зберігаються фіскальні номери касових апаратів;

zdata_c.cod_regdoc=zregdoc.cod_regdoc – для того, щоб зв'язати таблиці zdata_c і zregdoc по стовпцю cod_regdoc, в якому зберігаються фіскальні номери касових апаратів;

zdata_c.line=zdata_n.line – для того, щоб зв'язати таблиці zdata_c і zdata_n по стовпцю line, в якому зберігаються дані відповідного касового апарату відповідні рядкам поля c_doc_rowc таблиці zdata_n;

zdata_n.c_doc_rowc like'^R000%G4' – оскільки в стовпці c_doc_rowc розташовується опис даних поля zn. У цьому випадку нас цікавлять усі рядки стовпця c_doc_rowc, які починаються на ^R000, а закінчуються на G4, оскільки якщо у платника декілька касових апаратів, то суми виторгу розташовуються в декількох рядках стовпця zdata_n.zn;

zdata_c.zn – у цьому стовпці містяться фіскальні номери касових апаратів, що надали звіт у ДПІ, а для вибору даних по касових апаратах, які перебувають на обліку і здають звіт у ДПІ в м. Керч zdata_c.zn like'0106%'.

Орієнтовний результат вибірки з бази **zdata_n:**

C_DOC_ROWC	ZN	COD_REGDOC	CARD	TAB	LINE
^HZY	2006	1060000523718	523718	0	0
^HZM	1	1060000523718	523718	0	0
^R01G4	1636,5	1060000524307	524307	0	0
^R01G5	1636,5	1060000524307	524307	0	0
^R01G6	272,75	1060000524307	524307	0	0
^R01G4	8265,26	1060000524724	524724	0	0
^R0001G4	1636,5	1060000524307	524307	1	0
^R0001G5	1636,5	1060000524307	524307	1	0
^R0001G6	272,75	1060000524307	524307	1	0
^HZY	2006	1060000523722	523722	0	0
^HZM	1	1060000523722	523722	0	0
^HZY	2006	1060000524312	524312	0	0
^HZM	1	1060000524312	524312	0	0
^R0002G4	7172,46	1060000521612	521612	1	1
^R0002G5	7172,46	1060000521612	521612	1	1

Результат виконання SQL-запиту:

TIN	FISKAL	SUMMA
191493	0106002889	54159,61
1185757	0106002988	673647,62
1185757	0106002989	1126632,18
1185757	0106002991	6510012,68
1185757	0106002992	878242,6
1562562	0106002300	45239,45
1562562	0106002381	16729,93
1884797	0106002046	689198,82
1884797	0106002047	1434413,5
2470268	0106001884	14490,93
2470268	0106001933	16770,8
2470268	0106002077	117853,41
2470268	0106002266	328650,88

Оскільки таблиці **zdata_c, zdata_n, zdata_d, zregdoc** постійно використовуються для різних вибірок з програми «БЕСТ-ЗВІТ for Oracle», для подальшої роботи з даними створимо таблицю **VUTORGS** у табличному просторі ***BESTZVIT*** **(або *TAXANALYS*)** на основі попереднього запиту:

```
CREATE TABLE bestzvit.VUTORGS
AS SELECT DISTINCT zregdoc.tin, zdata_c.zn fiskal,
sum(zdata_n.zn) over
(PARTITION BY zdata_c.zn) summa
FROM bestzvit.zregdoc, bestzvit.zdata_n, bestzvit.zdata_c
WHERE (zdata_c.cod_regdoc=zdata_n.cod_regdoc
AND zdata_c.cod_regdoc=zregdoc.cod_regdoc
AND zdata_c.line=zdata_n.line
AND zdata_n.c_doc_rowc LIKE'^R000%G4'
AND zdata_c.zn LIKE'0106%')
ORDER BY zregdoc.tin,zdata_c.zn
```

Приклад 2.4.2. Завдання: написати запит для виведення максимальної і мінімальної суми виторгу. Приклади запитів подані нижче:

```
SELECT tin,fiskal,summa FROM bestzvit.vutorgs
WHERE summa=(SELECT MAX(summa) OVER FROM
bestzvit.vutorgs)
```

Результат виконання запиту:

TIN	FISKAL	SUMMA
33117746	0106002766	9652199,02

```
SELECT tin,fiskal,SUMMA FROM bestzvit.vutorgs
WHERE summa=(SELECT MIN(summa) OVER FROM
bestzvit.vutorgs)
```

Результат виконання запиту:

TIN	FISKAL	SUMMA
20753291	0106000326	38

Приклад 2.4.3. Завдання: написати запит для підрахунку кількості касових апаратів за кодом

SELECT tin,*COUNT(fiskal)* ***OVER FROM*** *bestzvit.vutorgs* ***GROUP BY*** *tin*

Результат виконання запиту:

TIN	COUNT(FISKAL)
191493	4
381686	8
446210	2
467873	3
700476	10
951994	3
1074673	1

Приклад 2.4.4. Завдання: вибрати суму виторгу за касовим апаратом, суму виторгу по підприємству в цілому, сумарний виторг по всіх підприємствах.

SELECT *fiskal, tin, summa,* ***SUM*** *(summa)* ***OVER*** *(****ORDER BY*** *tin,fiskal)*
itogo_all, ***SUM****(summa)* ***OVER*** *(****PARTITION BY*** *tin* ***ORDER BY*** *fiskal) itogo_po_tin,*
ROW_NUMBER() ***over*** *(****PARTITION BY*** *tin* ***ORDER BY*** *fiskal) number_row_tin*
FROM *bestzvit.vutorgs* ***ORDER BY*** *tin, fiskal*

Результат виконання запиту:

FISKAL	TIN	SUMMA	ITOGO_ALL	ITOGO_ PO_TIN	NUMBER_ ROW_TIN
0106002889	191493	54159,61	54159,61	54159,61	1
0106002929	191493	336548,47	390708,08	390708,08	2
0106003032	191493	24469,61	415177,69	415177,69	3
0106003058	191493	9071,05	424248,74	424248,74	4
0106001620	381686	31306	455554,74	31306	1
0106002429	381686	499729	955283,74	531035	2
0106002637	381686	664072	1619355,74	1195107	3
0106002708	381686	381818	2001173,74	1576925	4
0106002742	381686	431800	2432973,74	2008725	5
0106002746	381686	172439	2605412,74	2181164	6
0106002829	381686	186511	2791923,74	2367675	7
0106003081	381686	55177	2847100,74	2422852	8
0106000693	446210	30574	2877674,74	30574	1
0106002687	446210	53471,34	2931146,08	84045,34	2

Опис роботи програми: у представленому вище коді вдалося отримати значення ITOGO_ALL для запиту в цілому. Це було зроблено по всій впорядкованій результуючій множині за допомогою конструкції SUM(SUMMA) OVER (ORDER TIN,FISKAL). Також вдалося підрахувати проміжні суми по підприємствах, скидаючи їх у нуль при переході до наступного підприємства. Цього вдалося домогтися завдяки конструкції PARTITION TIN в SUM(SUMMA) – у запиті була вказана конструкція, що задає умову розбиття даних на групи.

Для послідовної нумерації рядків у кожній групі, відповідно до критеріїв впорядкування, використовувалася функція ROW_ NUMBER (для видачі цього номера рядка був доданий стовпець NUMBER_ROW_TIN). У результаті видно, що 0106003058 – четвертий за списком касовий апарат у платника

"191493" при впорядкуванні за фіскальним номером (FISKAL). Функція ROW_ NUMBER використовується і в багатьох інших ситуаціях, наприклад для транспонування або перетворення результуючих множин. Цей новий набір функціональних можливостей відкриває абсолютно нові перспективи роботи з даними.

Можна позбутися великого об'єму процедурного коду і складних (або неефективних) запитів, що вимагають багато часу на розробку, і отримати при цьому бажаний результат. Більше того, якщо розглянути синтаксис аналітичних функцій, виявляється, що записувати з їх допомогою запитів набагато простіше, ніж на стандартній мові SQL.

Питання для самоконтролю

1. Поясніть призначення АС «Податкова звітність».
2. Поясніть зміст таблиць zdata_c, zdata_n, zdata_d, zregdoc.
3. У якій схемі бази даних Oracle розміщені таблиці zdata_c, zdata_n, zdata_d, zregdoc?
4. Для чого в прикладі 2.4.1 застосовується умова zdata_c.cod_regdoc=zdata_n.cod_regdoc?
5. Запишіть умову прив'язки таблиць zdata_c і zregdoc по стовпцю cod_regdoc, в якому зберігаються фіскальні номери касових апаратів.
6. Запишіть умову вибірки, якщо необхідно вивести усі рядки стовпця c_doc_rowc, які починаються на ^R000 та закінчуються на G4.
7. Для чого створюється таблиця VUTORGS?
8. Для чого у прикладі 2.4.2 використовується ключове слово OVER ?
9. Яка функція використовується для послідовної нумерації рядків у кожній групі відповідно до критеріїв впорядкування?
10. У яких ситуаціях використовується функція ROW_ NUMBER?

РОЗДІЛ 3
МЕТОДИ ТА ЗАСОБИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ПОДАТКОВОЇ ІНФОРМАЦІЇ

3.1. Огляд сучасних платформ і напрямів розвитку технологій інтелектуального аналізу

3.1.1. Загальні відомості про інтелектуальний аналіз (Data Mining)

У контексті адміністрування податків все більшу вагу набирають *завдання класифікації* суб'єктів господарювання за рівнем податкових ризиків, формування плану перевірок, класифікація регіонів за податковим потенціалом і навантаженням, визначення класів і типів факторів, чинників, що впливають на здійснення податкової політики у регіонах, виявлення подібних за різними показниками суб'єктів господарювання, що ухиляються від оподаткування. Впровадження в податковому відомстві України ризикоорієнтованої системи адміністрування податків є значним прогресом у впровадженні нових методологій управління діяльністю податкового відомства. Цей процес є лише початком для побудови комплексної системи управління новоствореної Державної податкової служби України.

Регресійно-кореляційний аналіз досить активно використовується для прогнозування податкових надходжень з урахуванням сезонних факторів, проте величезні масиви даних потребують нових підходів для їхнього аналізу.

Кластерні методи, методи асоціацій, нейромережні та генетичні алгоритми також можуть розглядатись для застосування в аналітичній роботі функціональних підрозділів ДПС України, а методологія нечітких множин вже використовується для формування плану перевірок. Усе це вказує на те, що аналітичні підрозділи податкового відомства України стоять на порозі змістовної і технологічної модернізації в обробці даних та інформації.

Методи та засоби, що можуть бути використані для впровадження у цій сфері, об'єднані під поняттям Data Mining. За визначенням Data Mining – це процес виявлення в неструктурованих даних раніше невідомих, корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності.

Формально *Data Mining* – це побудова моделі даних. На сьогодні існує декілька підходів до побудови моделей даних, а саме:

– статистичний (англ. *Statistical Modelling*): базується на теорії статистики та зосереджується на перевірці гіпотез;

– на основі машинного навчання (англ. Machine Learning): базується на методах навчання нейромереж;

– обчислювальний (по суті – інтелектуальний аналіз даних): інтеграція теорії статистики та евристик, сконцентрований на єдиному процесі аналізу даних, включає штучний інтелект.

Інтелектуальний аналіз даних (Data Mining) – це компллкекс методів та засобів, призначених для пошуку у великих обсягах даних неочевидних, об'єктивних і потрібних на практиці закономірностей. Неочевидних – тобто таких, що не виявляються стандартними методами обробки інформації або експертним шляхом

Традиційні методи аналізу даних (статистичні методи) і аналітична обробка в реальному часі (*Online Analytical Processing*, далі – OLAP) в основному орієнтовані на перевірку заздалегідь сформульованих гіпотез (*verification-driven data mining*) і на «грубий» розвідувальний аналіз, що становить основу оперативної аналітичної обробки даних, у той час, коли основним завданням Data Mining є пошук неочевидних закономірностей.

Інструменти Data Mining можуть знаходити такі закономірності і будувати гіпотези про взаємозв'язки самостійно. Оскільки формулювання гіпотези щодо залежностей є найскладнішим завданням, перевага Data Mining, порівняно з іншими методами аналізу, є очевидною.

На рис. 3.1 показано зв’язок Data Minig з іншими сферами аналізу інформації.

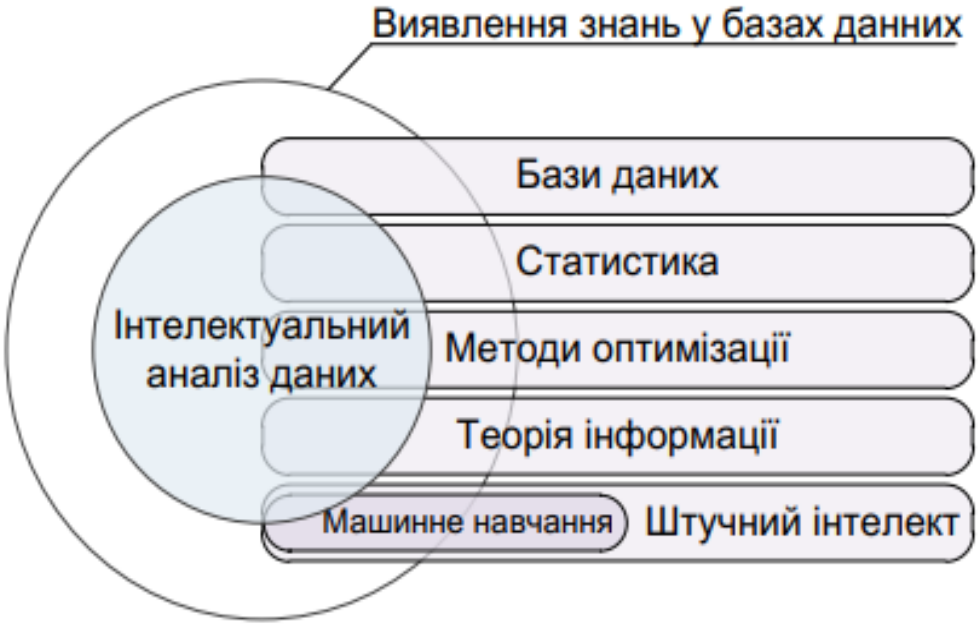


Рис. 3.1. Зв’язок Data Mining з іншими сферами аналізу інформації

Основні задачі, які можна вирішувати методами Data Mining, і їхні можливості застосування в аналітичній діяльності податкового відомства:

– **класифікація** – віднесення об’єктів (спостережень, подій) до одного із заздалегідь відомих класів. (Класифікація суб’єктів господарювання за рівнем податкового ризику, класифікація регіонів за рівнем податкового навантаження тощо);

– **регресія** (у тому числі задачі прогнозування) – встановлення залежності безперервних вихідних від вхідних змінних. (Прогнозування податкових надходжень, встановлення залежності податкових надходжень від різних факторів соціально-економічного розвитку регіону тощо);

– **кластеризація** – угруповання об’єктів (спостережень, подій) на основі даних (властивостей), що описують сутність цих об’єктів, у кластери. Об’єкти всередині кластера повинні бути «подібними» один на одного і відрізнятися від об’єктів, що ввійшли в інші кластери. Чим більше подібні об’єкти всередині кластера і чим більше відмінностей між кластерами, тим точніша кластеризація.

Виявлення груп подібних суб'єктів господарювання за різними показниками фінансово-економічної діяльності);

– **асоціація** – виявлення закономірностей між пов'язаними подіями. Прикладом такої закономірності служить правило, яке вказує, що з події X слідує подія Y. Такі правила називаються асоціативними. Вперше ця задача була запропонована для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її ще називають аналізом ринкової корзини (market basket analysis). (Виявлення закономірностей та причинно-наслідкових правил зв'язків щодо сплати податків);

– **послідовні шаблони** – встановлення закономірностей між пов'язаними в часі подіями, тобто виявлення залежності, що якщо відбудеться подія X, то через заданий час відбудеться подія Y;

– **аналіз відхилень і аномалій** – виявлення найбільш нехарактерних подій, явищ. (Виявлення значних відхилень у сплаті податків типових платників податків, типового виду діяльності тощо).

3.1.2. Класифікація методів Data Mining

Основна особливість Data Mining – це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій. У технології Data Mining гармонійно поєдналися строго формалізовані методи та методи неформального аналізу, тобто кількісний та якісний аналіз даних.

До методів і алгоритмів інтелектуального аналізу даних належать: штучні нейронні мережі, дерева рішень; символьні правила; методи найближчого сусіда і k-найближчого сусіда; метод опорних векторів; байєсовські мережі; лінійна регресія; регресійний-кореляційно аналіз; ієрархічні методи кластерного аналізу; неієрархічні методи кластерного налізу, в тому числі алгоритми k-середніх і k-медіани; методи пошуку асоціативних правил, у тому числі алгоритм Apriori; метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних тощо.

Більшість аналітичних методів, що використовуються в технології Data Mining – це відомі математичні алгоритми. Новим у їхньому застосуванні є можливість їхнього використання при вирішенні тих чи інших конкретних проблем, обумовлена можливостями, які з'явилися завдяки розвитку технічних і програмних засобів. Слід зазначити, що більшість методів інтелектуального аналізу даних були розроблені в рамках теорії *штучного інтелекту.*

Метод (англ. Method) являє собою норму або правило, певний шлях, спосіб, прийом рішень завдань теоретичного, практичного, пізнавального, управлінського характеру.

Усі методи інтелектуального аналізу даних можна розділити на *статистичні на кібернетичні.*

Перша група методів об'єднується в групу **статистичних методів,** до яких входять:

– попередній аналіз природи статистичних даних (перевірка гіпотез стаціонарності, нормальності, незалежності, однорідності, оцінка виду функції розподілу, її параметрів тощо);

– виявлення зв'язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз та ін.);

– багатовимірний статистичний аналіз (лінійний і нелінійний дискримінантний аналіз, кластерний аналіз, компонентний аналіз, факторний аналіз та ін.);

– динамічні моделі і прогноз на основі часових рядів.

Друга група методів інтелектуального аналізу – це множина підходів, об'єднаних **теорією штучного інтелекту.** До цієї групи відносимо такі методи:

– штучні нейронні мережі (розпізнавання, кластеризація, прогноз);

– еволюційне програмування (у т.ч. алгоритми методу групового обліку аргументів);

– генетичні алгоритми (оптимізація);

– асоціативна пам'ять (пошук аналогів, прототипів);

– нечітка логіка;

– дерева рішень;

– системи обробки експертних знань.

Методи Data Mining також *можна класифікувати за задачами, для* вирішення яких вони вибираються. Відповідно до такої класифікації виділяємо дві групи:

– методи на вирішення задач сегментації (тобто задачі класифікації і кластеризації);

– методи для вирішення задач прогнозування.

Відповідно до другої класифікації методи Data Mining можуть бути спрямовані на отримання описових і прогнозуючих результатів.

Перша група – *методи на вирішення задач сегментації (описові методи*) служать для знаходження шаблонів (або зразків), що описують дані, які піддаються інтерпретації з точки зору аналітика. До методів, спрямованих на отримання *описових результатів*, належать *ітеративні методи кластерного аналізу*, в тому числі: *алгоритм k-середніх, k-медіани, ієрархічні методи кластерного аналізу, самоорганізаційні карти Кохонена, методи крос-табличної візуалізації, різні методи візуалізації тощо.*

Друга група методів – це *прогнозуючі методи*, що використовують значення одних змінних для передбачення/прогнозування невідомих (пропущених) або майбутніх значень інших (цільових) змінних. До методів, спрямованих на отримання *прогнозуючих результатів*, належать такі, як: *нейронні мережі, дерева рішень, лінійна регресія, метод найближчого сусіда, метод опорних векторів та ін*.

Властивості методів Data Mining

Різні методи Data Mining характеризуються певними властивостями, які можуть бути визначальними у виборі методу аналізу даних. Методи можна порівнювати між собою, оцінюючи характеристики їхніх властивостей. Серед основних властивостей і характеристик методів Data Mining можна вказати такі: *точність, масштабованість, інтерпретованість, перевірність, трудомісткість, гнучкість, швидкість і популярність.*

Зокрема *масштабованість* – це властивість обчислювальної системи, що забезпечує передбачуване зростом системних характеристик, наприклад, швидкості реакції, загальної продуктивності тощо, при додаванні до неї обчислювальних ресурсів.

Більшість інструментів Data Mining, що пропонуються нині на ринку програмного забезпечення, реалізують одразу кілька методів, наприклад: дерева рішень, індукцію правил і візуалізацію, або ж нейронні мережі, карти Кохонена та візуалізацію.

В універсальних прикладних статистичних пакетах (наприклад, **SPSS, SAS, Statgrafics, Statistica** та ін.) реалізується широкий спектр найрізноманітніших методів (як статистичних, так і кібернетичних). Слід враховувати, що для можливості їх використання, а також для інтерпретації результатів роботи статистичних методів (кореляційного, регресійного, факторного, дисперсійного аналізу тощо) потрібні спеціальні знання в галузі статистики.

Універсальність того чи іншого інструменту часто накладає певні обмеження на його можливості. Перевагою використання таких універсальних пакетів є можливість відносно легко порівнювати результати побудованих моделей, отримані різними методами. Така можливість реалізована, наприклад, у пакеті Statistica, де порівняння засноване на так званій «конкурентній оцінці моделей». Ця оцінка полягає в застосуванні різних моделей до одного і того ж набору даних для порівняння їх характеристик та вибору найкращої з них.

Важливим аспектом у контексті впровадження технологій інтелектуального аналізу даних у діяльності органів ДФС є знання аналітичних можливостей платформи Oracle, на якій функціонує і розвивається інформаційно-аналітична система податкового відомства.

3.1.3. Технології Oracle Advanced Analytics

2012 року корпорація Oracle оголосила про випуск Oracle Advanced Analytics – нової опції для СУБД Oracle Database 11g, яка об'єднує середовища статистичних досліджень Oracle R Enterprise і Oracle Data Mining.

Використовуючи можливості Oracle Database, середовище Oracle R Enterprise пропонує продуктивність корпоративного класу для додатків, розроблених за допомогою мови статистичних досліджень і графічної візуалізації R, а також на кілька порядків збільшує обсяги даних, з якими можуть працювати ці додатки. Технологія Oracle Data Mining, що стала компонентом опції Oracle Advanced Analytics, дозволяє легко створювати і розгортати додатки для інтелектуального аналізу, які допомагають отримувати від інформаційних масивів цінні знання для бізнесу.

Oracle Advanced Analytics реалізує концепцію аналітики, вбудовану в базу даних, з переміщенням алгоритмів до даних, що зберігаються в Oracle Database і Oracle Exadata (традиційний підхід, заснований на переміщенні даних до алгоритмів, вимагає додаткових витрат на вилучення даних зі сховищ і їх завантаження в ноутбуки або спеціалізовані сервери). На основі Oracle Advanced Analytics можна створювати аналітичні додатки в режимі реального часу, які допомагають отримувати відповіді на ключові питання бізнесу.

Продукт Oracle Advanced Analytics разом з оптимізованими програмно-апаратними комплексами Oracle Big Data Appliance, Exadata Database Machine і Exalytics In Memory Machine надає найповнішу і інтегровану в галузі платформу для аналізу «великих даних».

Опція поглибленої аналітики **Oracle Advanced Analytics** доступна починаючи з версії 11gR2 Oracle Database Enterprise Edition і включає в себе два основних компоненти:

– **Oracle Data Mining** – опція, яка до Advanced Analytics Option була самостійною, її часто отримували разом з OLAP Option, яка як і раніше є окремою опцією;

– **Oracle R Enterprise** – корпорація Oracle інтегрувала мову статистичних досліджень R, що широко використовується в наукових колах і швидко замінює в комерційних організаціях мову Base SAS і SPSS.

3.1.4. Опція поглибленої аналітики Oracle Data Mining

Технологія Oracle Data Mining призначена для аналізу даних методами, що належать до класу вилучення знань (Data Mining). Основне завдання цієї технології полягає у виявленні у великих наборах даних прихованих закономірностей, залежностей і взаємозв'язків, корисних під час прийняття рішень на різних рівнях управління. Такі закономірності використовуються у вигляді моделей різного типу, що дозволяють проводити класифікацію ситуацій або об'єктів, прогнозувати їхню поведінку, виявляти групи подібних об'єктів тощо. Істотно, що моделі будуються автоматично на основі аналізу наявних даних про об'єкти, спостереженнях і ситуаціях за допомогою спеціальних алгоритмів.

Методи добування знань застосовуються в різних сферах для вирішення таких практичних завдань, як залучення нових клієнтів, виявлення найбільш перспективних груп і утримання клієнтів, підвищення ефективності маркетингової діяльності з просування продуктів і послуг, виявлення причин збоїв обладнання тощо.

Основу Data Mining становлять процедури, що реалізують різні алгоритми побудови моделей класифікації, регресії, кластеризації. Версія Data Mining 12c підтримує широкий спектр таких алгоритмів (рис 3.2).

Класифікація використовує історичні дані для побудови моделі, яка може бути використана для прогнозування належності нових даних до того чи іншого класу. Для вирішення завдань класифікації в Data Mining доступні алгоритми:

1) узагальнені лінійні моделі (GLM);

2) наївні класифікатори (Naive Bayes);

3) метод опорних векторів (SVM);

4) дерева прийняття рішень;

5) кластеризація – потрібна для дослідження даних і виявлення природних груп. Наприклад, побудови клієнтських сегментів;

6) покращений алгоритм k-means (на основі метрики відстані);

7) метод *Orthogonal Partitioning Clustering* (на основі щільності);

8) регресія – використовується для прогнозування недискретних величин:

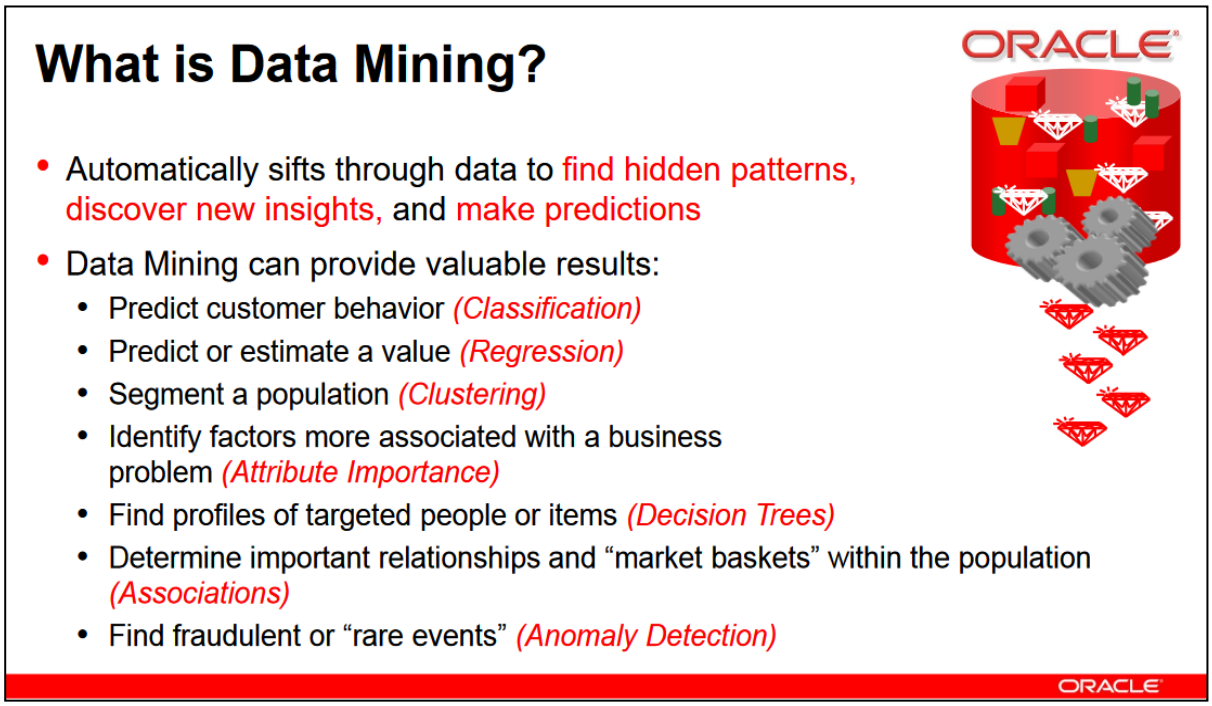


Рис. 3.2. Узагальнені групи алгоритмів, що підтримує Oracle Data Mining

9) визначення важливості атрибутів – пошук фактора, який впливає на позитивний відгук на пропозицію – вирішується за допомогою алгоритму Minimum Description Length;

10) виявлення аномалій за допомогою методу опорних векторів дозволяє побудувати профіль одного класу і потім, при застосуванні, відзначати випадки, що так чи інакше відрізняються від цього профілю (тобто «ненормальні» або «підозрілі»); це дає можливість виявляти рідкісні випадки, необов'язково пов'язані один з одним, виявити які з допомогою класифікації практично неможливо;

11) пошук асоціативних правил за допомогою алгоритму Apriori дозволяє виявляти події, що спільно зустрічаються. Наприклад, виявляти спільно придбані товари;

12) виділення ознак дозволяє створювати нові атрибути як лінійну комбінацію вже існуючих. Алгоритм застосовується для аналізу текстів, семантичного аналізу, розпізнавання образів.

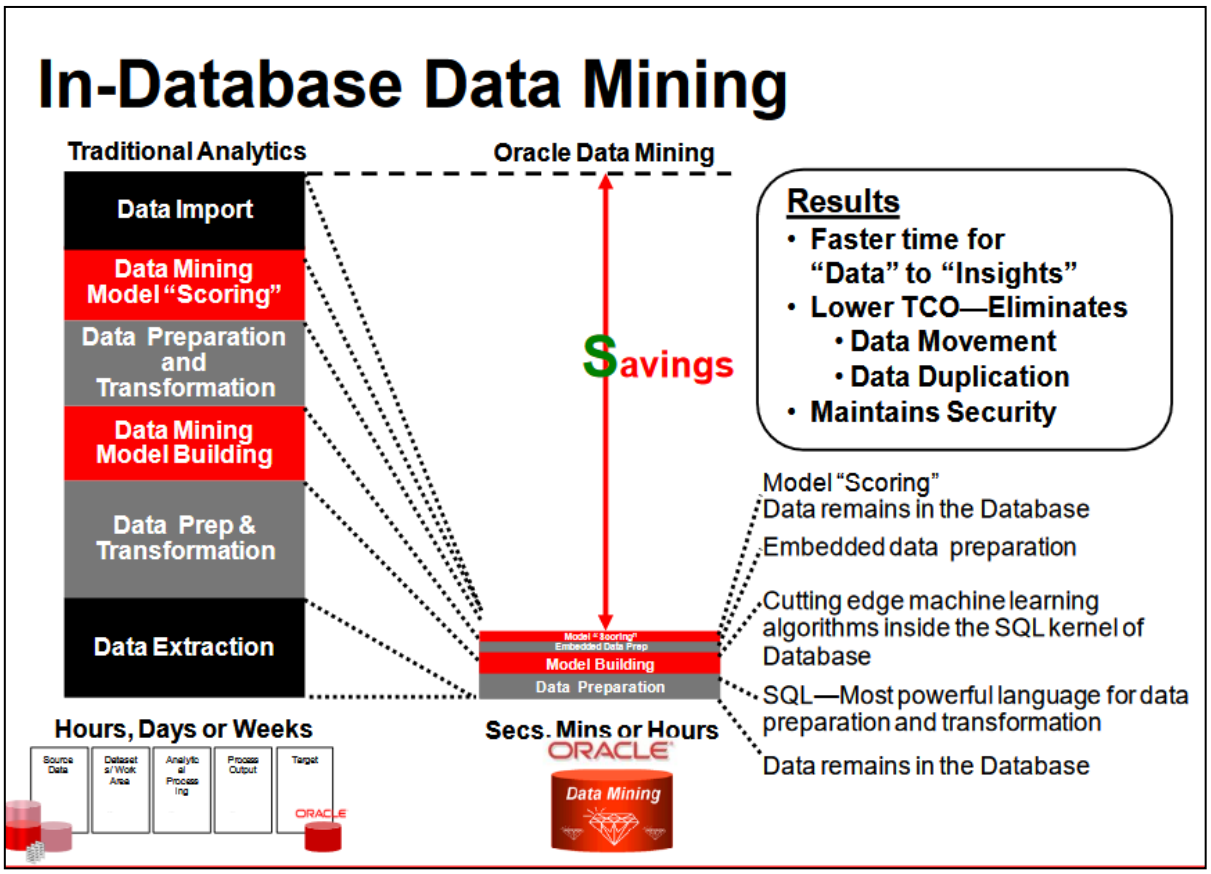


Рис. 3.3. Переваги Oracle Data Mining перед традиційним аналізом даних

Пояснення основних вузлів Oracle Data Mining (рис. 3.4.)

1.Tables and Views – даний вузол дозволяє створювати таблиці і уявлення та інші об'єкти бази Oracle Data Mining.

2.Transformations

Важлива особливість алгоритмів полягає в тому, що всі вони працюють безпосередньо в реляційній базі даних Oracle і не вимагають вивантаження і збереження даних у спеціальних форматах. Крім власне алгоритмів, Oracle пропонує засоби підготовки даних, оцінки результатів, застосування моделей до нових наборів даних **(transformations)**. Використовувати всі ці можливості

можна як на програмному рівні за допомогою Java API або PL/SQL API, так і за допомогою графічного середовища Oracle Data Miner, яке являє собою розширення для середовища SQL Developer і орієнтоване на роботу аналітиків, які вирішують завдання прогнозування, виявлення тенденцій, сегментації та ін. Набір алгоритмів класу Data Mining в компоненті Oracle Data Mining дозволяє вирішувати широкий спектр бізнес-завдань. Отримані за допомогою цих алгоритмів інтелектуальні моделі можна розгортати в Oracle Database 11g, використовуючи при цьому технології Oracle Exadata Smart Scan для швидкого скорингу величезних обсягів даних.

3. Explore Data – дослідження даних;
4. Modelling – застосування методів для моделювання;
5. Text – аналіз тексту.

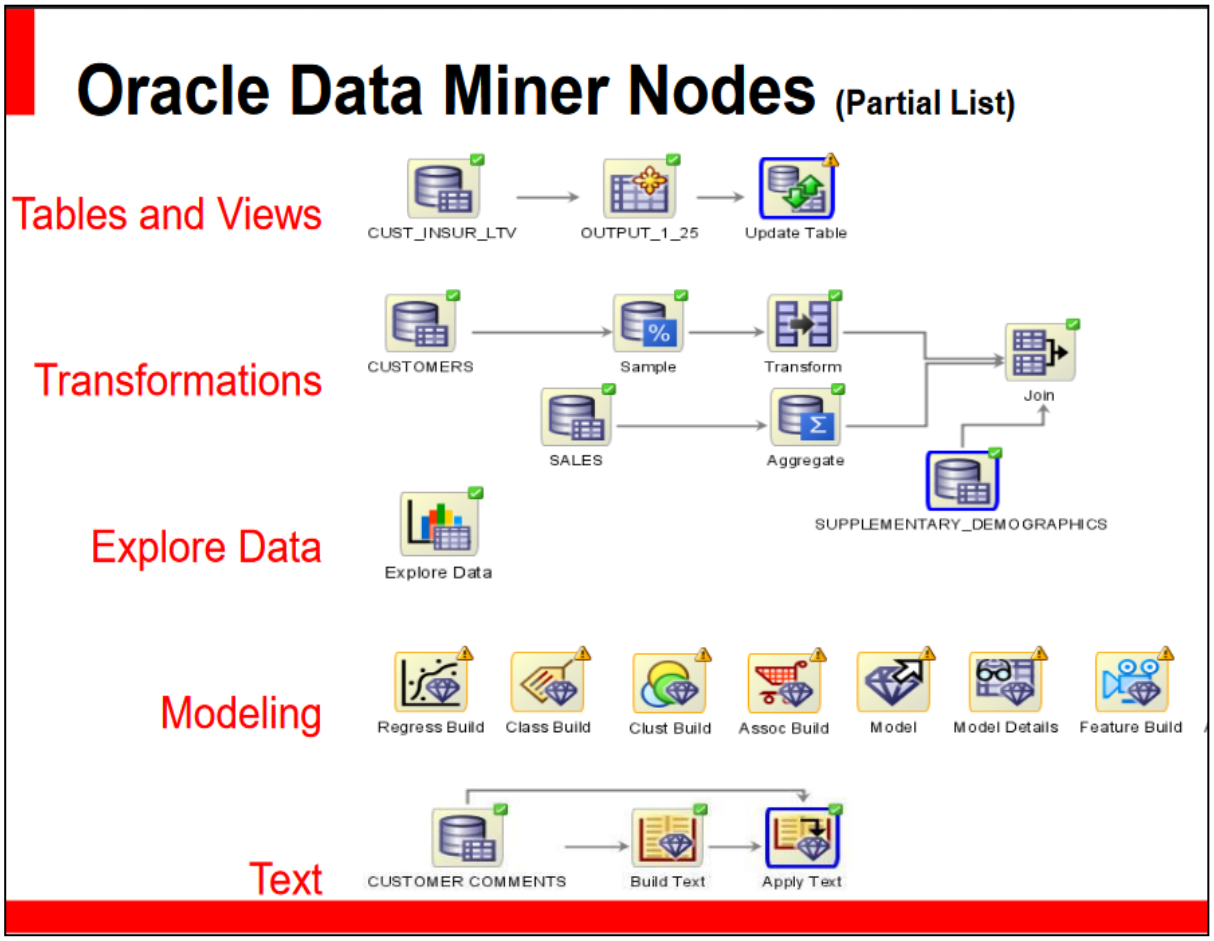


Рис. 3.4. Основні вузли Oracle Data Mining

3.1.5. Опція Oracle R Enterprise

Продукт Oracle R Enterprise дає аналітикам можливість застосовувати різноманітні складні обчислювальні методи до наборів даних у мільярди рядків за лічені секунди, роблячи реальним ітеративний, швидкий і якісний аналіз «великих даних». Він скорочує час розгортання моделей статистичного аналізу, позбавляючи від необхідності переводити моделі на інші мови програмування перед розгортанням і практичним застосуванням.

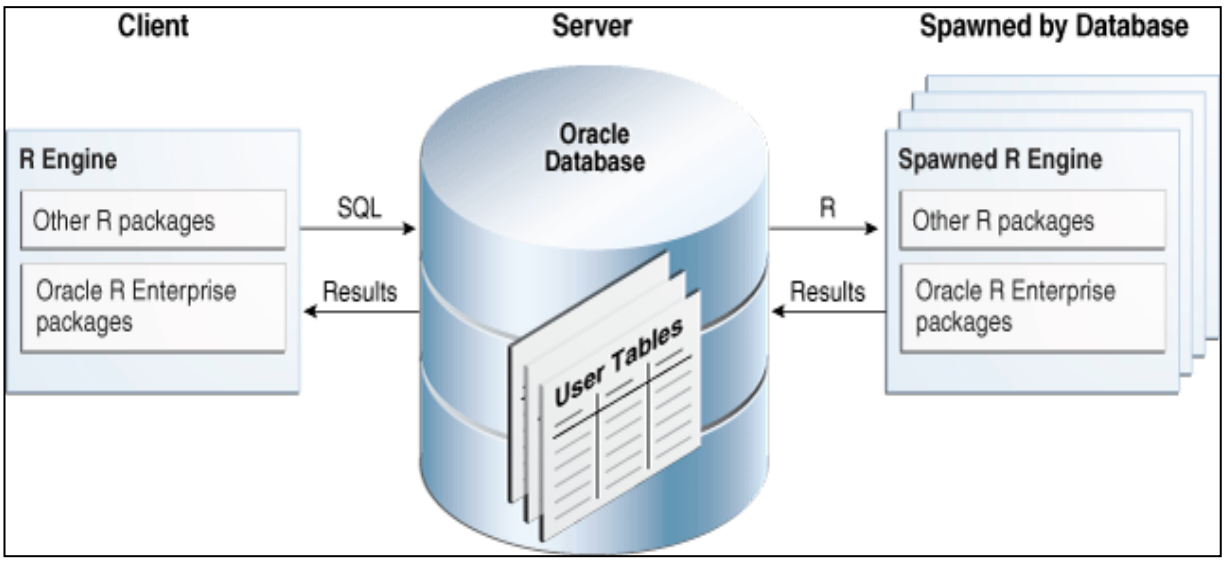


Рис. 3.5. Архітектура Oracle R Enterprise

Oracle R Enterprise інтегрує набір вбудованих в Oracle Database алгоритмів глибинного аналізу (data mining), аналітичних та статистичних функцій і можливість доступу до OLAP-кубів з мовою програмування R для «прозорого» застосування всіх цих технологій користувачами середовища R.

Пояснення до архітектури Oracle R Enterprise (рис. 3.5).

1. **Рушій клієнта R (Client) –** це набір пакетів R, що дозволяє підключатися до бази даних Oracle і взаємодіяти з даними у цій базі даних. Можна використовувати будь-які команди R від клієнта. Крім того, клієнт надає такі функції:

– рівень прозорості R SQL перехоплює функції R для масштабованого виконання в базі даних;

– функції перехоплюють перетворення даних, статистичні функції та функції Oracle R Enterprise;

– інтерактивне відображення графічних результатів і керування потоком як у відкритому джерелі R;

– подання R функцій для виконання в Oracle Database.

2. **Сервер (Server Oracla DataBase)** являє собою набір процедур і бібліотек PL/SQL, які доповнюють версію Oracle Database 12c Release 1 (12.1) можливостями, необхідними для підтримки клієнта Oracle R Enterprise. Рушій R також встановлюється на базі даних Oracle для підтримки вбудованого виконання R. Oracle Database порождає рушії R, які можуть забезпечити паралелізм даних. Механізм Oracle R Enterprise Database забезпечує таку функціональність:

– масштабується до великих наборів даних;

– доступ до таблиць, переглядів і зовнішніх таблиць у базі даних, а також доступних через посилання бази даних;

– паралельне виконання запитів SQL;

– використання статистичних функцій інтелектуального аналізу даних у базі даних.

3. **R механізми, створені БД Oracle (spawned by Oracle Database**), підтримують паралелізм, керований базою даних, і забезпечують виконання запланованого виконання сценаріїв R, тобто планування або ініціювання сценаріїв R, упакованих у запит PL / SQL або SQL. Oracle R Enterprise, забезпечує ефективний перехід до інших створених рушіїв. Вбудоване виконання R може використовуватися для емуляції програмування стилю MapReduce.

Глибока інтеграція між R, Oracle Database 11g і Hadoop дозволяє програмістам створювати один скрипт на мові R, який потім може виконуватися у трьох різних середовищах: у встановленому на ноутбуках вільно розповсюджуваному середовищі R; в середовищі Apache Hadoop з інтерфейсами Oracle Big Data Connectors і в СУБД Oracle Database 11g.

Результати застосування технологій і алгоритмів опції Oracle Advanced Analytics можна переглядати за допомогою інструментального середовища Oracle Business Intelligence Foundation Suite й оптимізованого програмно-апаратного комплексу Oracle Exalytics In Memory Machine, що дозволяє реалізувати можливості аналізу «великих даних» у масштабі всього підприємства.

Середовище Oracle R Enterprise (ORE), використовуючи можливості Oracle Database 12c, істотно підвищує продуктивність додатків, розроблених за допомогою некомерційної мови статистичних досліджень і графічної візуалізації R, а також на кілька порядків збільшує обсяги даних, з якими можуть працювати ці додатки.

Oracle R Enterprise дозволяє застосовувати статистичні та аналітичні моделі, розроблені на R, до даних, які зберігаються в базі даних Oracle, і використовувати можливості паралельного виконання і масштабованість бази даних. ORE робить таблиці та подання бази даних доступними для R-середовища так, якби вони були об'єктами R. R Enterprise, вміє транслювати команди маніпуляції даними R в SQL-оператори, виконувати їх у базі даних і повертати результат у звичне для аналітика середовище R.

Можна також використовувати вбудований в Oracle Database компонент R (R Embedded) і виконувати готові R-скрипти паралельно декількома екземплярами R Embedded на різних наборах даних, що дозволяє значно збільшити швидкість статистичної обробки даних, моделювання і прогнозування.

Ключовим принципом опції Advanced Analytics Option, що належать як до data mining (інтелектуальний аналіз даних), так і до R, є наближення обробки до даних. Замість того, щоб витягувати дані з бази, а також з файлів або інших джерел, а потім завантажувати їх у сервер статистичної обробки, наприклад, SAS, можна виконувати R-скрипти і алгоритми Data Mining безпосередньо в базі даних, що дозволяє легко застосовувати моделі і класифікувати дані в режимі реального часу, наприклад, у додатках call-центру або для реалізації частини ETL-процесу (рис. 3.5).

Насправді існують два R-пакети, які Oracle об'єднав разом; один з них – безкоштовний пакет, інший – це опція бази даних. Безкоштовний пакет є власне дистрибуцією Oracle відкрито розповсюджуваного R-пакета, того ж, що скачується з веб-сайту R Project's website, але доповненого спеціальними бібліотеками для оптимізації роботи R на апаратній платформі x86. Вільно розповсюджуваний R може також бути завантажений з веб-сайту

Oracle разом з ліцензованим Oracle R Enterprise і встановлений безпосередньо на Oracle Linux або іншу ОС Unix з використанням Oracle's Yum repository. Oracle R Enterprise є базовим R-пакетом, розширений для роботи всередині бази даних Oracle такими (ліцензованими) елементами:

– R-пакети для розширення стандартних пакетів, що входять до складу open-source R;

– бібліотеки бази даних для з'єднання з Oracle і запуску R-скриптів у базі даних;

– SQL-розширення для виклику R-функцій з SQL-запитів і PL/SQL-структур.

За допомогою цих елементів реалізовані чотири основні можливості Oracle Enterprise R:

– «Transparency» layer (рівень «прозорості»), який перехоплює стандартні R-функції і розширює їх, щоб розмістити певні R-функції і типи даних у базі даних Oracle;

– Statistics Engine (статистичний механізм), який надає набір статистичних функцій і процедур для найбільш поширених і часто використовуваних статистичних бібліотек з метою виконання цих бібліотек у базі даних Oracle;

– SQL-розширення, які дозволяють серверу бази даних виконувати R-скрипти, підтримують паралелізм, забезпечують SQL-доступ до функцій R і XML-результатами;

– Hadoop Connector (конектор до Hadoop), що забезпечує виконання R-скриптів на Oracle Hadoop кластері, файли якого можуть розміщуватися в будь-якій файловій системі HDFS, у базах даних Oracle або в локальних файлах.

Питання для самоконтролю до теми 3.1

1. Дайте визначення понять Data Mining, Інтелектуальний аналіз даних.

2. Дайте визначення поняття Online Analytical Processing, OLAP.

3. Опишіть зв'язок Data Mining з іншими сферами аналізу інформації

4. Опишіть основні задачі, які вирішують методи Data Mining.

5. Дайте визначення понять «метод», «методологія», «методика».

6. Назвіть дві основні задачі, за якими можна класифікувати методи Data Mining.

7. Опишіть концепцію використання алгоритмів і даних в Oracle Advanced Analytics.

8. Які два основні компоненти включає в себе Oracle Database Enterprise Edition?

9. Яке призначення опція поглибленої аналітики Oracle Data Mining?

10. Назвіть декілька алгоритмів класифікації, що інтегровані в Oracle Data Mining.

11. Назвіть основні вузли аналізу даних Oracle Data Mining.

12. Опишіть суть опції поглибленої аналітики Oracle R Enterprise.

13. Що собою представляє рушій клієнта R (Client)?

14. Що собою представляє Server Oracle Data Base?

15. Опишіть R механізми, створені БД Oracle (spawned by Oracle Database).

16. Яким засобом можна проглядати результати застосування технологій і алгоритмів опції Oracle Advanced Analytics?

17. Як можна використовувати вбудований компонент R (R Embedded)?

18. Назвіть декілька ліцензованих елементів розширення функціональності базового R-пакета для Oracle R Enterprise.

18. Яке функціональне призначення Hadoop Connector (Конектор до Hadoop)?

3.2. Регресійно-кореляційні моделі податкових процесів

3.2.1. Основні поняття регресійно-кереляційного аналізу

У Data Mining існує великий клас задач, де потрібно встановити залежність між ознаками (атрибутами, показниками), які описують досліджуваний процес або об'єкт предметної області. Для цього будуються різні моделі, в яких ці ознаки виступають як змінні. Якщо модель буде коректно відображати залежність між вхідними та вихідними змінними, то за допомогою такої моделі можна буде передбачати значення вихідної змінної за заданим значенням вхідних.

Регресійно-кореляційний аналіз – це побудова та аналіз математичної моделі у вигляді рівняння регресії (рівняння кореляційного зв'язку), що виражає залежність результативної ознаки від однієї або кількох ознак-факторів і дає оцінку міри щільності зв'язку.

Розглянемо приклад. Власник крамниці планує оптимізувати закупівлі і для цього збирає статистику, що відображає залежність обсягів продажів від ціни, яка встановлюється на продукти. Передбачається, що роздрібна ціна змінюється щомісяця залежно від ціни, яка встановлюється постачальником, на яку, в свою чергу, впливають сезонність, якість товару, ситуація на ринку тощо. Для роздрібного продавця оцінка того, скільки товару він зможе продати за певну ціну, являє собою великий інтерес. Якщо така оцінка буде отримана, то стане зрозуміло, скільки і яких продуктів буде потрібно закупити, наприклад, на місяць при певній цінової ситуації на ринку.

Результатом зібраних спостережень з'явилася залежність щомісячних продажів товару А від встановленої ціни (табл. 1).

Мета аналізу – оцінка очікуваних обсягів продажів товару А залежно від встановленої ціни. Побудуємо модель продажів товару А, де як вхідна змінна буде використовуватися ціна, а в

якості вихідної – обсяг продажів. На рис. 3.6. представлено діаграму розсіювання вихідних даних. Візуальний аналіз показує наявність зворотньої залежності між ціною і обсягами продажів, які зі збільшенням ціни продажу падають. І це не є несподіваним. Проте з практичного боку найбільший інтерес являє собою кількісний опис цієї залежності, а саме, якого падіння попиту слід очікувати при збільшенні ціни за одиницю.

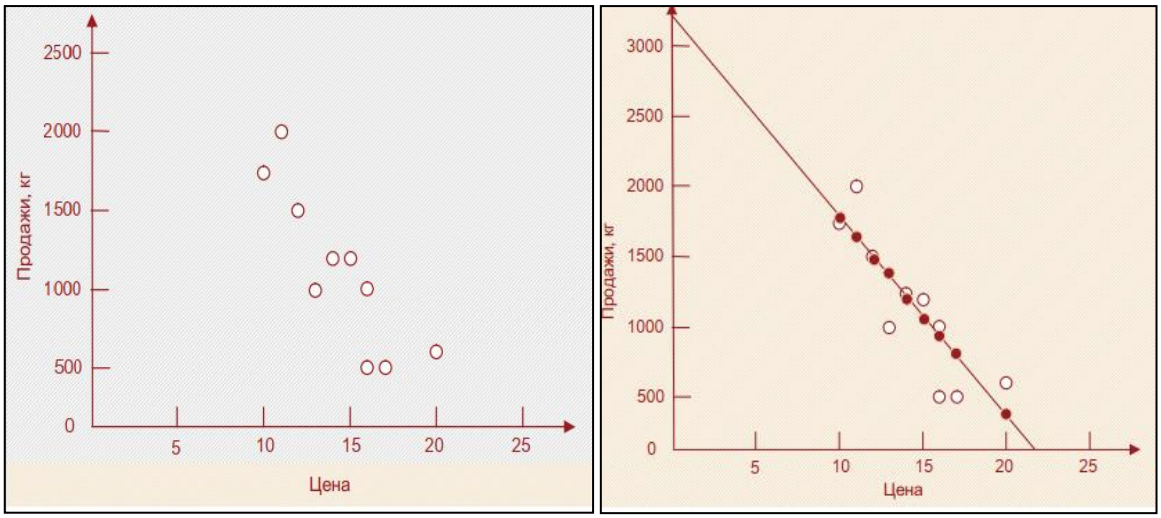


Рис. 3.6. Візуальний аналіз діаграми розсіювання і лінія регресії

Якщо хочемо змоделювати залежність обсягів продажів товару А від ціни за кілограм, то потрібно побудувати пряму, кожна точка якої буде являти собою оцінку обсягів продажів для заданої ціни. Але таких ліній можна побудувати нескінченно багато, і тільки одна з них забезпечить оптимальну оцінку обсягів продажів. Потрібно провести лінію таким чином, щоб розсіювання вздовж неї точок, що відповідають реальним спостереженням, було мінімальним. На практиці лінію будують так, щоб сума квадратів відхилень спостережуваних значень від оцінених за допомогою даної лінійної залежності була мінімальною, тобто:

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \to \min ,$$

де n – число спостережень;

$\hat{y}_i$ – оцінка вихідного значення і-го спостереження, отримана за допомогою моделі;

y_i – реально бачимо значення обсягу продажів.

Цей метод відомий як метод найменших квадратів (МНК), а лінія, побудована з його допомогою, називається *лінією регресії.*

Лінія регресії – це пряма найкращого наближення для набору пар значень вхідних і вихідної змінної (x, y), що обирається таким чином, щоб сума квадратів відстаней від точок (xi, yi) до цієї прямої, виміряних вертикально (тобто уздовж осі y), була мінімальна.

Рівняння, що описує лінію регресії, називається **рівнянням регресії**:

$$\hat{y}_i = b_0 + b_{1*}x,$$

де $\hat{y}_i$ – оцінка значення вихідної змінної; b_0 – точка перетину лінії з віссю y, звана також вільним членом. Це значення, яке приймає вихідна змінна y при x = 0. Коефіцієнт b_1 визначає нахил лінії щодо осі x (іноді його називають кутовим коефіцієнтом. *Тобто b_1 – це величина, на яку змінюється значення вихідної змінної y при зміні вхідної змінної x на одиницю.*

Коефіцієнти лінійного рівняння b_0 і b_1 називаються коефіцієнтами регресії. У зарубіжній літературі *регресію* часто називають *прогнозуванням* (prediction). При цьому вхідну змінну називають прогнозною змінною або предиктором *(Predictor variable),* а вихідну змінну, яка прогнозується, *(Predicted variable).*

Таким чином, завдання побудови моделі лінійної регресії зводиться до знаходження таких коефіцієнтів b_0 і b_1, для яких сума квадратів помилок, тобто різниць між реальними значеннями вихідної змінної y_i і їхніми оцінками була б мінімальна.

$$y_i = b_0 + b_{1*}x + \varepsilon$$

де ε – помилка

Використовуючи будь-який статистичний пакет, зокрема Excel, можна розрахувати коефіцієнти регресії. (В Excel, наприклад, потрібно активувати майстер функцій, в головному меню вибрати *Формули / Вставити функцію*/. Далі у вікні «Категорія» вибрати «Статистичні», у вікні «Функція» – ЛИНЕЙН. Ввести діапазон значень та необхідні параметри, як показано на рис. 3.7).

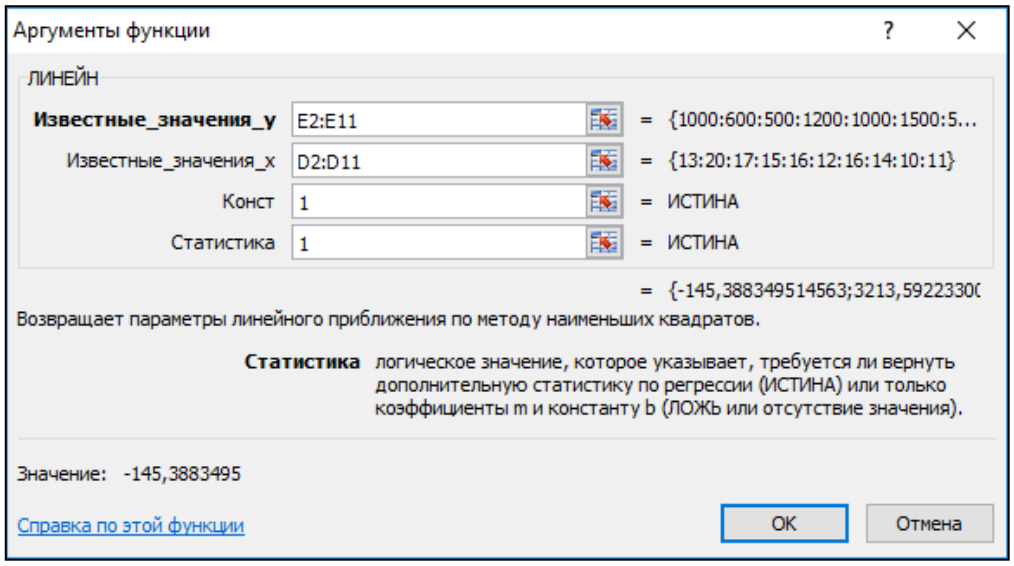


Рис. 3.7. Введення параметрів функції лінійної регресії

Відомі значення y – діапазон, що містить дані результативної ознаки (ціна товару А).

Відомі значення x – діапазон, що містить дані факторної ознаки (кількість проданого товару А);

Константа – логічне значення, яке вказує на наявність або відсутність вільного члена в рівнянні; якщо *Константа* = 1, то вільний член розраховується звичайним чином, якщо *Константа* = 0, то вільний член дорівнює 0.

Статистика – логічне значення, яке вказує, виводити додаткову інформацію з регресійного аналізу чи ні. Якщо *Статистика* = 1, то додаткова інформація виводиться, якщо *Статистика* = 0, то виводяться тільки оцінки параметрів рівняння.

Натискуємо «ок». У лівій верхній клітинці виділеної області з’явиться перший елемент підсумкової таблиці. Щоб розкрити всю таблицю, натисніть на клавішу <F2>, а потім на комбінацію клавіш <Ctrl> + <Shift> + <Enter>.

Додаткова регресійна статистика буде виводитися в порядку, зазначеному такою схемою:

Значення коефіцієнта b_1	Значення коефіцієнта b_0
Стандартна помилка b_1	Стандартна помилка b_0
Коефіцієнт детермінації R2	Стандартна помилка y
F-статистика	Число ступенів свободи df
Регресійна сума квадратів	Залишкова сума квадратів
$\sum(\hat{y}_x - \bar{y})^2$	$\sum(y - \hat{y}_x)^2$

Для нашого прикладу буде виведено інформацію:

-145,3883495	3213,592233	
30,02445334	440,8587726	
0,745613197	272,5453111	
23,4481722	8	
1741752,427	594247,5728	

Тобто рівняння регресії можна записати так:

$$\hat{y}_i = b_0 + b_{1*}x = 3213{,}6\text{-}145{,}4*x.$$

Суть коефіцієнтів рівняння регресії:

– b_0 – це значення вихідної змінної y при значенні вхідної змінної x = 0. Тобто у нашому випадку при ціні товару А, рівній 0, оцінка обсягів продажів становитиме 3213,6 кг. Однак ця формальна інтерпретація суперечить здоровому глузду, оскільки якщо роздавати товар безкоштовно, то куплять будь-яку його доступну кількість. Така ситуація виникла через те, що у вихідній вибірці спостережень відсутні значення x, близькі до нуля. Звідси випливає одне з обмежень лінійної регресії: лінію регресії (і, відповідно, що описує її рівняння) слід вважати придатною для апроксимації деякої реальної функції тільки в тому діапазоні змін вхідної змінної x, в якому розподілені вихідні спостереження. Інакше результати можуть виявитися непередбачуваними. Проте у деяких випадках надання цьому коефіцієнту економічного змісту має зміст, як буде показано у наступному параграфі.

– b_1 – значення коефіцієнта нахилу лінії регресії b_1 можна інтерпретувати як середню величину зміни значення вихідної змінної при зміні значення вхідної змінної на одиницю. У нашому прикладі це означає, що при збільшенні ціни товару А за один кілограм на одну грошову одиницю можна очікувати зменшення попиту в середньому на 145,4 кг. Лінію регресії для знайденого нами рівняння подано на рис. 3.6 вище.

3.2.2. Оцінка значимості регресійної моделі

Для багатьох практичних застосувань знайти рівняння регресії і побудувати відповідну лінійну залежність недостатньо. Щоб вирішити задачу, потрібно побудувати регресійну модель. Відмінність простої лінійної регресії від регресійної моделі полягає в тому, що рівняння регресії здебільшого будується на основі вибіркових даних. При цьому з множини спостережень, званих генеральною сукупністю, виділяється деяка вибірка, на основі властивостей якої передбачається робити висновки про властивості всієї сукупності. Для цієї вибірки і будується рівняння регресії. Потім проводиться спроба узагальнити отриману лінійну залежність на всю генеральну сукупність.

Тобто *проста лінійна регресія* будується на основі вибірки спостережень, кожне з яких містить відповідність x → y, а *регресійна модель* передбачає, що виявлена лінійна залежність має місце для всіх можливих пар x → y, містяться в генеральній сукупності. Однак при побудові регресійної моделі виникає ряд проблем, що виходять за рамки побудови рівняння регресії.

По-перше, потрібно визначити, чи справедливе припущення про те, що *лінійна залежність*, виявлена для вибіркових спостережень, *буде істинна* для всієї сукупності.

По-друге, ситуація ускладнюється тим, що регресія, власне кажучи, не є детермінованою завданням, оскільки для реальних даних залежність зміни вихідної змінної від зміни вхідної носить випадковий характер. Отже, коефіцієнти регресії можуть розглядатися як статистика, що описує розподіл вихідної змінної.

Проста лінійна регресійна модель задається таким чином. нехай є вибірка даних, що містить n парних спостережень, в кожному з яких значенням незалежної змінної x_i відповідає значення залежної змінної y_i, пов’язаних через лінійне відношення:

$$y = \beta_0 + \beta_1 + \varepsilon.$$

Для рівняння лінійної регресії наданого прикладу коефіцієнт детермінації $R^2 = 0{,}75$, що вказує на досить суттєвий зв’язок залежної і незалежної змінних. Але якщо для деякої вибірки визначили

коефіцієнт детермінації $R^2 = 0,2$, це значення показує, що відповідна модель навряд чи є значущою, тобто лінійна залежність між вхідною і вихідною змінними відсутня. Але чи відсутня вона повністю? Можливо, лінійна залежність між змінними має місце навіть при досить малих значеннях коефіцієнта детермінації. Чи існує системний підхід, що дозволяє визначити наявність лінійної залежності між вхідною і вихідною змінними? Методичну основу для оцінки значущості таких залежностей створюють *гіпотези лінійної регресії.*

У багатьох задачах інтерес становить не оцінка значення параметра моделі β_1 для генеральної сукупності, а деяке твердження про нього. У цьому випадку завдання зводиться до перевірки гіпотези, за результатами якої гіпотеза буде або підтверджена, або відхилена.

Вводять у розгляд дві гіпотези – *нульову (основну) і альтернативну*. Нульова гіпотеза H_0 передбачає, що лінійні зв'язки між змінними відсутні, водночас як альтернативна гіпотеза H_a стверджує, що такі зв'язки мають місце. Таким чином:

H_0: $\beta_1 = 0$ – лінійний зв'язок між змінними відсутній;

На: $\beta_1 \neq 0$ – лінійний зв'язок між змінними має місце.

Якщо правильна гіпотеза H_0, то лінія регресії буде розташована горизонтально і побудована регресійна модель марна. З таким же успіхом можна використовувати для оцінки просте середнє значення вихідної змінної, обчислене за всіма наявними спостереженнями. Навпаки, якщо вдасться відкинути нульову гіпотезу, підтвердиться значимість моделі, тобто наявність лінійної залежності між вхідною і вихідною змінними. Але щоб зробити остаточний висновок про придатність моделі до вирішення тієї чи іншої проблеми, *необхідно визначити ступінь її значимості*. Для цього можна використовувати *силу відхилення нульової гіпотези*. Для підтвердження або відхилення гіпотези використовується деякий статистичний показник, значення якого повинно бути обчислено на основі наявних спостережень.

Наприклад, якщо значення, прийняте показником, досить малоймовірне у припущенні істинності гіпотези, то істинність гіпотези також малоймовірна і її слід відхилити. При цьому впевненість у відхиленні гіпотези тим сильніша, чим менша ймовірність появи цього значення показника. Таке перевірочне значення часто називають *p*-значенням. Чим менше *p*-значення, тим вища очікувана значимість моделі і сильніша лінійна залежність між вхідною і вихідною змінними.

Для перевірки істинності нульової та альтернативної гіпотез існують різні критерії.

Регрессионная статистика						
Множествен	0,863488967					
R-квадрат	0,745613197					
Нормирован	0,713814846					
Стандартная	272,5453111					
Наблюдения	10					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	1	1741752	1741752,427	23,4481722	0,001284315	
Остаток	8	594247,6	74280,9466			
Итого	9	2336000				
	Коэффициенты	*артная о*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечен	3213,592233	440,8588	7,289391598	8,4753E-05	2196,97008	4230,214386
Переменная	-145,3883495	30,02445	-4,842331278	0,001284315	-214,6248631	-76,15183595

Рис. 3.8. Результат регресійного аналізу «Пакет анализа/ Регрессия»

Наприклад, гіпотеза H_0 може стверджувати, що вибірковий розподіл значень змінної є нормальним, а альтернативна гіпотеза – що він таким не є. Тоді ймовірність значення перевірочної статистики або підтвердить, що дані мають нормальний розподіл, або покаже ступінь його відхилення від нормального. Одними з найбільш популярних критеріїв оцінки значущості регресійних моделей *є t-критерій*, побудований на основі *t-розподілу Стьюдента*, і *F-критерій*, використовує *F-розподіл Фішера.*

Для розрахунків значення *t-статистика й* і розрахункового значення *F-критерію* скористаємося Excel «Пакет анализа/ Регрессия».

Із рисунка маємо t-статистика = -4,84, розрахункове значення F-критерію =23,5.

З таблиці t-розподілу можна побачити, що для n – 2 = 10 – 2 = 8 ступенів свободи значення 3,55 відповідає p = 0,005. Це вказує на ймовірність справедливості гіпотези, що дорівнює 0,5 %. Значення статистики t = –4,84, отримане на основі регресії, істотно менше ніж 3,55. Отже, для цього значення ймовірність справедливості гіпотези ще менше. Можна зробити висновок про високу ймовірність неспроможності гіпотези H_0, оскільки поява значення t = –4,8 дуже малоймовірна, якщо гіпотеза H_0 істинна. Тоді припущення про те, що $\beta_1 = 0$ і лінійна залежність між змінними відсутня, також не має сенсу. Перевіримо значущість моделі на основі *F критерію*.

У цьому випадку використовуються два показники – середній квадрат регресійної квадратичної суми S^2_R (у разі простої лінійної регресії m = 1) і середній квадрат помилок регресії S^2_E.

$$s_R^2 = \frac{Q_R}{m} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}, \qquad s_E^2 = \frac{Q_E}{n-m-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}. \quad (3.1)$$

На рис. 1 S^2_R *і* S^2_E знаходяться у розділі «Дисперсионный анализ» колонка MS. З їх допомогою можна обчислити значення, зване *F-статистикою*. Значення $F = S^2_R / S^2_E$ підпорядковується F-розподілу з (1, n - m – 1) ступенями свободи.

Якщо гіпотеза H_0: $\beta_1 = 0$ істинна і лінійний зв'язок між змінними відсутній, то помилка оцінювання зростає і її середній квадрат S^2_R теж. Тому якщо нульова гіпотеза справедлива, то значення F-статистики буде мале через збільшення знаменника S^2_E. Навпаки, якщо справедлива альтернативна гіпотеза H_a: $\beta_1 \neq 0$, то помилка оцінювання падає, а значення F зростає. Таким чином, більше значення F узгоджується з істинністю альтернативної гіпотези.

Для нашого прикладу, значення *F-статистика*, розраховане для даних про продажі товару А, буде F = 1741752/74281 = 23,5. По таблиці F-розподілу можна встановити, що для значень

ступенів свободи чисельника і знаменника, рівних відповідно 1 і n – m – 1 = 10 – 1 – 1 = 8 розраховане значення F = 23,5 набагато перевищує Fα = $F_{0,01}$ = 11,26. Таким чином, ймовірність того, що гіпотеза H_0: $\beta_1 = 0$ істинна для розрахованого значення F-статистики, буде менше 0,01. Тому відкидаємо нульову гіпотезу, що дозволяє зробити висновок про *високу значимість нашої регресійній моделі.*

3.2.3. Регресійно-кореляційна модель оцінювання тіньової економіки і податкових втрат

При оцінюванні ефективності та загальної якості адміністрування податків необхідно враховувати рівень тіньової економіки, відстеження її динаміки. Методика розрахунку тіньової економіки, затверджена наказом Міністерства економіки у 2009 р., передбачає обчислення групи індексів, які потребують великої кількості точних даних про фінансово-господарську діяльність економічних структур, які завідомо є не повною мірою достовірними й отримати які не завжди можливо. Тому доцільним є застосування статистичних та економетричних методів як допоміжних засобів для уточнення офіційних розрахунків, перевірки гіпотез та припущень щодо рівня тіньової економіки. Пропонується метод визначення усередненого рівня неоподатковуваного *валового регіонального продукту ВРП* та податкових втрат на основі оцінювання параметрів лінійної регресії між регіональними податковими надходженнями та ВРП у розрахунку на душу населення.

ВРП є узагальнюючим показником, що характеризує рівень розвитку економіки регіону. Його величина у ринкових цінах визначається як сума валової доданої вартості усіх видів економічної діяльності, включаючи чисті податки на продукти. Фактично показник характеризує все, що створено у регіоні протягом року.

Повне (абсолютне) оподаткування економіки регіону можна представити сумою нарахованого податкового зобов’язання на офіційну діяльність господарюючих суб’єктів (ГС) та податків і платежів, що не сплачуються ними через виведення діяльності у «тінь». Податкове зобов’язання регіону на основі даних ДФС України можна обчислювати як суму фактичних податкових надходжень, офіційного податкового боргу та наданих податкових пільг.

Для побудови регресійної моделі та її аналізу можна використати згадані показники у розрахунку на душу населення у регіонах України. Таким чином, формулу всіх можливих податків і платежів ($Тдн_i$) *i*-го регіону та податкового зобов'язання ($Тз_{дні}$) у розрахунку на душу населення можна визначати за формулами (3.2, 3.3):

$$Тдн_і = Тз_{дні} + Т_{Тдні}, \quad (3.2)$$

$$Тз_{дні} = Т_{Фдні} + Т_{Бдні} + Т_{Пдні}, \quad (3.3)$$

де $Тдн_i$ – вся сума можливих податків і платежів і-го регіону у розрахунку на душу населення (на д.н.);

$Тз_{дні}$ – податкове зобов'язання *i*-го регіону у розрахунку на д.н.;

$Т_{Тдні}$ – податки і платежі, що не сплачуються господарюючими суб'єктами через виведення діяльності у «тінь» у *i*-му регіоні у розрахунку на д.н.;

$Т_{Фдні}$ – фактичні податкові надходження і-го регіону у розрахунку на д.н.;

$Т_{Бдні}$ – податковий борг і-го регіону у розрахунку на д.н.;

$Т_{Пдні}$ – податкові пільги, отримані платниками податків і-го регіону у розрахунку на д.н.

Для спрощення для побудови моделі будемо використовувати формулу простої лінійної регресії – рівняння зв'язку двох змінних (3.4):

$$y_i = \alpha + \beta * x_i + \varepsilon_i, \; i = 1, n, \quad (3.4)$$

де x_i – значення незалежної змінної;

y_i – значення залежної змінної;

α, β, ε_i – параметри моделі та помилка (залишок) регресії.

Загальну формулу повного оподаткування ВРП на основі вищезазначених показників можна записати так (3.5):

$$Тдн_і = Тдн_{Фі} + Тдн_{Бі} + Тдн_{Пі} + Тдн_{Ті}, = Т_{ВРП} * ВРПдн_і + \varepsilon_і, \quad (3.5)$$

де $ВРП_{дні}$ – валовий регіональний продукт *i*-го регіону у розрахунку на д.н.;

$Т_{ВРП}$ – усереднена ефективна ставка оподаткування ВРП у державі.

Припустимо, що при сплаті всіх можливих податків і платежів, відсутності боргу і пільг модель не повинна мати постійного члена регресії α (при $ВРПдн_i = 0$ податки $Тдн_i=0$). Для розрахунків параметрів моделі на фактичних даних приймаємо, що податкові втрати $T_{Вдні}$ включають податковий борг (недоїмка, штрафи тощо), податкові пільги і податки не сплачені на тіньовий ВРП (3.6):

$$T_{Вдні} = T_{днТ} + T_{днБі} + T_{днПі}. \qquad (3.6)$$

Тоді загальні податкові втрати ($T_{Вдні}$) та номінальний (офіційний) ВРП, що не оподатковується ($ВРП_{Ндн}$) на душу населення, можна визначити на основі оцінених параметрів регресії (3.7):

$$T_{днФі} = T_{Вдн} + \beta * ВРПдн_i \,, \; ВРП_{Ндн} = - T_{Вдні} / \beta. \qquad (3.7)$$

На рис. 3.9 для прикладу наведено регресію (лінія Б) та оцінені параметри на основі формули (5) для даних 2008 року.

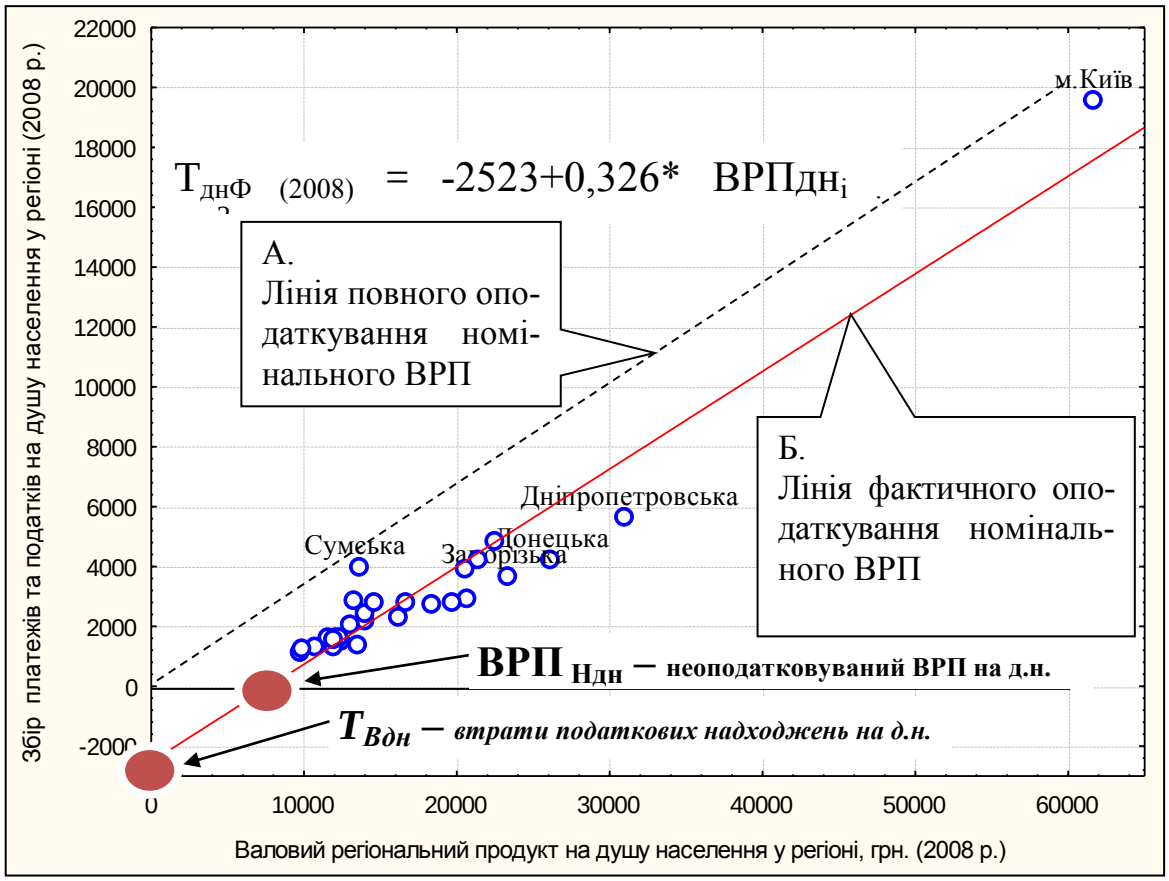


Рис. 3.9. Лінійна регресія регіональних податкових надходжень/дн та ВРП/дн у 2008 р.

Оцінки параметрів регресій показують, що економетричні моделі пояснюють від 70 % до 80 % дисперсії. Розраховані значення критерію Фішера для різних років дають можливість відкинути гіпотезу про рівність коефіцієнта β=0 з ризиком помилитися не більше ніж у 5 % випадків (Fрозр >Fтабл).

Результати вказують на наявність у 2008 році неоподаткованого регіонального валового продукту у розрахунку на д.н. у розмірі ~ 7800 грн. Це може означати, що у загальнодержавному вимірі у 2008 році неоподатковувалося (у тому числі знаходилося у тіні) ~ 7800 грн*46 млн нас. ~ 359 млрд грн. Відповідно, податкові втрати, включаючи податки і платежі, що не сплачуються господарюючими суб'єктами через виведення діяльності у «тінь», дорівнювали ~2523 грн д.н., що може означати недоотримання до бюджету 2523 грн *46 млн нас. ~ 116 млрд грн.

Відповідно, усереднену оцінку податкових втрат через наявність неоподатованого ВРП у регіоні можна здійснити на основі оцінки параметрів $Тдн_Т$ та β регресії (3.8):

$$Тз_{дні} = Тдн_{Фі}+Тдн_{Бі}+Тдн_{Пі},= – Тдн_Т + \beta * ВРПдн_і + \varepsilon_і. \quad (3.8)$$

Середнє значення неоподатковуваного ВРП у розрахунку на душу населення у державі визначається на основі формули (3.9)

$$ВРПндн = Тдн_Т / \beta. \quad (3.9)$$

На основі значень $Тдн_Т$ та *ВРПндн* можна розрахувати орієнтовні суми податкових втрат ВРП у регіонах та по державі загалом, обчислюючи добутки їх значень на кількість населення. Розрахунки на основі запропонованого методу наведені у таблиці 3.1.

На основі отриманих параметрів регресійної моделі розраховані орієнтовні суми податкових втрат і тіньового ВРП у регіонах та по державі загалом. Наприклад, для 2008 року отримані коефіцієнти моделі:

$$a = – 2523, b = 0{,}326. \quad (3.10)$$

Таблиця 3.1

Розрахунки економічних макропоказників податкової політики

№	Наближені результати розрахунків за методом / роки	2007	2008	2009
1.	ВРП*, млрд грн	721	948	915
2.	Податкові надходження (дані ДПС України), млрд грн	121	148	137
3.	Неоподатковуваний+ тіньовий ВРП у розрахунку на д.н., грн	5400	7800	>7315
4.	Загальний рівень неоподатковуваного, тіньового ВРП, млрд грн (п.3 * N*)	252	359	340
5.	Відсоток тіньового ВРП щодо офіційного, %	35	38	37
6.	Загальні податкові втрати у розрахунку на д.н., грн	1670	2523	2300
7.	Недоотримані податки, млрд грн (п.6 *N)	77	116	105
8.	% недоотриманих податків від суми податкових надходжень (п.7/п.2)	63	78	>77%

*N – кількість населення України за даними Державної служби статистики України.

У загальнодержавному вимірі у 2008 році база оподаткування, яка не оподатковувалася (у тому числі пільги і тіньова неоподатковувана діяльність), становила орієнтовно 359 млрд грн.

Відповідно, податкові втрати, включаючи податки і платежі, що не сплачуються господарюючими суб'єктами через пільги, заборгованість, виведення діяльності у «тінь», становили (у розрахунку на душу населення) ~2 523 грн. За розрахунками у 2008 році загалом недоотримано податків ~ 116 млрд грн.

Якщо порівняти результати розрахунків за цією моделлю, можна зробити невтішні висновки про негативні тенденції, що формувалися при здійсненні податкової політики впродовж 2008–2013 років. Приблизний неоподаткований ВРП по державі сягає близько 40 % від сумарного ВРП. Орієнтовані втрати податків через пільги, борг, тіньову економіку сягають близько 80 % податкових надходжень.

Цей метод дозволяє здійснювати наближену усереднену оцінку неоподаткованого, у тому числі тіньового ВРП, як на регіональному, так і на загальнодержавному рівні. Подальшого дослідження, уточнення та розкриття економічного змісту потребують розраховані значення показників, що наведені у табл. 3.1. Це важливо для розуміння процесів адміністрування податків.

3.2.4. Податкове навантаження як фактор ризику економічної безпеки держави

Теорія регресійно-кореляційного аналізу широко і достатньо повно висвітлена в науково-навчальній літературі. У цьому підрозділі буде розглянуто приклади застосування такого методу в аналізі податкової інформації.

З метою підвищення ефективності адміністрування податків, створення стабільної системи державних доходів податковим відомством упродовж 2001–2011 років здійснювався Проект модернізації операційної діяльності, реорганізація інформаційної та організаційної інфраструктури. Зазначений проект містив, зокрема, завдання щодо забезпечення умов для добровільного виконання вимог податкового законодавства платниками податків та забезпечення повного, своєчасного надходження законодавчо встановлених в Україні податків та зборів, впровадження в практичну діяльність ризик-орієнтованої системи та стандартних процедур перевірок до всіх платників податків.

Моніторинг проекту здійснювався на підставі контракту, укладеного між Державною податковою адміністрацією України та іноземним підприємством «ГФК ЮКРЕЙН» щодо проведення опитування платників податків у 2001–2010 роках.

У 2001 році ТОВ «Міжнародний центр розвитку підприємництва та менеджменту» на основі даних проведеного в ряді міст України опитування репрезентативної вибірки платників податків були розраховані базові ключові показники, на основі яких встановлені планові і контрольні значення, визначені Міжнародним

банком реконструкції та розвитку як обов'язкові до відстеження проекту. Зокрема такі показники:

– добровільність виконання платниками податків своїх податкових зобов'язань (далі – ДВПЗ);

– витрати з виконання вимог податкового законодавства платниками податків;

– ймовірність виявлення випадків ухилення від сплати податків та неплатежів;

– якість функціонування ДПС з адміністрування податків та надання послуг;

– показник ділової активності;

– доброчесність платників.

Вищезазначені показники можна розглядати як фактори податкових ризиків держави. Для дослідження був обраний показник, що відображає суб'єктивне сприйняття підприємцями рівня податкового навантаження і характеризує величину сукупної ставки оподаткування діяльності підприємців, а саме: показник «Добровільність виконання податкових зобов'язань», що розраховувався ГФК Юкрейн за результатами опитування платників податків. Значення показника обчислювалося як відсоток відповідей «оптимальний» на питання: «Візьміть до уваги всі податки і збори, які сплачує ваше підприємство. На вашу думку, яким є чинний рівень податкового навантаження вашого підприємства порівняно до оптимального (прийнятного для вашого підприємства)?». База для розрахунку – всі відповіді за винятком «важко сказати».

Як фактор ризику можна обчислювати обсяг податкових втрат (податковий розрив (tax gap) – як суму обсягів податкового боргу (нарахованих, узгоджених, але не сплачених податкових зобов'язань) та податкових втрат від тіньового сектору економіки (3.11):

$$ПВ = ПБ + ПВ_{ТЕ}, \qquad (3.11)$$

де ПВ – податкові втрати держави;

ПБ – податковий борг платників, що подають звіти;

$ПВ_{ТЕ}$ – податкові втрати держави від неоподатковуваного ВВП (тіньовий сектор економіки).

Податковий борг звітуючих платників розраховувався на основі даних, наведених за формулою (3.12):

$$ПБ = НП – СП, \quad (3.12)$$

де НП – сума нарахованих податків;

СП – сплачені податки та збори до державних цільових фондів.

Податкові втрати від тіньової економіки розраховувалися за формулою (3.13):

$$ПВ_{ТЕ} = ПТЕ * ВВП * ЕС, \quad (3.13)$$

де ПТЕ – показник тіньової економіки у % ВВП;

ВВП – обсяг валового внутрішнього продукту України;

$ЕС_{ВВП}$ – ефективна ставка оподаткування ВВП.

Дані, що використовувалися для регресійно-кореляційного аналізу у 2001–2010 рр., наведені у таблиці 3.2. Результати побудови кореляційних залежностей для 2001–2010 рр. показано на рис. 3.10–3.11.

Таблиця 3.2

Показники для регресійно-кореляційного аналізу економічної безпеки держави та податкових втрат у 2001–2010 рр.

Показник	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
ВВП, млрд грн	204	226	267	345	442	544	721	948	913	1083
$ЕС_{ВВП}$, %	0,18	0,20	0,20	0,18	0,22	0,23	0,22	0,24	0,23	0,22
ДВПЗ,%	54,5	59,5	66	69,2	74,7	86,8	88,9	81,6	90,6	91,8
ПТЕ, %	51,4	50,8	49,7	49	48	47	47	46	46	45
ІЕБУ	0,49	0,48	0,49	0,48	0,49	0,51	0,52	0,48	0,47	0,48
$ПВ_{ВВП}$, %	23,6	22,4	19	15,7	16,5	13,6	12,6	14,56	12,04	11,25

Аналізуючи кореляційну залежність питомої ваги податкових втрат у ВВП (далі – $ПВ_{ВВП}$) та показника добровільності виконання податкових зобов'язань (далі – ДВПЗ), можна побачити тісний, майже функціональний зв'язок (рис. 3.10). На рис. 3.11 наведено розраховані інструментом надбудови Excel «Пакет анализа\Регрессия» параметри отриманої залежності.

$$ПВ_{ВВП} = 39{,}5\text{-}0{,}31 * ДВПЗ,\ R^2 = 0{,}94. \qquad (3.14)$$

Коефіцієнт детермінації (R-квадрат) 0,94 означає, що варіація питомої ваги податкових втрат у ВВП на 94 % визначається варіацією показника «Добровільності виконання податкових зобов'язань» і лише 6 % – варіацією інших факторів. Зв'язок, і це цілком логічно, є обернено пропорційним.

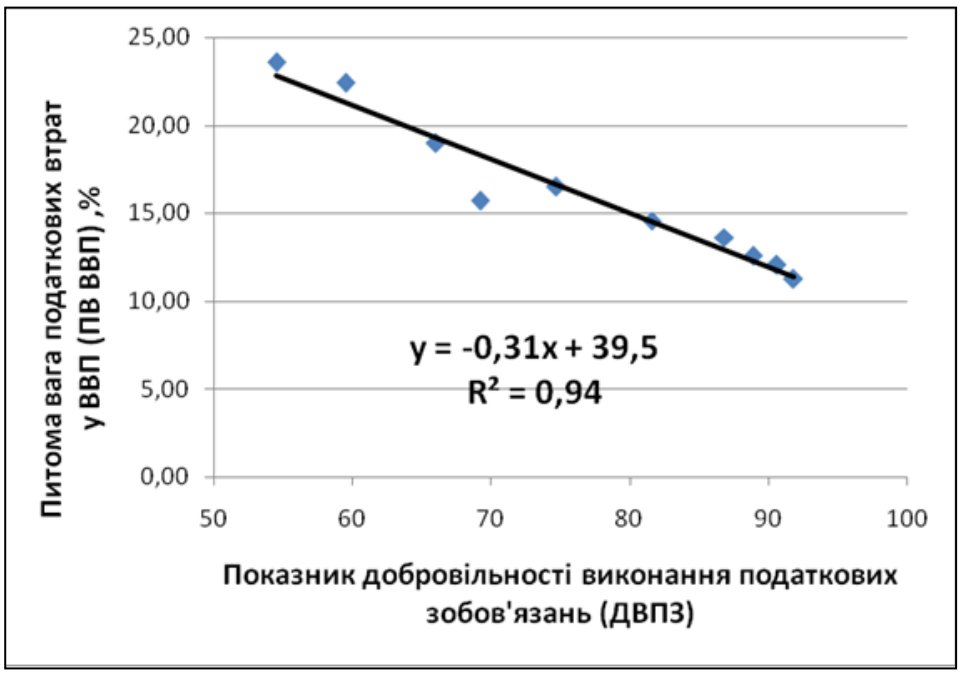


Рис. 3.10. Кореляція між показником добровільності виконання податкових зобов'язань і питомою вагою податкових втрат у ВВП у 2001–2010 рр.

Регрессионная статистика						
Множественный	0,971					
R-квадрат	0,942					
Нормированный	-1,25					
Стандартная оши	1,089					
Наблюдения	1					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	10	155,566	15,557	131,075	#ЧИСЛО!	
Остаток	8	9,495	1,187			
Итого	18	165,061				
	Коэф.	*Станд. ошибка*	*t-стат.*	*P-Значение*	*Нижн. 95%*	*Верхн. 95%*
b	39,51	2,072	19,07	5,92E-08	34,737	44,284
a	-0,306	0,0268	-11,45	3,06526E-06	-0,3687	-0,245

Рис. 3.11. Регресійна статистика та дисперсійний аналіз залежності (3.14), рис. 3.10

Коефіцієнт Фішера F = 131, що значно вище табличного (при даних параметрах f1=1 і f2=8 для рівня значущості p = 0.05 (5 %) табличне значення критерію Фішера дорівнює F= 5,32).

Нагадаємо, *критерій Фішера* для регресійної моделі відображає, наскільки добре ця модель пояснює загальну дисперсію залежною змінною. Розрахункове значення критерію Фішера можна визначити за рівнянням:

$$F = (R^2 / (1 - R^2) * (f2/f1), \qquad (3.15)$$

де R – коефіцієнт кореляції;

f1 і f2 – число ступенів свободи.

Перший дріб у рівнянні дорівнює відношенню поясненої дисперсії до непоясненної. Кожна із цих дисперсій ділиться на свій ступінь свободи (другий дріб у виразі). Число ступенів свободи поясненої дисперсії f1 дорівнює кількості пояснюючих змінних (наприклад, для лінійної моделі виду Y = A * X + B отримуємо f1 = 1). Число ступенів свободи непоясненної дисперсії f2 = N-k-1, де N – кількість експериментальних точок, k – кількість пояснюючих змінних.

Приклад: для лінійної моделі виду:

$$Y = A0 + A1 * X1 + A2 * X2,$$

побудованої по 20 експериментальних точках, отримуємо:

f1 = 2 (дві змінних X1 і X2), f2 = 20–2–1 = 17.

Для перевірки значимості рівняння регресії обчислене значення критерію Фішера порівнюють з табличним, узятим для числа ступенів свободи f1 (більша дисперсія) і f2 (менша дисперсія) на обраному рівні значущості (зазвичай p = 0,05). Якщо розрахований критерій Фішера вищим ніж табличний, то пояснена дисперсія істотно більша ніж можна було пояснити, і модель є значущою. Коефіцієнт кореляції і F-критерій, поряд з параметрами регресійній моделі, як правило, обчислюються в алгоритмах, що реалізують метод найменших квадратів.

Відповідно до оцінених коефіцієнтів лінійної залежності можна зробити узагальнені висновки щодо стану податкових втрат під час зростання показника ДВПЗ. А саме зростання показника ДВПЗ на 10 % приведе у середньому до зменшення питомої ваги податкових втрат у ВВП на 3,1%.

Аналізуючи кореляційну залежність відношення обсягів тіньової економіки до ВВП (далі – ТЕВ) та ДВЗ, також можна побачити тісний, майже функціональний зв'язок: коефіцієнт детермінації, який дорівнює 0,92 (рис. 3.12). Зв'язок обернено пропорційний: зростання показника ДВПЗ на 10 % може привести до зменшення рівня тіньової економіки ($ТЕ_{ВВП}$) у середньому на 1,5 %.

$$ПТЕ = 59,6–0,15*ДВПЗ, R^2 = 0,92. \quad (3.16)$$

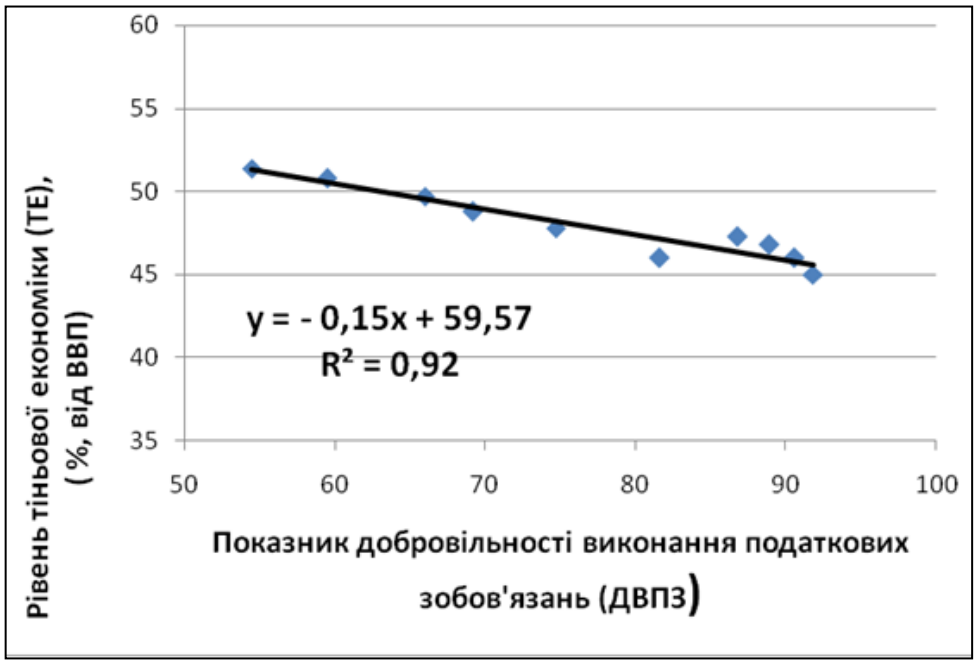


Рис. 3.12. Кореляція між показником добровільності виконання податкових зобов'язань і рівнем тіньової економіки України у 2001–2010 рр.

Для аналізу кореляційної залежності інтегрального індексу економічної безпеки України від обсягів тіньової економіки у 2007–2014 роках використовувалися дані Міністерства економічного розвитку і торгівлі України про стан економічної безпеки України станом на 2014 рік (табл. 3.3).

Регрессионная статистика						
Множественный R	0,961					
R^2	0,924					
Нормированный R^2	-1,25					
Станд. ошибка	0,634					
Наблюдения	1					
Дисперс. анализ						
	df	*SS*	*MS*	*F*	*ачимость F*	
Регрессия	10	39,08	3,908	97,322	#ЧИСЛО!	
Остаток	8	3,212	0,402			
Итого	18	42,289				
	Коэф.	*Станд. ошибка*	*t-стат.*	*P-Знач.*	*Ниж. 95%*	*Верх. 95%*
b	59,711	1,205	49,555	3,04E-11	56,933	62,49
a	-0,154	0,0156	-9,865	9,39E-06	-0,189	-0,118

Рис. 3.13. Регресійна статистика та дисперсійний аналіз залежності (3.16), рис. 3.12

Таблиця 3.3

Динаміка середньої ставки оподаткування по Україні (ССО), обсягів тіньової економіки та інтегрального індексу економічної безпеки України у 2007–2014 рр.

Показ-ник	2007	2008	2009	2010	2011	2012	2013	2014
$ЕС_{ВВП}$, %	22	24	23	22	26	25	23	23
ПТЕ, %	29	35	40	38	34	34	35	43
ІЕБУ	52	48	47	48	50	48	49	46

Кореляційна залежність (рис. 3.14) Інтегрального індексу економічної безпеки України (далі – ІЕБУ) та показника тіньової економіки (далі – ПТЕ) вказує на тісний лінійний зв'язок зазначених показників. Коефіцієнт детермінації 0,85 означає, що варіація ІЕБУ на 85 % визначається варіацією показника ПТЕ, і лише на 15 % – варіацією інших факторів. Зв'язок є обернено пропорційним: зростання показника тіньової економіки ПТЕ на 10 % погіршує інтегральний показник економічної безпеки держави (ІЕБУ) на 4 %.

$$ІПЕБ = 62,8 – 0,4* ПТЕ. \qquad (3.17)$$

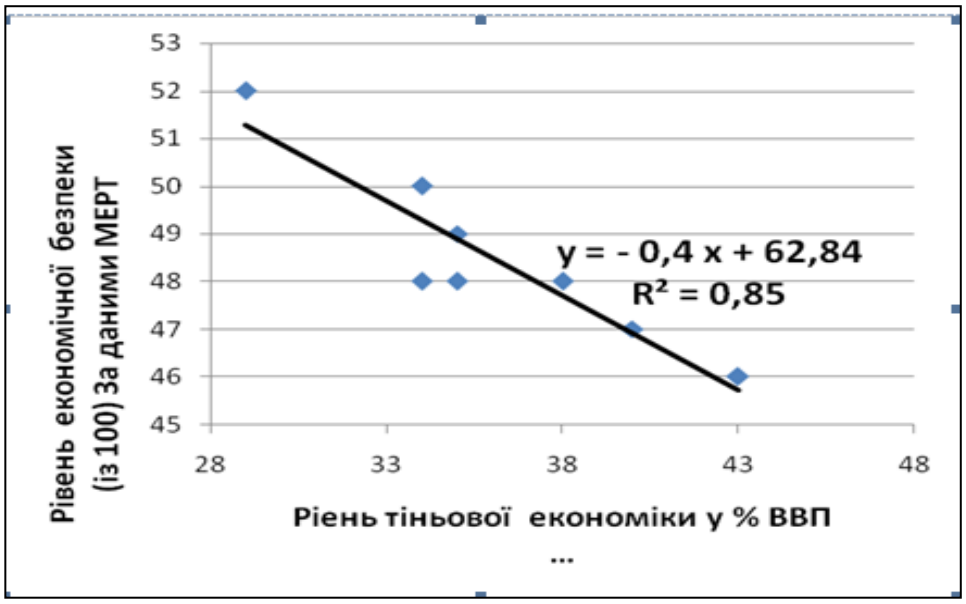


Рис. 3.14. Кореляція економічної безпеки (ІЕБУ) та рівня тіньової економіки (ПТЕ) у 2007–2014 рр.

Перевірка статистичних гіпотез вказує, що отримані кореляційні залежності не є випадковими і що існує тісний лінійний зв'язок між розглянутими показниками.

3.2.5. Прогнозування податкових надходжень на основі авторегресійних моделей динамічних рядів

ARIMA (англ. **Autoregressive integrated moving average**, іноді модель Бокса – Дженкінса, методологія Бокса – Дженкінса) – інтегрована модель авторегресії – змінного середнього – модель і методологія аналізу часових рядів. Є розширенням моделей ARMA для нестаціонарних часових рядів, які можна зробити стаціонарними взяттям різниць деякого порядку від вихідного часового ряду (так звані інтегровані або різні стаціонарні тимчасові ряди).

Податкові надходження мають чітко виражений сезонний тренд, що пов'язаний як з періодичністю надання суб'єктами господарювання звітності, так і сезонністю сплати податкових зобов'язань, коли у кінці кварталу, року мобілізуються основні платежі, проводиться реструктуризація боргів, здійснюються взаємозаліки. Під час прогнозування нестаціонарних економічних процесів

з чітко вираженою сезонністю рекомендується використовувати інтегровані авторегресійні моделі ковзного середнього.

Більш точні прогнози дає модель ARIMA із урахуванням сезонних компонент (ARIMAS).

Порядок моделі ARIMA задається трьома параметрами ARIMA (p, d, q):

1) (p) порядок авторегресії – AR nonseasonal autoregressive parameters (вибираємо значення – «0»);

2) (d) порядок оператора різниці (вибираємо значення – «0»);

3) (q) для приведення процесу до стаціонарного виду; порядок моделі ковзного середнього – MA onseasonal moving-average parameters (вибираємо значення – «0»).

Модель ARIMA (при $d = 0$) записується у вигляді:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + \delta + \varepsilon_t - K_1 \varepsilon_{t-1} - K_2 \varepsilon_{t-2} - ... - K_q \varepsilon_{t-q},$$

де y_t – величина досліджуваного процесу на час t $(t=1,2,...T)$;

$K_1, K_2 \ldots K_q$ – коефіцієнти моделі ковзного середнього;

$A_1, A_2, ..., A_P$ – авторегресійний коефіцієнт;

δ – параметр, що визначає середнє значення;

T – кількість спостережень;

ε_t – «білий» шум.

Таким чином, модель ARIMA прогнозує значення відповідного економічного показника на час t на підставі значень, які спостерігалися у попередні періоди. Кількість значень часового ряду визначається порядком авторегресії, проте не повинна бути менше 50.

Використовуються також похибки, одержані у процесі прогнозування попередніх значень. Кількість цих похибок визначається порядком моделі ковзного середнього. Авторегресійна складова і модель ковзного середнього утворюють модель ARIMA, яку можна використати для прогнозування стаціонарного випадкового процесу, тобто процесу, де математичне сподівання, дисперсія та моменти вищого порядку не залежать від часу. Оскільки у більшості часових рядів економічних показників випадки існування стаціонарних випадкових процесів зустрічаються надзвичайно рідко,

для приведення процесу до стаціонарного вигляду використовується оператор різниці, що перетворює модель ARMA в модель ARIMA.

Особливістю статистичних характеристик рядів податкових надходжень є відсутність значущого зв'язку між надходженнями в сусідні місяці. Авторегресійна функція надходжень за окремими типами податків, полягає у тому, що зв'язок між суміжними значеннями порівняно незначний (дискретність рядів дорівнює 1 місяць). У більшості випадків автокореляційна функція має екстремум при значенні величини лагу у 12 місяців. Це вказує на існування сезонності, період якої дорівнює 1 рік (12 місяців). Так, при наявності надходжень за січень попередніх років точніше можна визначити обсяг надходження на січень прогнозованого року. Цю особливість процесу добре описує модель *ARIMAS (P,D,Q),* де параметри P, D, Q мають зміст, аналогічний p, d, q із зрушенням на 12 кроків:

1) (P) сезонний параметр авторегресії – SAR seasonal autoregressive parameters (вибираємо значення – «3», квартальні дані);

2) (D) порядок оператора різниці (вибираємо значення – «1»);

3) (Q) сезонний параметр ковзної середньої – SMA seasonal moving-average parameters (вибираємо значення – «0»).

Для ідентифікації моделі *ARIMA*ARIMAS* проводиться розрахунок і візуальний аналіз автокорелограм, на основі якого робиться висновок про характер тренду, наявність сезонності, порядок авторегресійного і ковзного середнього, вибираються параметри моделі (p, d, q). Вибір моделі може також здійснюватися методом зменшення похибок, отриманих згідно з моделлю на базисному інтервалі.

Під час використання моделі ARIMA розраховуються такі похибки:

MSE – середній квадрат похибки:

$$MSE = \sum_{i=1}^{T} (x(i) - x(i))^2 / T,$$

де $x(i)$ – фактичне, $\hat{x}(i)$ – прогнозне значення;

MAE – середня абсолютна похибка:

$$MAE = \sum_{i=1}^{T} \left| \; x(i) - x(i) \; \right| / T \; ;$$

MPE – середня відсоткова похибка:

$$MPE = \sum_{i=1}^{T} \frac{(x(i) - x(i))}{x(i)} * 100\% / T \; ;$$

MAPE – середня абсолютна відсоткова похибка:

$$MAPE = \sum_{i=1}^{T} \left| \frac{x(i) - x(i)}{x(i)} \right| * 100\% / T \; .$$

Прогнозування передбачає наукове, засноване на системі факторів і доказів, встановлення причинно-наслідкових зв'язків, виявлення можливих шляхів і результатів майбутнього розвитку явищ і процесів, оцінку показників, що характеризують ці явища і процеси для більш-менш віддаленого майбутнього.

Оскільки часові ряди податкових надходжень характеризуються вираженою сезонністю, рекомендується використовувати інтегровані авторегресійні моделі ковзного середнього (ARIMA). Модель ARIMA прогнозує значення відповідного економічного показника на час *t* на підставі значень, які спостерігалися у попередні періоди. Побудова моделі здійснювалася за даними за 2008–2011 роки, точність прогнозу перевірялася за даними ДФС України про сплату податків за 2012 рік. Якщо за даними ДФС у 2012 році звітуючими платниками було сплачено 289319034 грн, то модель із встановленими межами довіри у 95 % дала дані про сплату податків у сумі 277421444 грн ± 5 %, що показує високу точність у межах 95 %. Експериментальним шляхом були підібрані такі параметри моделі: (p) порядок авторегресії =1; (d) порядок оператора різниці =1; (q) порядок моделі ковзкового середнього =0 (рис. 3.15).

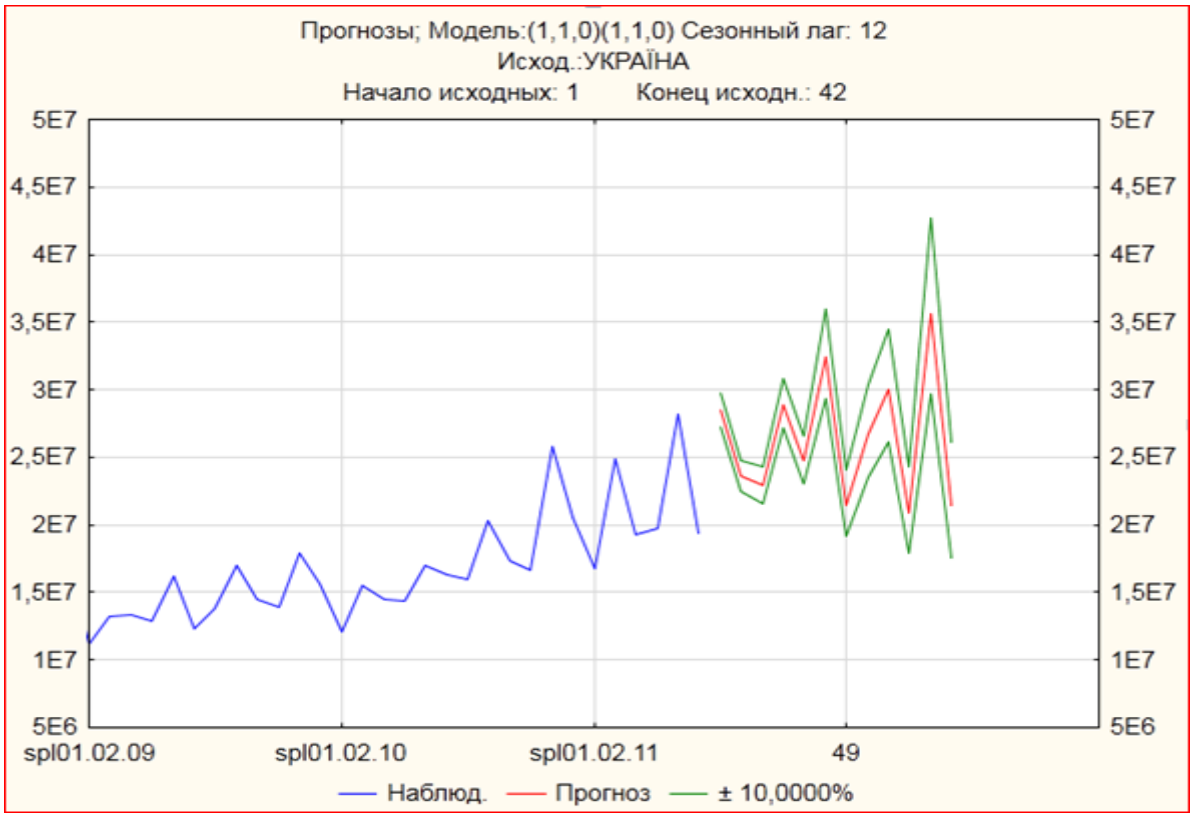


Рис. 3.15. Результати застосування моделі ARIMA для прогнозування сплати податків звітуючими платниками у 2012 році

Економетрична модель пояснює від 70 до 80 % дисперсії. Розраховані значення критерію Фішера для різних років дають можливість відкинути гіпотезу про рівність коефіцієнта β=0 з ризиком помилитися не більше ніж у 5 % випадків (Fемп>Fкр).

Існуючою на сьогодні проблемою для застосування описаних методів є відсутність в інформаційних масивах ДПС України інформації Державної служби статистики про соціально-економічний розвиток регіонів. Результати застосування прогнозних та регресійних моделей Statistica та Oracle Data Miner на базі даних про сплату податків та величину регіонального ВРП показують широкі перспективи для інтеграції даних інструментів у інформаційно-аналітичній системі. Розширення аналітичних можливостей інформаційної системи України сприятиме формуванню своєчасних, аргументованих управлінських рішень, виконання прогнозних показників податкових надходжень і своєчасного реагування фіскальними органами на негативні тенденції.

3.2.6.Нелінійна оптимізація у визначенні точок Лаффера

Результати кластерного аналізу показують наявність груп регіонів із відмінною динамікою соціально-економічного розвитку. Враховуючи цей факт, доцільним є оцінювання ефективності здійснення податкової політики у кожній із груп окремо. Класифікація груп регіонів за ознаками податкового зобов'язання на душу населення та ставки оподаткування ВРП дасть можливість перевірити гіпотезу про наявність ефекту Лаффера й оцінити ефективність податкової системи у кожному із регіонів.

Використовуючи апріорно відому нелінійну залежність між сумарною ставкою оподаткування ВРП та податковими надходженнями (ефект Лаффера), актуальною є безпосередня оцінка нелінійної двопараметричної формули на відповідність фактичним податковим даним (розділ 1).

Пошук параметрів нелінійної двопараметричної моделі можна оцінити за допомогою методів нелінійної оптимізації, що дозволяють отримувати екстремуми випуклих функцій. На сьогодні одним із найефективніших методів пошуку оптимуму випуклих функцій є група *Квазі-Ньютонівських методів* (*Quasi Newton methods*), основу яких складає узагальнена формула Ньютона, що використовується в ітераційних методах пошуку оптимуму нелінійних функцій:

$$x^{k+1} = x^k + \lambda_k p_i, \qquad (3.18)$$

де x^{k+1} – поточне наближення до рішення ;

k – кількість кроків ітерації, $k = 1, \ldots, k$;

λ_k – параметр, що характеризує величину кроку під час обчислення;

p_i – напрям пошуку оптимуму.

Напрям пошуку оптимуму функції визначається за формулою:

$$p_i = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k), \qquad (3.19)$$

де $\nabla f(x^k)$ – вектор перших частинних похідних функції $f(x^k)$, градієнт, або так званий Якобіан;

$\nabla^2 f(x^k)$ – матриця других частинних похідних функції або так званий Гессіан за іменем математика Гессе, що ввів її у використання.

Принципи заміни матриці Гессе на матрицю перших частинних похідних саме і покладені в основу методів, що отримали назву *Квазі-Ньютонівських (Quasi Newton Methods). Квазі-Ньютонівські методи* реалізовані в багатьох програмних продуктах для реалізації навчання нейромереж, нелінійного оцінювання, рішення системи нелінійних рівнянь.

Зокрема, у модулі «Нелінійне оцінювання» *(Nonlinear Estimation)* пакета *Statistica 10*) використовуються методи пошуку екстремуму функції Гауса Ньютона та Левенберга-Марквардта, які показали фактично однакову точність.

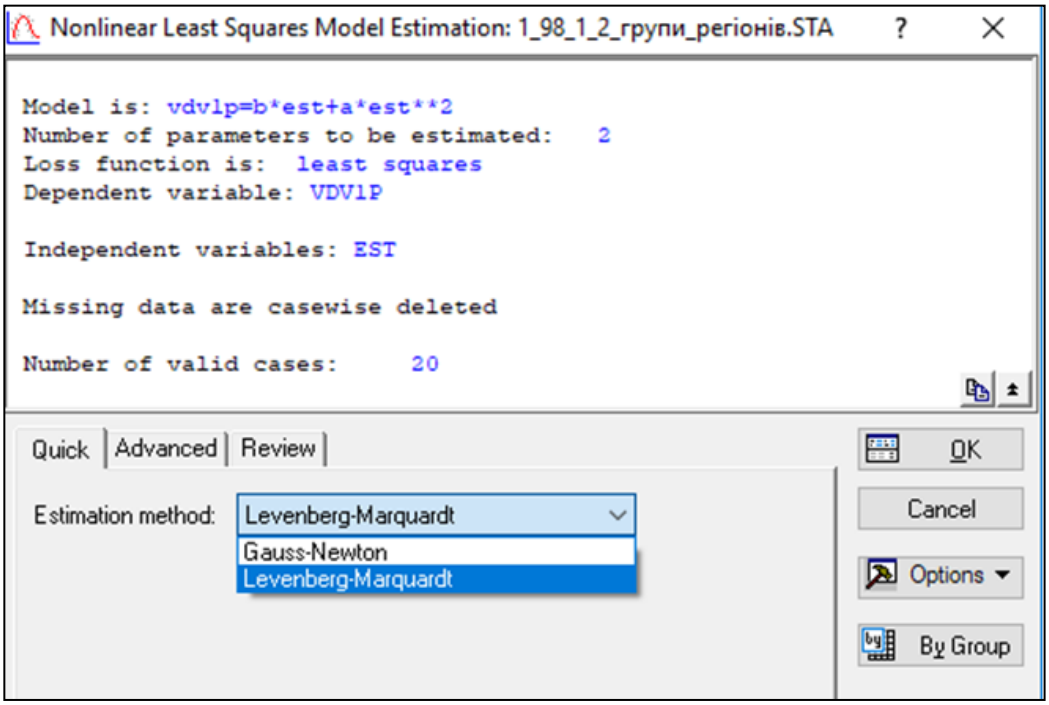


Рис. 3.16. Вікно вибору методу нелінійного оцінювання в *Statistica 10*

Модуль *нелінійного оцінювання* пакета *Statistica 10* дає можливість дослідити як залежність між експериментальними даними відповідає заданій нелінійній функції, визначити її параметри і процент поясненої дисперсії.

На рис. 3.17 для прикладу показаний вигляд виробничої та фіскальної кривих для групи регіонів у 2009 році, отриманий за допомогою пакета *Statistica 10.* Ідеальним для господарюючих суб'єктів був би варіант, коли $T_{ВДВij} < T_j^*$, або $T_{ВДВij}/T_j^*<1$, тобто коли держава стимулює зростання виробництва у регіоні, що, у

свою чергу, повинно приводити до зростання податкового зобов'язання регіонів за рахунок зростання податку на прибуток і зміщення з часом точки T_j^* вправо.

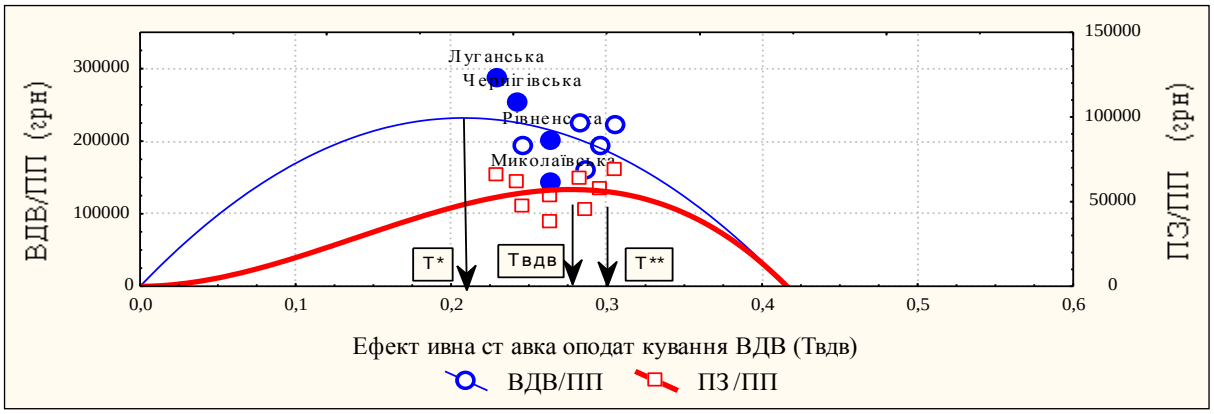


Рис. 3.17. Виробнича та фіскальна функції для групи регіонів у 2009 році

Одним із основних висновків аналізу точок Лаффера за 2009 рік є те, що для більшості річних регіональних даних $T_{ij}^* < Твдв_j < T_{ij}^{**}$, тобто податкове навантаження не стимулювало зростання виробництва.

Для оцінювання ефективності здійснення державою податкової політики у регіонах доцільно визначити позиціонування точки $T_{ВДВij}$ відносно точок T_j^*, T_j^{**} та допустимі межі її зміщення.

Рівень фіскальності податкової політики, здійснюваної у регіоні можна оцінювати на основі розрахунку *індексу фіскальності податкової політики* $I_ф$ (3.20), який характеризує позицію сукупної ставки оподаткування ВДВ у регіоні щодо точок T_j^*, T_j^{**}:

$$I_{фj} = 2T_{ВДВij} / (T_j^* + T_j^{**}). \qquad (3.20)$$

Отримання навичок використання методів нелінійної оптимізації для оцінювання параметрів моделі Лаффера в середовищі Statistica передбачено у лабораторному практикумі 3. Лабораторна робота 3.3.

Питання для самоконтролю до теми 3.2

1. Дайте визначення регресійно-кореляційного аналізу.

2. Охарактеризуйте метод найменших квадратів: математична суть, задачі, результат.

3. Як можна охарактеризувати лінію регресії?

4. Охарактеризуйте коефіцієнти рівняння регресії $\hat{y}i = b0 + b1*x$.

5. Вкажіть у рівнянні регресії $\hat{y}i = b0 + b1*x$ коефіцієнт, на яку змінюється значення вихідної змінної у при зміні вхідної змінної х на одиницю.

6. Як обчислюються коефіцієнти лінійного рівняння b0 і b1?

7. Які існують назви вхідних змінних та вихідних змінних?

8. Як змінюється лінія регресії при наявності або відсутності вільного члена в рівнянні (константи)?

9. Назвіть декілька показників додаткової регресійної статистики, яка буде виводитися під час використання функції Excel «ЛИНЕЙН».

10. Що характеризує коефіцієнт детермінації?

11. У чому відмінність побудови лінії регресії і побудови регресійної моделі?

12. Для чого проводиться оцінка значимості регресійної моделі?

13. Опишіть суть двох гіпотез, які перевіряються при здійсненні регресійного аналізу – нульова (основна) і альтернативна.

14. Яке перевірочне значення використовують для оцінювання істинності нульової гіпотези?

15. Які критерії використовують для перевірки істинності нульової та альтернативної гіпотез?

16. Якщо гіпотеза H0: $\beta 1 = 0$ істинна, то помилка оцінювання зростає чи зменшується?

17. Отримана наступна регресія Тдні = -2523+0,326* ВРПдні. Яке значення валового регіонального продукту на душу населення у регіоні (ВРПдн), якщо податкові надходження на душу населення у регіоні Тдн= 5120 грн?

18. Де можна застосовувати і які оцінки в оподаткуванні дозволяє робити метод регресійно-кореляційного аналізу?

19. Охарактеризуйте модель податкового навантаження як фактора ризику економічної безпеки держави.

20. Які показники використовувались Міжнародним банком реконструкції та розвитку, як обов'язкові до відстеження проекту модернізації ДПС України?

21. Охарактеризуйте показник, що відображає суб'єктивне сприйняття підприємцями рівня податкового навантаження, який використовувався «ГФК ЮКРЕЙН» для відстеження проекту модернізації ДПС України?

22. Запишіть формулу податкових втрат на основі податкового боргу та втрати держави від неоподатковуваного ВВП.

23. Якщо ПТЕ – показник тіньової економіки у % ВВП, ВВП – обсяг валового внутрішнього продукту України, ЕСВВП – ефективна ставка оподаткування ВВП, як можна записати формулу податкових втрат від тіньової економіки?

3.3. Побудова класифікаційних моделей на базі технологій Data Miming

3.3.1. Нейронні мережі у задачах формування і виявлення образів суб'єктів оподаткування

Історія нейронних мереж

З появою сучасної електроніки почалися перші спроби апаратного відтворення процесу мислення людини. Перший крок був зроблений у 1943 році з виходом статті нейрофізіолога Уоррена Маккалоха та математика Уолтера Піттса про роботу штучних нейронів і представлення моделі нейронної мережі на електричних схемах. Паралельно з прогресом у нейроанатомії і нейрофізіології психологами було створено моделі людського мислення, навчання. У 1950–1960-ті роки група дослідників, об'єднавши ці біологічні та психофізіологічні підходи, створила перші штучні нейронні мережі за керівництва Натаніеля Рочестера.

З 1980-х років інтерес до штучних нейронів знову починає зростати у зв'язку із зацікавленістю фахівців різних галузей науки можливостями технологій машинного навчання щодо застосування їх у своїх дослідженнях.

Наступним поштовхом у розвитку технологій нейромереж став 2007 рік, коли Джефрі Хінтон в університеті Торонто розробив алгоритми глибокого навчання багатошарових нейронних мереж. Це ознаменувало активізацію застосування штучних нейронних мереж у різних сферах людської діяльності та дало новий поштовх розробкам технології штучного інтелекту, зокрема для актуальних задач в напрямі аналізу Big Data.

Біологічний нейрон – це нервова клітина, яка є основним елементом нервової системи живих організмів. Саме нейрони, взаємодіючи між собою, забезпечують процеси пошуку, передачі та обробки інформації в нервовій системі. Біологічний нейрон складається з *тіла клітини* або *соми* – оболонки, що містить речовини, необхідні для забезпечення життєдіяльності клітини.

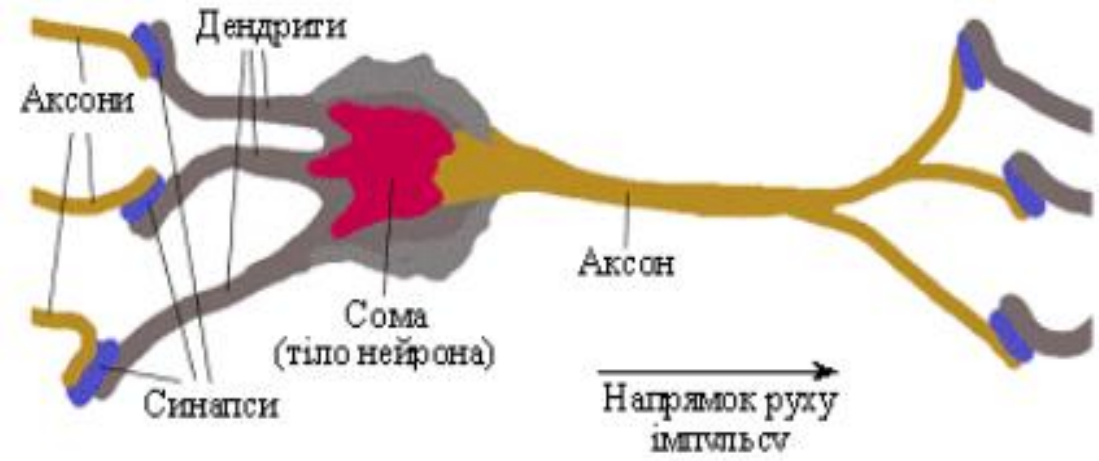


Рис. 3.18. Модель біологічного нейрона

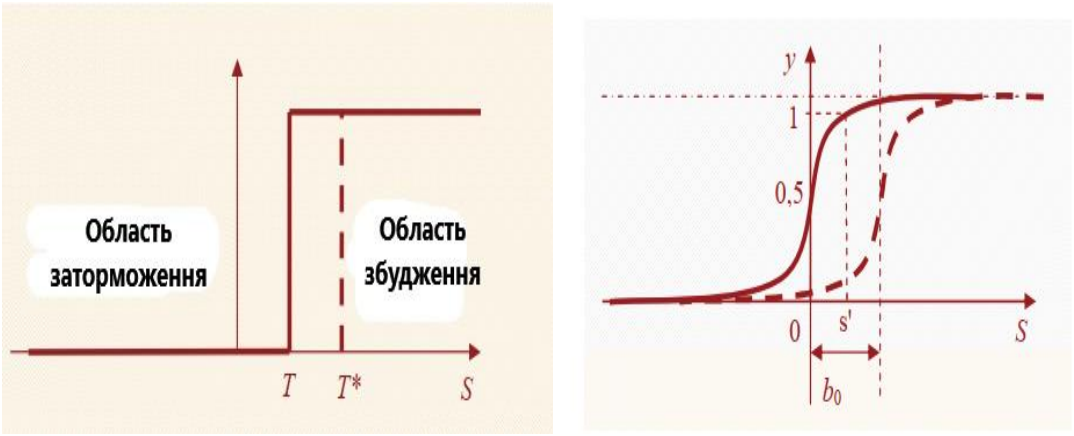
Розмір тіла клітини становить від 3 до 100 мкм. Усередині розташовано ядро. Нейрон з'єднується з іншими нейронами через відростки двох видів: численні тонкі, сильно розгалужені *дендрити* і єдиний *аксон*, більш товстий і розгалужується на кінці. Сигнали від інших нейронів надходять у клітку через так звані *синапси*, що утворюються у місцях контакту дендритів одного нейрона з тілом іншого, а передаються через *аксон.*

Передача сигналів усередині нервової системи являє собою складний електрохімічний процес. Нервові імпульси передаються між нейронами за допомогою спеціальних біохімічних речовин – нейромедіаторів, які є подразниками, змушують нейрон переходити в збуджений стан. Якщо розглядати процес спрощено, то під впливом нейромедіаторів синапси можуть змінювати здатність передавати сигнал. Іншими словами, кожному міжнейронному зв'язку (синапсу) можна поставити у відповідність деякий коефіцієнт або вагу, на який має множитися значення сигналу, що надходить через синапс. Ці ваги можуть приймати як *негативні*, так і *позитивні* значення. Зв'язки, які мають позитивні ваги, називають *збуджуючими*, а які мають негативні – *гальмуючими*. Сигнали, прийняті через синапси, надходять у тіло нейрона, де відбувається їх сумування. При цьому одні зв'язки є збуджуючими, а інші – гальмуючими.

Залежно від балансу збуджуючих і гальмуючих зв'язків нейрон сам може перейти в збуджений стан: як тільки сумарне збудження перевищує певний поріг активації, нейрон починає через аксон передавати сигнали інших нейронів. Інтенсивність

сигналу ролі не відіграє, так як працює принцип «все або нічого». Якщо співвідношення збуджуючих і гальмуючих зв'язків таке, що поріг активації перевищено, нейрон переходить у збуджений стан, якщо немає – то в заторможений (рис. 3.19 (А).

Ця ситуація пояснюється на рис. 3.19, де ілюструється так звана *активаційна функція нейрона*. Вона має вигляд сходинки. Якщо сума збуджуючих і гальмуючих сигналів, що надходять на вхід нейрона через синапси, не перевищує деякого порога T, то сигнал на виході нейрона, відповідно до активаційної функції, дорівнює 0. Якщо сума перевищує цей поріг, то нейрон переходить у стан збудження і передає сигнали іншим нейронам.



А Б

Рис. 3.19. Ступінчаста (А) і сигмоїдальна (Б) функції активації

Незважаючи на величезну кількість нейронів у мозку, вони становлять лише кілька відсотків від його обсягу; інше місце зайнято міжнейронними зв’язками, при цьому число зв’язків кожного нейрона не має аналогів в електронній техніці. Дослідження показали, що саме строго організована взаємодія величезної кількості синаптичних зв’язків забезпечує високу продуктивність мозку у плані обробки інформації.

Штучний нейрон (рис. 3.20) є або електронним процесорним елементом або програмним елементом, на основі якого будуються штучні нейронні мережі. Подібно біологічному прототипу штучний нейрон виконує зважене підсумовування своїх входів з подальшим нелінійним перетворенням результату, аналогічно

порівняно з порогом активації. Штучний нейрон складається з таких основних елементів, як:

– *набір вхідних зв'язків* (синапсів) x_i, кожна з яких має вагу w_i. Значення ваг нейронів можуть бути як позитивними, так і негативними;

– *суматор* для підсумовування вхідних сигналів xi, зважених вагами wi;

– *функція активації* f (s), що виконує перетворення (зазвичай нелінійне) значень з виходу суматора.

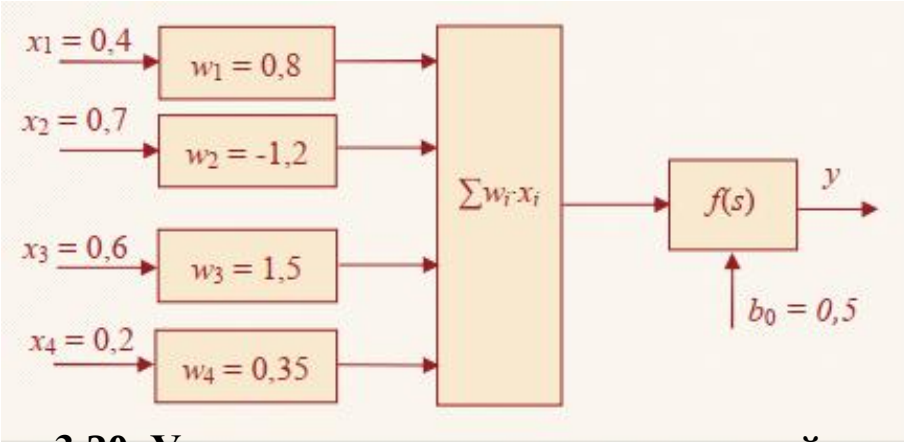


Рис. 3.20. Узагальнена модель штучного нейрона

Розглянемо приклад обчислення виходу нейрона із входами та вагами, показними на рис. 3.20. Будемо використовувати сигмоїдальну функцію активації (рис. 3.19 Б).

Нехай на вхід нейрона надходить 4 сигнали xi (i = 1 ... 4), де x1 = 0,4, x2 = 0,7, x3 = 0,6 і x4 =0,2 (рис. 3.19). Значення входів нейронів нормуються до 1. Кожен зв'язок має деяку вагу w_i. Значення ваг задамо таким чином: w_1 = 0,8, w_2 = -1,2, w_3 = 1,5 і w_4 = 0,35. Зсув нехай буде b_0 = 0,5.

Розрахуємо вихід нейрона S відповідно до формули:

$$S = w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n + b_0 = \sum_{i=1}^{n} w_i \cdot x_i + b_0$$

$$S = \sum_{i=1}^{n=4} w_i \cdot x_i + b_0 = 0{,}4 \cdot 0{,}8 + 0{,}7 \cdot (-1{,}2) + 0{,}6 \cdot 1{,}5 + 0{,}2 \cdot 0{,}35 = 0{,}32 - 0{,}84 + 0{,}9 + 0{,}07 + 0{,}5 = 0{,}95$$

Тепер обчислимо результат перетворення отриманої суми з допомогою сигмоїдальної функції, приймаючи крутизну a = 1:

$$y = \frac{1}{1 + e^{-as}} = \frac{1}{1 + e^{-1 \cdot 0{,}95}} = 0{,}72 .$$

3.3.2. Вибір структури мережі і методу навчання

Наступні питання, які необхідно вирішити при застосуванні нейромереж, це – ***структура мережі і метод навчання***.

Одним із найскладніших та найважливіших етапів у застосуванні нейромереж є вибір їх структури та методів навчання. Кількість шарів нейронів обмежується ресурсами комп'ютера або спеціалізованих мікросхем, на яких функціонує нейронна мережа. Зокрема, залежно від завдання, методу і призначення мережі вибирається і відповідний метод навчання (рис. 3.21).

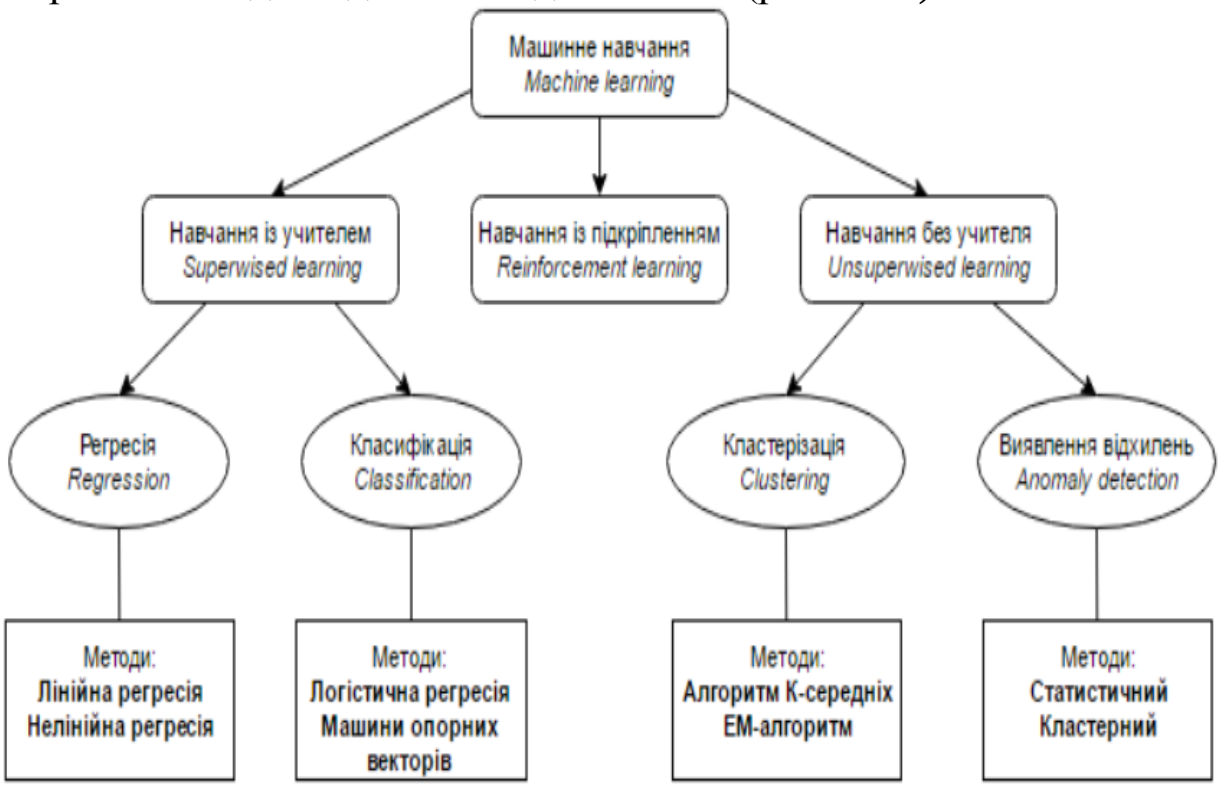


Рис. 3.21. Класифікація методів машинного навчання

Оскільки проблема синтезу нейромережі пов'язана із поставленими завданнями, у більшості випадків оптимальний варіант отримується на основі інтуїтивного підбору. Саме на можливості дослідження структури нейромереж та отримання оптимальної структури мережі необхідно звертати увагу при виборі нейроемуляторів.

Для визначення числа нейронів у прихованих рядах однорідних нейромереж можна скористатися формулою (3.21), що оцінює необхідне число синаптичних ваг (зв'язків) L_w у багатошаровій мережі із сигмоїдними функціями активації нейрону.

$$\frac{mN}{1+\log_2 N} \le L_W \le m\left(\frac{N}{m}+1\right)(n+m+1)+m, \quad (3.21)$$

де n – розмірність вхідного сигналу (наприклад, кількість вибраних макроекономічних факторів впливу на надходження до бюджету держави (ВВП, індекс інфляції тощо), m – розмірність вихідного сигналу, N – число фактів навчальної вибірки (наприклад, кількість звітних податкових періодів регіональних ДПІ перед ДПА України).

Після оцінки необхідної кількості ваг можна розрахувати число нейронів у прихованих шарах.

Число нейронів у прихованому шарі для двошарової мережі визначається за формулою (3.21):

$$L = \cdot\frac{L_W}{n+m}, \frac{N}{10} - n - m \le L \le \frac{N}{2} - n - m$$

На практиці часто користуються емпіричними залежностями, що позитивно себе зарекомендували. Наприклад, рекомендується використовувати такі формули для підбору нейронів прихованого шару:

$$L = (n+m) / 2. \quad (3.22)$$

$$Lmin = N/10 - n - m. \quad (3.23)$$

$$Lmax = N/2 - n - m. \quad (3.24)$$

Зазвичай, навчання мережі розподіляється на такі етапи:

1)ваговим коефіцієнтам і зміщенням мережі присвоюються малі випадкові значення із діапазону (w_{min}, w_{max});

2) входи (x_0, x_1,..., x_{N-1}) повинні розрізнятися для усіх елементів вибірки. При використанні багатошарової мережі як класифікатора бажаний вихідний сигнал (d_0,d_1,…,d_{N-1}) складається із нулів за виключенням одного елементу, що відповідає класу, до якого належить поточний вхідний сигнал;

3)подати на входи однин із вхідних векторів, які мережа повинна навчитися розрізняти й обчислити її вихід;

4)для налагодження вагових коефіцієнтів використовується рекурсивний алгоритм, що спочатку застосовується до вихідних нейронів мережі, а потім проходить мережу у зворотньому напрямі до першого шару (back propagation). Синаптичні ваги налагоджуються за формулою:

$$w_{ij}(t+1)=w_{ij}(t)+rg_j x'_j,$$

де w_{ij} – вага від нейрона i або від елемента вхідного сигналу $\underline{i}$ до нейрона j у момент часу t, x'_j – вихід нейрона i або i-й елемент вхідного сигналу, r – крок навчання, g_j – значення помилки для нейрона j. Якщо нейрон з номером належить останньому шару, то помилка визначається так:

$$g_j = y_j(1 - y_j)(d_j - y_j),$$

де d_j – бажаний вихід нейрона j, y_j – поточний вихід нейрона j.

3.3.3. Огляд структур та типів нейромереж

Задачі класифікації вирішуються за допомогою мереж таких типів: багатошарового перцептрона *(Multilayer Perceptron)*, радіальної базисної функції *(Radial basis function networks)*, мереж Кохонена *(Kohonen networks)*, імовірнісних Байєсових нейронних мереж *(Probabilistic (Bayesian) networks)*.

Мережа типу радіальної базисної функції (RBF) має проміжний шар із радіальних елементів, кожен із яких відтворює Гаусову поверхню відклику:

$$\Phi(x, a, \sigma^2) < x\} = \frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}},$$

де a – середнє;

σ – стандартне відхилення;

e – число Эйлера (2.71...);

π – число Пі (3.14...);

$-\infty < x < \infty$.

Оскільки ці функції нелінійні, для моделювання довільної функції немає необхідності брати більше одного проміжного шару нейронів. Для моделювання довільної функції необхідно взяти достатню кількість радіальних елементів. Для комбінації виходів прихованого шару радіальних елементів, як показали дослідження, достатньо взяти їх лінійну комбінацію, тобто зважену суму гаусових функцій. Мережа *RBF* має вихідний шар, що складається із елементів з лінійними функціями активації.

Мережі *RBF* мають ряд переваг перед багатошаровими перцептронами. Вони можуть моделювати довільні нелінійні функції за допомогою лише одного проміжного шару, параметри лінійної комбінації у вихідному шарі можна повністю оптимізувати за допомогою добре відомих методів лінійного програмування, які працюють набагато швидше і не мають проблем із локальними мінімумами функцій.

Мережі Кохонена (Kohonen Neural Networks).

Самоорганізуючі карти ознак Кохонена *(Kohonen's Self Organizing Feature Map (SOM)), м*ережі, що принципово відрізняються від інших методів некерованого навчання. Вони дозволяють представляти результати кластеризації у вигляді двовимірних карт, де відстані між об'єктами відповідають відстаням між їх векторами в багатовимірному просторі, а самі значення ознак відображають різними кольорами і відтінками. Можна провести аналогію між SOM і звичайною географічною картою, де розміщення об'єктів і відстані між ними відповідають їхнім розташуванням на земній поверхні. Однак, крім горизонтальних координат, необхідно показати і рельєф – вершини високих гір, пагорбів, а також глибину водойм.

Для цього використовується спеціальна колірна гамма. Так, висота місцевості зображується за допомогою відтінків коричневого, глибина морів і океанів – синього: чим вище або глибше об'єкт, тим більш темним кольором він забарвлюється. Таким чином, двовимірна карта дозволяє представляти тривимірні дані.

Якщо розмірність простору ознак набору даних, на якому будується SOM, дорівнює двом (наприклад, вік і дохід), то і вектори ваг нейронів також будуть двокомпонентними, і відображення

результатів кластеризації в цьому випадку проблем не викличе. Але в більшості практичних застосувань доводиться мати справу з даними, для яких розмірність простору ознак більше, ніж два.

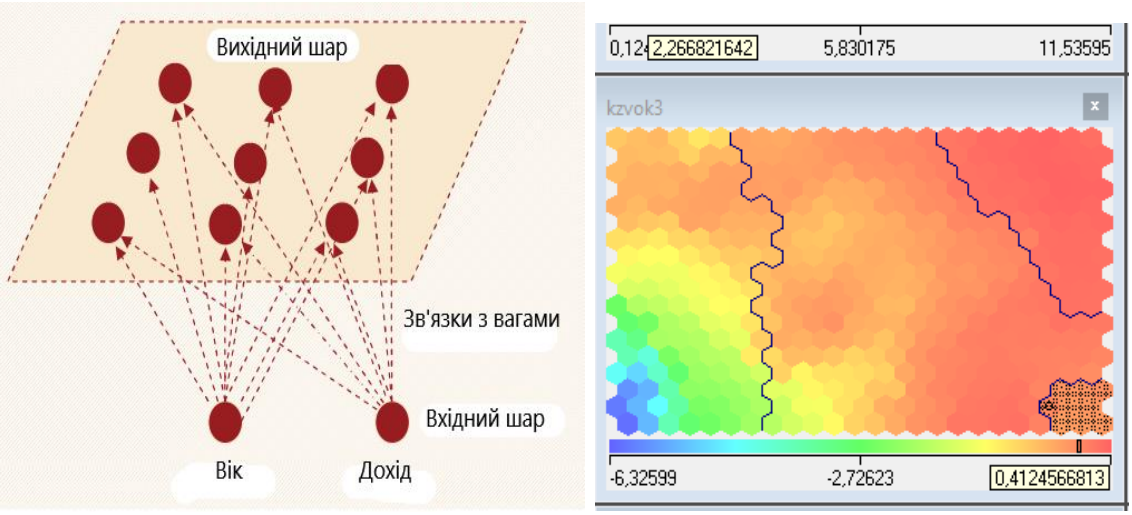


Рис. 3.22. Двовимірні карти Кохонена

При керованому навчанні на вхід мережі подаються разом із вхідними змінними вихідні значення, і мережа повинна навчитися відображати спостереження. При некерованому навчанні на входи мережі подаються тільки значення вхідних змінних, тобто мережі Кохонена навчаються розуміти саму структуру даних.

Чим відрізняються поняття *«мережа Кохонена»* і *«карта Кохонена»? По-перше, мережа Кохонена* використовується тільки для кластеризації об’єктів, а візуалізація результатів буде проводитися традиційними способами, тобто за допомогою таблиць або діаграм. Водночас карта Кохонена дозволяє візуалізувати результати кластеризації, у тому числі багатовимірні.

По-друге, у мережі число вихідних нейронів відповідає кількості кластерів, яке повинно бути отримано, а в карті – кількості сегментів, з якого буде складатися карта, або, іншими словами, розміром карти. З вищезазначеного можна зробити висновок, що карта Кохонена – це окремий випадок мережі Кохонена, коли число вихідних нейронів набагато перевищує число кластерів.

Розмір карти визначається ступенем її детальності: чим вище число сегментів, тим докладніше представлено розподіл ознак об’єктів. До звичайної мережі Кохонена можна повернутися, зменшивши число сегментів карти до потрібної кількості кластерів.

Мережа Кохонена може розпізнавати кластери в даних, а також установлювати подібність класів. Якщо в даних розпізнані класи, то їх можна позначити, після чого мережа може вирішувати задачі класифікації. Іншою можливістю застосування мереж Кохонена є виявлення нових явищ (знань). Мережа розпізнає кластери в навчальних вибірках і відносить усі дані до того чи іншого кластеру. Якщо після навчання мережа зустрінеться з невідомим набором, що не можна віднести до будь-якого кластеру, вона не зможе його класифікувати і тим самим виявить його новизну.

Для реалізації алгоритму мережі Кохонена необхідно визначити міру подібності нейронів. Алгоритм Кохонена можна записати таким чином:

– Крок 1. Ініціалізація мережі: Ваговим коефіцієнтам мережі, число яких дорівнює розміру мережі M N, присвоюються малі випадкові значення. $D_j(t)$ – множина нейронів, що вважаються сусідніми до нейрона j момент часу t. Розміри зони сусідства зменшуються з часом. $D_j(o)$ – початкова зона сусідства (рис. 3.21).

– Крок 2. Пред'явлення мережі нового вхідного вектору вхідних сигналів.

– Крок 3. Обчислення відстані d_j від вхідного сигналу до кожного нейрона j за формулою: $d_j = \sum_{i=1}^{N} (x_i(t) - w_{ij}(t))^2$.

– Крок 4. Вибір нейрона j, для якого відстань d_j є найменшою.

– Крок 5. Налагодження ваг для нейрона j* і усіх нейронів із його зони сусідства Dj*(t) за формулою:

$$w_{ij}(t+1) = w_{ij} + \eta (x_i(t) - w_{ij}(t),$$

де η – крок навчання ($0 < \eta < 1$), що зменшується з часом до нуля.

– Крок 6. Повернення до кроку 2.

Недоліки карт Кохонена

Карти Кохонена – потужний засіб візуалізації і розвідувального аналізу даних, який відрізняється від класичних статистичних процедур, під час яких перевіряється деякий набір висунутих гіпотез. Але у карт Кохонена є ряд недоліків і обмежень. Хоча мінуси не зменшують переваги методу, в деяких випадках потрібно

обережне ставлення як до самого процесу побудови карт, так і до інтерпретації його результатів.

Одним із найбільш серйозних недоліків карт Кохонена, що пред'являються деякими дослідниками в галузі Data Mining, є те, що самі по собі карти завдання кластеризації не вирішують, а дозволяють за кольором карт висувати гіпотези про наявність кластерної структури і залежностей в наборі даних. *Висунуті гіпотези потрібно підтверджувати іншими методами*, тому що вони можуть виявитися помилковими.

Імовірнісні нейронні мережі Байєса (Probabilistic (Bayesian) networks)

Під час вирішення задач класифікації за допомогою імовірнісних мереж можна оцінити щільність імовірності для кожного класу, порівняти між собою імовірності належності різних класів і вибрати з них найімовірніший. Тобто імовірнісна мережа апроксимує щільність імовірності.

Традиційним підходом до вирішення задачі є отримання оцінки для щільності імовірності за наявними даними. При цьому передбачається, що щільність має нормальний закон розподілу. Після цього оцінюються параметри моделі.

Найвідомішим алгоритмом навчання нейромереж є *метод зворотнього поширення* (back propagation), що був запропонований у 1986 р. Рухмельхартром, Макклеландом та Вільямсом. Іншими, більш сучасними і ефективнішими у реалізації, є *Квазіньютоновські методи*.

Для задач класифікації суб'єктів оподаткування можна використовувати **імовірнісні мережі Байєса** *(Probabilistic Neural Networks, PNN),* для дослідження якого у вибірці була сформована одна номінальна вихідна величина.

У нашому випадку це стовпець із попередньо визначеними класами підприємств – ЗЗ, СЗ, СП, ЗП. Коефіцієнт згладжування *(Smoothing)*, що визначає ширину гаусових функцій із центром у кожній навчальній вибірці, які є функціями активації першого шару, був підібраний на рівні 0,39. Використовуючи можливість PNN визначати ціну різних типів помилок за допомогою четвертого шару, було вибрано структуру 4-и шарової PNN (рис. 3.23).

Для знаходження рішення з найменшими втратами, уникнення «неясних» класифікацій і «невизначених рішень» у режимі *Пре/пост-обробка (Pre/post Processing)* вибираємо рівень *Прийняття-Acccept* рівним *1*, а рівень *Відхилення-Reject* рівним 0.

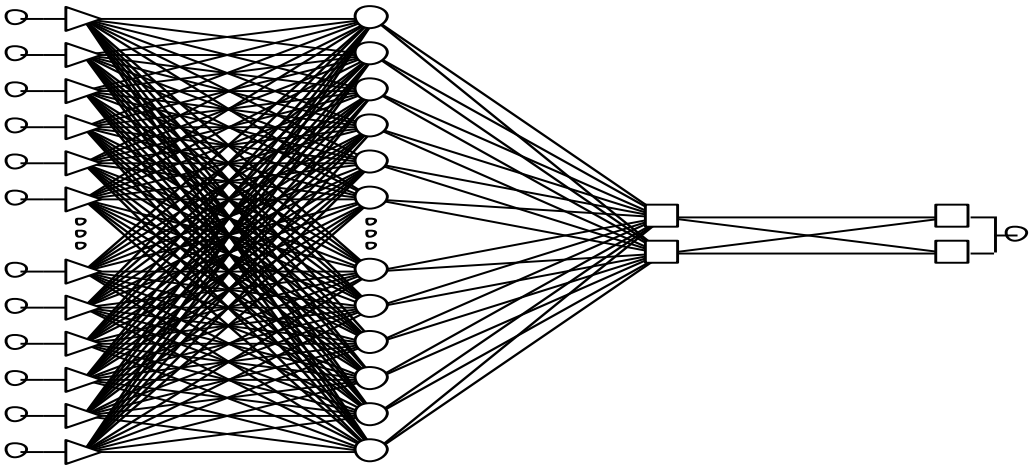


Рис. 3.23. Структура імовірнісної чотиришарової мережі Байєса

Для навчання мережі була сформована навчальна (90 підприємств) і контрольна вибірки (62 підприємства). На навчальній вибірці відсоток правильно класифікованих ризикових підприємств (ЗЗСЗ), тобто тих, що є прибутковими, а стануть збитковими або залишаться збитковими, рівний 82,2 %, що значно вище ніж при застосуванні дискримінантного аналізу (69,7 %). Прогін мережі на контрольній вибірці дає дещо менший відсоток точності, проте також значно вищий результату класифікації дискримінантними функціями – 79,2 %. Результати застосування імовірнісних баєсових мереж наведені у таблиці 3.4.

Отже, методи дискримінантного аналізу, алгоритми ймовірнісних мереж Байєса та самонавчальних карт Кохонена дають змогу вирішувати задачі класифікації і розпізнавання образів суб'єктів господарювання, отримувати імовірності їх станів на основі фінансово-статистичної та податкової звітності.

Таблиця 3.4

Результат класифікації за допомогою імовірнісної мережі Баєса

	% правильно класифікованих			
	Навчальна вибірка (90 підпр.)		Контрольна вибірка (62 підпр.)	
	0,822	0,600	0,792	0,553
	ЗЗСЗ	СПЗП	ЗЗСЗ	СПЗП
Усього	45	45	24	38
Правильно класифіковано	37	27	19	21
Неправильно класифіковано	8	18	5	17
Не визначено	0	0	0	0
Підприємства групи ОЗСЗ	37	18	19	17
Підприємства групи СПОП	8	27	5	21

Таблиця 3.5

Порівняння результатів застосування дискримінантного аналізу і нейроалгоритмів у класифікації підприємств

№	Метод	Похибка навчання мережі	Правильно класифіковано, %		Правильно класифіковано, %			
			ЗЗСЗ	СПЗП	ЗЗ	СЗ	СП	ЗП
1.	Дискримінант-ний аналіз	-	68,1	71,0	71,7	21,7	34,5	87,0
2.	Імовірнісні мережі Байєса	0,25	82,2	60,0	-	-	-	-
3.	Карти Кохонена	0,29	-	-	75,0	52,0	51,0	69,0

Якщо перевагою дискримінантного аналізу є отримання аналітичних записів класифікаційних функцій, які можна інтегрувати у будь-яку інформаційну систему алгоритмічними мовами програмування, то перевагами нейромережних технологій є можливість пошуку нових явищ і знань у зашумлених і неповних даних.

3.3.4. Кластерний аналіз і класифікація регіонів на основі параметрів податкового навантаження

Розбиття економічних об'єктів, зокрема регіонів України, на кластери, подібних за різними економічними параметрами, наприклад показниками соціально-економічного розвитку, є важливим завданням у прийнятті рішень. Оскільки результати аналізу і прийняте рішення може стосуватися всієї групи об'єктів, що входять до кластеру.

Кластерний аналіз (англ. *Data clustering*) – задача розподілу заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Завдання кластеризації належать до статистичної обробки, а також до широкого класу завдань навчання без вчителя групи статистичних медотів Data Mining.

Для апробації методу кластеризації можна перевіряти різні гіпотези, наприклад, щодо наявності кластерів регіонів України, які відрізняються між собою за показниками податкового зобов'язання регіону ($ПЗ_{ДН}$) та валового регіонального продукту у розрахунку на душу населення ($ВРП_{ДН}$) (або величини сплати податків і величини ВРП у розрахунку на душу населення тощо). Наявність таких кластерів та динаміка і напрям зміни відстані між такими кластерами може вказувати на позитивні чи негативні тенденції соціально-економічного розвитку регіонів, ефективність здійснення податкової політики у регіонах.

Для кластерного аналізу були сформовані дані за вищевказаними показниками.

Як було зазначено, рішенням задачі кластерного аналізу є розподіл одиниць сукупності на групи, що відповідає деякому критерію оптимальності. Критерій може являти собою деякий функціонал, що відображає відповідні рівні бажаних розподілів на групи. Цей функціонал будемо називати *цільовою функцією.* Однорідність сукупності задається правилом обчислення певної

метрики, що характеризує ступінь подібності *i*-ої та *j*-ої одиниць сукупності. Найпопулярнішою і найуживанішою є Евклідова метрика, у основі якої лежить розрахунок Евклідової відстані між одиницями сукупності:

$$d\left(X_i, X_j\right) = \left[\sum_{h=1}^{p}\left(x_{hi} - x_{hj}\right)^2\right]^{1/2}, \qquad (3.25)$$

де X_i, X_j – одиниці сукупності,

x_{hi}, x_{hj} – стандартизовані значення *h*- ї ознаки і- ї та *j*-ї одиниць сукупності.

Використовуючи *модуль кластерного* аналізу Ststistica 10 було сформовано три групи регіонів. Ознаками є податкове зобов'язання та ефективна ставка оподаткування ВРП, створена у регіоні у розрахунку на душу населення ($ПЗ_{дн}$ і $Т_{ВДВ}$ відповідно). Показники розраховані на основі даних ДФС України за податковими надходженнями та даними Міністерства економіки України по ВРП.

Перевірка гіпотези про наявність кластерів регіонів України, що розрізняються за показниками величини валового регіонального продукту та величини сплати податків у розрахунку на одного платника податку (звітуючого), здійснювалася у пакеті Statistica за даними за 2008–2013 роки. Як метрики використовувалася Евклідова відстань. Приклад дендрограми кластерного аналізу для 2013 року показано на рис. 3.24.

Якщо вважати метою економіки держави зближення регіонів за рівнем економічного розвитку, то спостерігаємо, навпаки, негативну динаміку, а саме збільшення міжгрупової відстані. Для 2013 року практично регіони були віднесені до трьох кластерів. В окремий кластер увійшли Черкаська та Донецька області. Два чітко виражені кластери утворили такі регіони: 1-й кластер (Херсонська, Миколаївська, Чернівецька, Львівська, Запорізька, Закарпатська, м. Севастополь, Хмельницька, Одеська, Кіровоградська, Волинська області); 2-й кластер (Тернопільська, Житомирська, Рівненська, Харківська, Луганська, Дніпропетровська, Київська, м. Київ, Івано-Франківська, Вінницька, Чернівецька, АР Крим).

Більш детальний аналіз показу розходження центрів кластерів, що вказує на негативні тенденції в соціально-економічному розвитку держави в цілому.

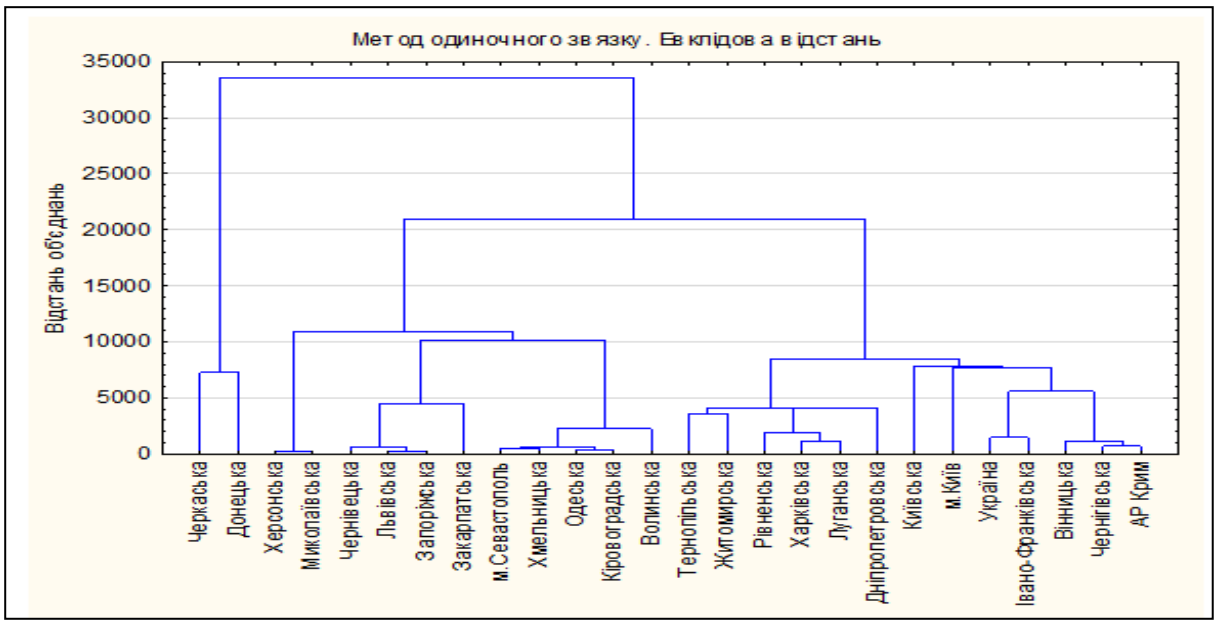


Рис. 3.24. Кластерна дендрограма регіонів за показниками валового регіонального продукту та сплати податків у розрахунку на одного платника податків (2013 рік)

Апробація кластерного аналізу у середовищі Oracle Data Miner також показала доцільність використання в інформаційно-аналітичній підсистемі ДФС України даного аналітичного інструменту, оскільки інформаційна система ДФС функціонує на платформі Oracle. Результати застосування класифікаційних моделей кластерного аналізу за допомогою Statistica та Oracle Data Miner на даних про сплату податків та величину регіонального ВРП показують доцільність інтеграції цих інструментів в інформаційно-аналітичній системі податкового відомства.

Економічні висновки щодо отриманих результатів показують негативну тенденцію протягом 2008–2013 років щодо віддалення кластерів регіонів за показниками ВРП та сплати податків у розрахунку на одного звітуючого платника. Для побудови корпоративного інформаційно-аналітичного середовища доцільно організувати серверне середовище для накопичення даних Державної служби статистики про соціально-економічний розвиток регіонів.

3.3.5. Дискримінантний аналіз у прогнозуванні платоспроможності суб'єктів оподаткування

Завдання перетворення фіскальної служби держави на сервісну службу напряму пов'язано із запобіганням банкротства підприємницьких структур, що передбачає їх консультування, сприяння спрощенню податкового адміністрування. Важливим аспектом у цьому напрямі є можливість аналізу податковим відомством статистичної звітності платників податків, їх фінансових показників як для виявлення ухилення від оподаткування, так і для консультування щодо уникнення збитковості.

Дискримінантний аналіз – один з методів багатовимірного аналізу, метою якого є класифікація об'єктів, тобто віднесення об'єкта до однієї з відомих груп деяким оптимальним способом (наприклад, роподіл сукупності підприємств на кілька однорідних груп за значеннями будь-яких показників виробничо-господарської діяльності).

Дискримінáнтний· анáліз – різновид багатовимірного аналізу, призначеного для вирішення завдань розпізнавання образів.

Отримана у результаті фінансового аналізу система показників, що характеризують стан підприємства, дозволяє виявити слабкі сторони у економіці підприємства. Одні показники можуть бути у критичній зоні, інші – мати задовільні значення. Проте на основі такого аналізу неможливо зробити однозначний висновок про те, чи втратить платоспроможність підприємство найближчим часом. Висновки про імовірність банкротства можна зробити тільки на основі зіставлення показників одного підприємства й аналогічних підприємств, що збанкрутували, або уникли її. Очевидно повних аналогів підприємств бути не може, що ускладнює цю проблему. Завданням у цьому випадку є пошук емпіричного рівняння деякої дискримінантної границі, що розподіляє всі можливі комбінації показників на два класи: 1) комбінація показників,

при яких підприємство збанкрутує (втратить платоспроможність); 2) комбінація показників, при яких підприємству банкрутство не загрожує.

Завдання класифікації, тобто поділ деякої сукупності аналізованих об'єктів на класи шляхом побудови так званої класифікаційної функції у вигляді кореляційної моделі, як правило, реалізуються за допомогою методів дискримінантного аналізу, який є розділом класичного факторного статистичного аналізу і реалізується більшістю статистичних пакетів.

Дискримінантний аналіз прийнято використовувати у діагностиці і прогнозуванні банкротства підприємств. Ці питання є актуальними і для підрозділів податкової служби – офісів великих платників, а у майбутньому адміністрацій податкових округів. Одним із основних завдань цих підрозділів є моніторинг фінансового стану бюджетоутворюючих СПД, підтримка їх платоспроможності і, відповідно, податкоспроможності.

Вперше задача прогнозування банкрутства була вирішена американським ученим Е. Альтманом, який на основі аналізу 19 підприємств отримав рівняння кореляційної лінійної функції, що описує положення дискримінантної межі між двома класами підприємств:

$$Z=a_0 + a_1K_п + a_2K_{фз},$$

де Z – показник класифікаційної функції;

a_0 – константа;

$K_п$ – коефіцієнт покриття (ліквідності);

$К_{фз}$ – коефіцієнт фінансової залежності у процентах;

a_1, a_2 – параметри рівняння, що показують ступінь впливу коефіцієнтів покриття і коефіцієнта фінансової залежності на імовірність банкрутства.

У результаті обробки статистичних даних Е. Альтманом була отримана залежність:

$$Z= –0,3877 –1,073\ K_п + 0,0579\ K_{фз}, \qquad (3.26)$$

де $Z = 0$ є значенням дискримінантної межі. Для підприємств у яких Z=0, ймовірність збанкрутувати дорівнює 50 %. Якщо $Z < 0$,

то ймовірність банкротства менша 50 % і далі зменшується відповідно до зменшення Z. При Z > 0 ймовірність банкротства більша 50 % і зростає зі зростанням Z. Відомі також п'ятифакторні та семифакторні моделі Альтмана. П'ятифакторна модель, на думку Альтмана, дає точність до 95 % на горизонті в один рік і 85 % на горизонті в 2 роки. Семифакторна модель Альтмана дозволяє прогнозувати банкрутство на горизонті у 5 років з точністю до 70 %.

Оскільки ці моделі відповідають умовам так званої нормативної економіки, застосовувати їх в умовах України неможливо. Найбільш прийнятною для оцінки ризику банкрутства підприємств, що функціонують в умовах перехідної економіки України, є чотирифакторна R-модель [58]:

$$R = 8{,}3 \cdot K_1 + K_2 + 0{,}054 \cdot K_3 + 0{,}63 \cdot K_4, \qquad (3.27)$$

де R – показник ризику банкрутства організації; K_1 – відношення оборотного капіталу до активів; K_2 – відношення чистого прибутку до власного капіталу; K_3 – відношення виручки від реалізації до активів; K_4 – відношення чистого прибутку до затрат.

За допомогою R-моделі можна прогнозувати банкрутство на три квартали з імовірністю 81 % (табл. 3.6).

Таблиця 3.6

Імовірність банкрутства організацій відповідно до значення моделі «R»

Значення R	Імовірність банкрутства, %
Менше 0	Максимальна (90–100)
0–0,18	Висока (60–80)
0,18–0,32	Середня (35–50)
0,32–0,42	Низька (15–20)
Більше 0,32–0,42	Мінімальна (до 10)

Для забезпечення умов соціально-економічного розвитку регіонів, виконання планових показників надходжень до бюджетів усіх рівнів соціально значимими і важливими для регіональних фіскальних органів є підтримка бюджетоутворюючих підприємств, визначення і прогнозування на ранніх етапах можливості їх

переходу у розряд збиткових, запобігання зменшенню податкового зобов'язання за галузевою ознакою та видом діяльності.

Наявність фінансово-статистичної та податкової звітності СПД дає змогу поставити завдання їх класифікації за вибраними ознаками.

Розглянемо приклад дискримінантного аналізу за допомогою модуля *Statistica Discriminant Function Analysis* .

Є дані по 377 підприємствах, які поділені на такі: 1 група діючих підприємств – усього 320; 2-га група із 46 підприємств із призупиненими рахунками; 3-тя група 11 підприємств-банкротів (станом на 1 січня календарного року). Для дискримінантного аналізу будемо застосовувати фінансові такі коефіцієнти, взяті на дату за рік до оцінювання їх стану (у дужках дається позначення, що використовується в аналізі): покриття (kp); загальної ліквідності (kzl); абсолютної ліквідності (kal); обіговості оборотних коштів (kook), забезпечення власними оборотними коштами (kzvok); оборотності дебіторської заборгованості (kodz); оборотності кредиторської заборгованості (kokz); фінансової незалежності (kfn); фінансової стабільності (kfs).

Завданням є побудова дискримінантних функцій, які будуть пов'язувати стан підприємства (тобто ознаками груп 1–3) із фінансовими коефіцієнтами, взятим за рік до оцінювання стану, тобто побудова прогнозних функцій стану підприємств на період до 1 року на основі фінансових коефіцієнтів. Зрозуміло, що незбалансована вибірка не може дати достаньої точності класифікації, зокрема для кількості підприємств 2-ї групи 46 підприємств із призупиненими рахунками і 3-ї групи із 11 банкротів.

Задачею є знаходження дискримінантної границі (дискримінантної функції) f_j, що може визначатися для кожного із станів підприємства *(j=1,2, ..., p)* (3.28)*:*

$$f_j = a_{0j} + a_{1j} x_{1j} + a_{2j} x_{2j} + \ldots + a_{mj} x_{mj}, \qquad (3.28)$$

де a_{ij} – коефіцієнти функції, x_{ij} – середнє значення *i*-ої ознаки в *j*-му класі.

Значення f_j для p груп є центрами їх тяжіння і називаються *центроїдами.* Процедура класифікації ґрунтується на визначенні геометричної близькості h-ї одиниці до центроїдів виділених груп. Належність об'єкта до групи визначається на основі відстані Махаланобіса, яка визначається так:

$$D^2 = (n-p)\sum_{i=1}^{m}\sum_{k=1}^{m} b_{ik}(x_{ijh} - \overline{x_{ij}})(x_{kjh} - \overline{x_{kj}}). \qquad (3.29)$$

Дискримінантна функція максимізує різницю між класами і мінімізує дисперсію всередині класу. Критерієм оптимального поділу сукупності на класи є максимум відношення між класової варіації до внутрішньокласової. Міжкласову варіацію характеризує квадрат різниці центроїдів $(f_j - f_s)$, а внутрішньокласову – середній квадрат відстаней між точками, що належить j-му класу x_{ijh} центроїдами цих класів f_j:

$$\sigma_f^2 = \frac{\sum_{j=1}^{p}\sum_{h=1}^{n_j}(a_{ijh} - x_{ij})^2}{\sum_{j=1}^{p} n_j - p},$$

де n_j – кількість одиниць j-го класу.

Таким чином, критерієм оптимального поділу на класи є відношення:

$$D^2 = \frac{(f_{j-} f_k)^2}{\sigma_f^2},$$

що прийнято називати узагальненою міжкласовою відстанню Махаланобіса.

Для оцінювання спроможності дискримінантної функції розпізнавати класи у багатовимірному ознаковому просторі використовують λ-статистику Вілкса (*Wilks lambda):*

$$\lambda = \prod_{j=1}^{p} \frac{1}{1+\lambda_j}, \qquad (3.30)$$

де λ_j – властиві значення матриці коваріацій.

На підготовчому етапі коефіцієнти покриття (kp) і фінансової незалежності (kfn) були відібрані як значимі ознаки. Отримані класифікаційні дискримінантні функції, наведені у табл. 3.7, дають можливість з імовірністю, що дорівнює якості класифікації, відносити до одного із класів підприємств нові дані нерозпізнаних підприємств.

Таблиця 3.7

Класифікаційні функції на основі коефіцієнта покриття (kp) і фінансової незалежності (kfn)

	G1	G2	G3
	p=0,83784	p=0,13514	p=0,02703
KP	-0,0816	-0,1849	-0,2071
KFN	12,1871	12,8309	9,0971
Constant	-4,3820	-6,4869	-5,8065

У табл. 3.8 показано результати класифікації підприємств за допомогою дискримінантного аналізу.

Таблиця 3 .8

Результати класифікації підприємств за допомогою дискримінантного аналізу

Функція	Якість класифікації (%)	G1	G2	G3
		P = 0,85	P = 0,12	P = 0,03
G1 (діючі)	98,5	327	1	4
G2 (призупинені рахунки)	2,2	43	1	3
G3 (банкроти)	36,4	7	0	4
Усього	85,1	377	2	11

Як видно із табл. 3.8, загальна якість класифікації підприємств становить 85,1 %. Якість класифікації діючих підприємств 98,5 %. Класифікація інших типів підприємств є недостатньою.

Розрахунок канонічних коренів у Statistica дає можливість проводити графічний аналіз результату класифікації. На рис. 3.25 показано результат канонічного аналізу на основі дискримінантних функцій G1, G2, G3 у середовищі Statistica.

Як видно із рис. 3.25, із двох коренів *Канонічний корінь 1* (КК1) дискримінує (розподіляє) сукупність підприємств за значенням КК1 ≈ 0, тобто підприємства, для яких значення канонічного кореня КК1≥0, перебувають у зоні ризику переходу у групу збиткових підприємств.

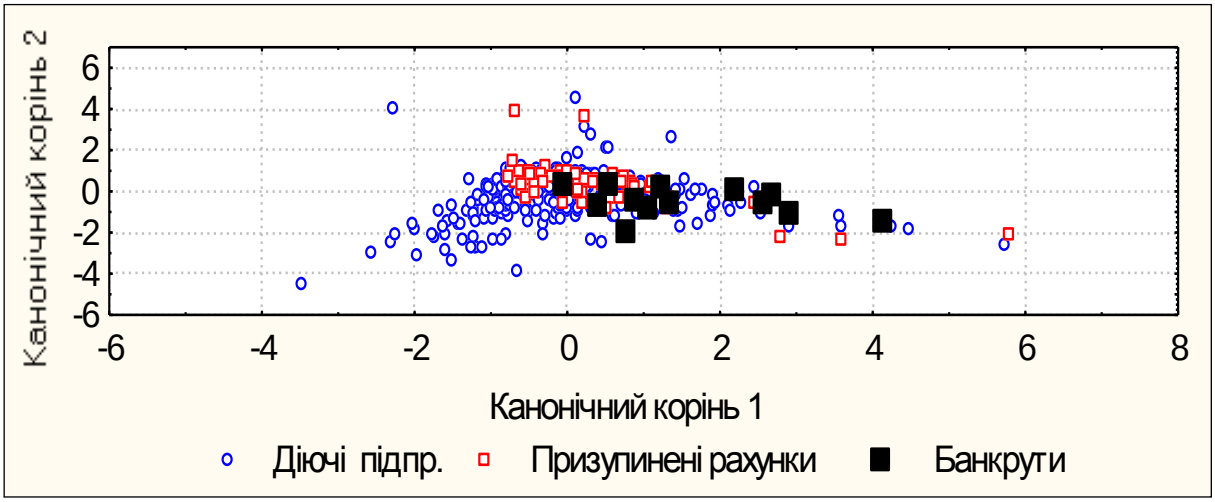


Рис. 3.25. Результати канонічного аналізу на основі функцій G1–G3

Відповідно до методології дискримінантного аналізу нове підприємство належить до того класу підприємств, значення функції для якого (G1-G3) приймає найбільше значення. Наявність імовірностей стану підприємств у майбутньому дозволить приймати обґрунтовані рішення щодо видів та типів упереджувальних консультаційних або контролюючих заходів, що передбачені законодавством.

Питання для самоконтролю до теми 3.3

1. Опишіть історію появи штучних нейронних мереж.

2. Назвіть основні елементи біологічного нейрона. Як вони відображаються у математичній моделі нейрона?

3. Для чого служить активаційна функція нейрона (функція активації)?

4. Наведіть модель та основні елементи штучного нейрона.

5. Які основні питання потрібно рзглянути під час застосування нейромереж?

6. Охарактеризуйте методи машинного навчання. Наведіть приклади.

7. Якщо n – розмірність вхідного сигналу і n = 10, m – розмірність вихідного сигналу і m =3, N – число фактів навчальної вибірки і N=1000. Яке число нейронів у прихованому ряді для дворядної мережі потрібне для нейромережі?

8. Перерахуйте основні методи та етапи навчання нейромережі.

9. Охарактеризуйте нейромережу типу радіальної базисної функції (RBF).

10. Опишіть особливості, переваги та недоліки мережі Кохонена (Kohonen Neural Networks).

11. Охарактеризуйте імовірнісні нейронні мережі Байєса (Probabilistic (Bayesian) networks).

12. Охарактеризуйте метод кластеризації. Математична суть і призначення.

13. Що описує графічне подання результату кластеризації дендрограма?

14. У чому суть дискримінантного аналізу. Формула дискримінації, коефіцієнти, використання.

15. Опишіть суть дискримінантного аналізу у прогнозуванні платоспроможності суб'єктів оподаткування.

16. Якщо: Кп – коефіцієнт покриття (ліквідності) = 0,8; Кфз – коефіцієнт фінансової залежності у відсотках = 25. Яка ймовірність банкротства підприємства за моделлю Альтмана?

17. У чому відмінність моделей дискримінантного аналізу Альтмана і R-моделі?

3.4. Застосування асоціативних правил в аналізі діяльності господарюючих об'єктів

3.4.1. Основні відомості про асоціативні правила

Аффінітивний аналіз (affinity analysis) – один з поширених методів Data Mining. Його назва походить від англійського слова affinity, яке в перекладі означає «близькість», «схожість».

Мета пошуку асоціативних правил – дослідження взаємного зв'язку між подіями, які відбуваються спільно. Різновидом аффінітівного аналізу є аналіз ринкового кошика (market basket analysis), мета якого – виявити асоціацію між різними подіями, тобто знайти правила для кількісного опису взаємного зв'язку між двома або більше подіями. Такі правила називаються *асоціативними правилами* (*association rules*).

Приклади застосування асоціативних правил

– виявлення наборів товарів, які в супермаркетах часто купуються разом або ніколи не купуються разом;

– визначення частки клієнтів, що позитивно ставляться до нововведень у їх обслуговуванні;

– визначення профілю відвідувачів веб-ресурсу;

– визначення частки випадків, у яких нові ліки показують небезпечний побічний ефект.

Базовим поняттям у теорії асоціативних правил є транзакція – деяка множина подій, що відбуваються спільно.

Типова транзакція – придбання клієнтом товару в супермаркеті. У переважній більшості випадків клієнт купує не один товар, а набір товарів, який називається ринковим кошиком. При цьому виникає питання: чи є купівля одного товару в кошику наслідком або причиною купівлі іншого товару, тобто чи пов'язані ці події? Цей зв'язок і встановлюють асоціативні правила, наприклад, може

бути виявлено асоціативне правило, яке стверджує, що клієнт, який купив молоко, з ймовірністю 75 % купить і хліб.

Наступне важливе поняття – предметний набір або ринковий кошик.

Це непорожня множина предметів, що з'явилися в одній транзакції.

Аналіз ринкового кошика – це аналіз наборів даних для визначення комбінацій товарів, пов'язаних між собою. Іншими словами, проводиться пошук товарів, присутність яких у транзакції впливає на ймовірність наявності інших товарів або комбінацій товарів.

Сучасні касові апарати в супермаркетах дозволяють збирати інформацію про покупки, яка може зберігатися в базі даних. Потім накопичені дані можуть використовуватися для побудови систем пошуку асоціативних правил.

На основі вищезазначеного сформулюємо завдання, для вирішення яких можна застосовувати асоціативні правила в адмініструванні податків.

Наприклад:

– виявлення груп платників, які присутні у схемах взаєморозрахунків, що призводять до порушень податкового законодавства;

– визначення груп та типів господарюючих об'єктів, що мають позитивну податкову історію;

– визначення профілю типових господарюючих суб'єктів, що належать до відповідних груп ризику;

– визначення типових випадків, у яких зміна ставок податків вказує на небезпечну тенденцію надходжень до бюджету.

Розглянемо приклад, який при достатньому досвіді в адмініструванні податків, контрольно-перевірочній роботі можна трансформувати у відповідні завдання, що належать до податкової сфери.

У таблиці 3.9 наведено приклад, що містить дані транзакцій покупок у магазині ряду продуктів. Наш приклад для спрощення буде обмежений 10 транзакціями, що містять 13 видів продуктів.

Зокрема, з точки зору адміністрування податків можна сформулювати постановку завдання таким чином: побудувати асоціативні правила, які на основі результатів перевірок (зафіксованих

порушень законодавства) можуть сформувати типовий образ платника податків за ознаками порушень.

Якщо брати специфіку контрольно-перевірочної роботи, то як трансакцію можна буде розглядати результати перевірок, а як продуктів – зафіксовані порушення податкового законодавства. Асоціативне правило з точки зору адміністрування податків можна сформулювати так: пошук певної комбінації порушень податкового законодавства, фіксація яких у результатах перевірок впливає на ймовірність наявності інших (незафіксованих) порушень.

Візуальний аналіз прикладу показує, що всі чотири транзакції, у яких фігурує салат, також включають помідори, і що чотири із семи покупок, що містять помідори, також містять салат. Салат і помідори здебільшого купуються разом. Асоціативні правила дозволяють виявляти й кількісно описувати такі збіги.

Таблиця 3.9

№	Транзакції
1.	Сливи, салат, помідори
2.	Селера, цукерки
3.	Цукерки
4.	Яблука, морква, помідори, картопля, цукерки
5.	Яблука, апельсини, салат, цукерки, помідори
6.	Персики, апельсини, селера, помідори
7.	Квасоля, салат, помідори
8.	Апельсини, салат, морква, помідори, цукерки
9.	Яблука, банани, сливи, морква, помідори, цибуля, цукерки
10.	Яблука, картопля

Асоціативне правило складається із двох наборів предметів, що мають назву ***умова*** (antecedent) та ***наслідок*** (consequent) й записуються у вигляді. Таким чином, асоціативне правило формулюється у вигляді: «*Якщо* ***умова,*** *то* ***наслідок***». Умова може обмежуватися тільки одним предметом. Правила відображаються за допомогою стрілок, спрямованих від умови до наслідку, наприклад, *помідори → салат*.

3.4.2. Основні характеристики асоціативних правил

Основними **характеристиками**, що описують асоціативне правило, є ***підтримка*** (support) і ***вірогідність*** (confidence).

Якщо позначити базу даних транзакцій через D , а число транзакцій у цій базі N, то кожна транзакція di, i =1…N являє собою певний набір предметів. Позначимо підтримку правила через S, а вірогідність – через C. *Підтримка асоціативного правила* – це число транзакцій, які містять як умову, так і наслідок. Наприклад, для асоціації $A \to B$ можна записати

$$S(A \to B) = P(A \cap B) = \frac{n(\{A;B\} \in d_i)}{N}$$

Вірогідність асоціативного правила $A \to B$ являє собою міру точності правила й визначається як відношення кількості транзакцій, що містять умову і наслідок, до кількості транзакцій, що містять тільки умову

$$C(A \to B) = P(A \mid B) = \frac{n1(\{A;B\} \in d_i)}{n1(\{A\} \in d_i)}$$

Якщо підтримка й вірогідність досить високі, можна з великою ймовірністю стверджувати, що будь-яка майбутня транзакція, що включає умову, буде містити й наслідок. Для прикладу розглянемо правило *салат → помідори*, що напрошується з попередніх спостережень. Для нього:

S(*салат* →*помідори*)=4/10 =0,4;
C(*салат* →*помідори*) =4 / 4 =1.

Це правило, що зустрічається в 40 % транзакцій, є для вихідних даних абсолютно правильним – у всіх випадках, коли клієнт купує салат, він разом з тим купує й помідори. Це легко пояснити логічно – обидва продукти використовуються для готування овочевих блюд і дійсно часто купуються разом. Тепер розглянемо асоціацію *конфеты → помидоры,* у якій містяться мало сумісні в гастрономічному плані продукти.

Підтримка цієї асоціації S = 4/10 = 0,4 – та ж, що в попереднього правила, а вірогідність C = 4/6 = 0,67. Таким чином, порівняно невисока вірогідність даної асоціації дає привід засумніватися в тому, що вона є правилом.

Останнє зауваження при досить великій номенклатурі товарів призводить до необхідності подвоювати кількість розрахунків. На практиці аналітики можуть віддавати перевагу правилам, які мають високу підтримку (вище певного рівня, наприклад, 0,3) або високу вірогідність (не менше 0,8–0,85).

Вірогідініість 0,67 здається не такою уже й малою. Чому ж говоримо про «незначну вірогідність» цієї асоціації? Справа в тому, що помідори зустрічаються в 7 чеках з 10 (P(B)= 0,7). Прийнято, що всі правила з вірогідністю меншою ніж проста ймовірність випадання наслідку не повинні розглядатися, адже вони, по суті, випадкові. Для прийняття асоціативного правила його вірогідність повинна бути не меншою ніж імовірність наслідку.

Висування одночасних вимог щодо підтримки й вірогідності дозволяють значно пом’якшити критерії (підтримку до 0,1–0,15, вірогідність – до 0,67–0,75). Правила, для яких значення підтримки й вірогідності перевищують певні, задані користувачем пороги, називають ***сильними правилами*** (strong rules). Усі наведені вище числові значення – емпіричні.

Наприклад, у задачах виявлення шахрайських операцій значення підтримки може знижуватися й до 1 %, оскільки із шахрайством пов’язане порівняно невелике число транзакцій. Як видно з викладеного вище, знайти асоціації не важко. Але як визначити, чи є кожне окреме правило значимим, таким, що дозволяє не вгадувати, а прогнозувати наслідок залежно від умови?

Наприклад, у кожних 100 аваріях у Росії одним з автомобілів є ВАЗ. Але правило *аварія* $\rightarrow$ *ВАЗ* , як і зворотне йому, буде малозначимим, оскільки в деяких регіонах ВАЗ становить більше 80 % автомобільного парку. Аналогічно в жіночих бутиках, де,

крім одягу й білизни продаються шоколад і цукерки, вводити правило *білизна→ шоколад* не раціонально, адже, за статистикою, шоколад і цукерки в них купують 87 % відвідувачок, які при цьому можуть не купувати білизни. Крім об'єктивних оцінок (підтримки й вірогідності) кожного зі згенерованих правил, різні джерела радять використовувати деякі суб'єктивні оцінки.

Усі вони так чи інакше базуються на об'єктивних оцінках. ***Ліфт*** (від interest lift – підвищення інтересу) обчислюється в такий спосіб:

$$L(A \rightarrow B)=C(A \rightarrow B) / P(B).$$

Ліфт – це відношення частоти появи умови в транзакціях, які містять і умову, і наслідок, до частоти появи наслідку в цілому. Чим більше значення ліфта, тим частіше наслідок визначається умовою порівняно з випадками, коли умова відсутня. **Якщо ліфт дорівнює 1**, зв'язок відсутній, близькі ж до нуля значення свідчать про сильну зворотну залежність.

Для нашого прикладу в таблиці 3.9 візьмемо два правила з однаковою вірогідністю:

P(*салат*)=4/10 = 0,4. *C*(*помідори* →*салат*)=4/7 = 0,57.
P(*цукерки* 6/10 = 0,6. *C*(*помідори* →*цукерки*)=4/7 = 0,57.

Здавалося б, правила однаково достовірні. Після розрахунку ліфта все стає на свої місця:

L(*цукерки*→*салат*)= 0,57/0,4 = 1,425;
L(*помідоры* →*цукерки*)= 0,57/0,6 = 0,95.

Левередж (leverage – важіль, плече) – це різниця між спостережуваною частотою, з якої умова й наслідок з'являються спільно, та добутком частот появи умови й наслідку окремо

$$T(A \rightarrow B)= S(A \rightarrow B) – P(A)\,P(B).$$

Проте не потрібно вважати ліфт універсальним мірилом адекватності. Справа в тому, що правило з меншою підтримкою й більшим ліфтом може бути менш значимим ніж альтернативне правило з більшою підтримкою й меншим ліфтом. Це пов'язане з тим, що останнє застосовується для більшої кількості покупців.

Левередж дозволяє впоратися із ситуаціями, коли й підтримка, й ліфт у правил ідентичні, але їх значимість відрізняється. Наприклад, у нашому овочевому магазині у правила *салат* $\rightarrow$ *помідори*

$$C = 1;\ S = 0{,}7;\ L = 1/0{,}7 = 1{,}43.$$

Поліпшення (improvement) – це відношення частоти спостережуваних виконань правила до добутку частот появи умови й наслідку окремо.

$$I(A \rightarrow B) = \frac{S(A \rightarrow B)}{P(A) \cdot P(B)} \quad .$$

Фактично, поліпшення показує, у скільки разів розглянуте правило забезпечує правильний прогноз краще, ніж випадкове вгадування. Усі правила $I(A \rightarrow B) \leq 1$ не є значимими.

Морква, як показує таблиця 3.9, також продається тільки з помідорами, і зустрічається чотири рази, тому й у правила *морква* $\rightarrow$ *помідори*

$$C = 1;\ S = 0{,}7;\ L = 1/0{,}7 = 1{,}43.$$

А левередж у цих правил відрізняється на 30 %:

$T(\textit{морква} \rightarrow \textit{помідори}) = 0{,}3 - 0{,}3 * 0{,}7 = 0{,}09.$
$T(\textit{салат} \rightarrow \textit{помідори}) = 0{,}4 - 0{,}4 * 0{,}7 = 0{,}12.$

Таким чином, значимість другої асоціації більша ніж першої. Альтернативою *левередж* є *поліпшення*.

Такі міри, як ліфт, левередж і поліпшення можуть використовуватися для обмеження набору розглянутих асоціацій шляхом встановлення граничних значень значимості, нижче яких асоціації відкидаються.

3.4.3. Пошук асоціативних правил

У процесі пошуку асоціативних правил може виконуватися виявлення всіх асоціацій, підтримка й вірогідність для яких перевищують заданий мінімум. Найпростіший алгоритм пошуку асоціативних правил розглядає всі можливі комбінації умов і наслідків, оцінює для них підтримку й вірогідність, а потім виключає всі асоціації, які не задовольняють заданим обмеженням. Число можливих асоціацій зі збільшенням предметів зростає експоненційно. Якщо в базі даних транзакції присутні k предметів і всі асоціації є бінарними (тобто містять по одному предмету в умові й наслідку), потрібно буде проаналізувати $k \cdot 2k-1$ асоціацій. Оскільки реальні бази даних транзакцій, розглянуті під час аналізу ринкового кошика, звичайно, містять тисячі предметів, обчислювальні витрати при пошуку асоціативних правил величезні.

Наприклад, якщо аналізувати вибірку, що містить усього 100 предметів, то кількість асоціацій, утворених цими предметами, становитиме $100 \cdot 299 \approx 6{,}4 \cdot 1031$. Природно, що побудова такої кількості правил і перевірка вірогідності кожного з них – завдання нелегке. Тому для її вирішення використовуються деякі алгоритмічні прийоми, що дозволяють скорочувати простір можливих рішень. Однією з найпоширеніших є методика, заснована на *виявленні так званих частих наборів*, коли аналізуються тільки ті асоціації, які зустрічаються досить часто. На цій концепції заснований відомий алгоритм *пошуку асоціативних правил Apriori.*

Ієрархічні асоціативні правила

Якщо провадити пошук асоціативних правил серед окремих предметів (наприклад товарів у магазині), то можна виявити, що в багатьох випадках асоціації з високою підтримкою для окремих товарів практично відсутні. Особливо це характерно для супермаркетів, де асортимент товарів кожного виду досить великий. Наприклад, у продажу можуть бути десятки різновидів таких товарів, як *кетчуп* і *макарони.* Тому, незважаючи на те, що в цілому підтримка асоціації *макарони* → *кетчуп* може бути дуже високою, підтримка асоціацій між окремими видами цих товарів, швидше за все, буде низькою. Отже, такі асоціації хоча й можуть

становити інтерес, будуть вилучені з розгляду, оскільки не задовольнятимуть деякому мінімальному порогу підтримки Smin.

Для вирішення цієї проблеми при пошуку асоціативних правил розглядають не окремі предмети, а їх ієрархію. Якщо на нижніх ієрархічних рівнях подібні цікаві асоціації відсутні, то на більш високих вони можуть мати місце. Іншими словами, підтримка окремого предмета завжди буде меншою ніж підтримка групи, до якої він входить:

$$S(I) > S(i_j),$$

де *I* – група в ієрархії; i_j - *j* – той предмет, що входить у цю групу. Причини цього очевидні: загальна підтримка групи дорівнює сумі підтримки включених у неї предметів.

В ієрархічній схемі кетчуп і макаронні вироби мають множину підвидів. Найчастіше під час обробки касових чеків може бути отримана інформація лише про конкретний товар. Якщо ж структура бази даних транзакцій дозволяє відображати ієрархію товарів, можна досліджувати всі утворені ними ієрархічні рівні. Асоціативні правила, виявлені для предметів, розташованих на різних ієрархічних рівнях, отримали назву *ієрархічні асоціативні правила.*

У закордонній літературі вони також відомі як **багаторівневі правила** (multilevel rules) або **узагальнені правила** (generalized rules). Ієрархія предметів, пов'язаних з комп'ютерною технікою, одна з найбільш складних – вона містить 5 рівнів. Звичайно, ієрархічні рівні нумеруються згори донизу, починаючи з нульового. Вузол, що відповідає нульовому рівню, називають **кореневим** (root node). У комп'ютерній техніці перший рівень відповідає всім товарам, другий – категоріям комп'ютерного забезпечення, третій – конкретним предметам (товарам), четвертий – фірмі-виробникові, п'ятий – конкретній моделі того або іншого товару.

Наприклад, товар (1) – принтер (2) лазерний (3) фірми Xerox (4) модель FX3100 (5). Шукати навіть у величезній базі даних транзакцій його асоціацію з іншим товаром (1) – мишею (2)

бездротовою (3) фірми Genius (4) модель KW-09 (5) марно. Підтримка такого правила навряд чи перевалить за 1 %, та й то лише з урахуванням того, що ця конкретна модель миші або принтера мають велику абсолютну популярність. Універсальним методом для подібних випадків є послідовний спуск за рівнями ієрархії від першого до останнього з виписуванням усіх асоціацій на кожному рівні. Логічно, що на вищих рівнях буде високою підтримка (наприклад, правил «*ноутбук → миша*») і вірогідність.

Звертаючи увагу на більш низькі рівні абстракції, аналізують нащадки лише тих категорій і підкатегорій, які є частими наборами, тобто зустрічаються не менше наперед заданої кількості разів, де k – номер рівня. Як на недолік ієрархічного підходу до пошуку асоціативних правил іноді вказують на те, що отримані правила в більшості випадків стосуються не окремих предметів, а їхніх груп, а це не завжди відповідає вимогам аналізу. Однак у деяких предметних областях кількість окремих найменувань може бути настільки значною, що виявлення асоціативних правил навіть на високому рівні узагальнення – це вже успіх. Крім того, технологія бізнес-аналізу більшою мірою орієнтована на узагальнені даних, тому виробники й конкретні моделі рідко цікавлять дослідника під час побудови асоціацій. Існує кілька підходів до пошуку ієрархічних асоціативних правил.

Найчастіше використовуються спадні методи, коли часті предметні набори досліджуються на кожному ієрархічному рівні, починаючи з першого й закінчуючи рівнем з найбільшою деталізацією. Іншими словами, щойно виявляються всі часті предметні набори на першому рівні, починається пошук популярних предметних наборів на другому тощо. На кожному рівні для відкриття частих наборів може використовуватися будь-який алгоритм, наприклад Apriori та його модифікації.

3.4.4. Стратегії ведення пошуку правил

Відомі кілька стратегій ведення пошуку правил.

1. ***Використання однакового порога мінімальної підтримки*** $S\min k = const$ на всіх ієрархічних рівнях. При пошуку правил одноразово задається певний поріг мінімальної підтримки (наприклад 5 %), при досягненні якого набір вважається частим, у протилежному випадку – не включається до списку правил. Перевага підходу: висока швидкість аналізу предметної галузі, обумовлена відмовою від оцінки часткових наборів, отриманих з тих з них, які недостатньо часто зустрічаються. Недолік: ризик пропустити тонкі асоціації на нижніх рівнях ієрархії, а саме: перевагу користувачів одного виробника або зв'язку певних моделей товарів. Іноді цей недолік намагаються обійти за рахунок зниження мінімального рівня підтримки. Як наслідок, поява десятків і сотень незмістовних правил з низькою підтримкою й вірогідністю.

2. ***Зниження порога мінімальної підтримки*** під час переходу до нижніх рівнів ієрархії. Може здійснюватися індивідуально для кожної підгрупи (наприклад, S («*Комп'ютери*») = 0,1, тоді при розподілі комп'ютерів на категорії «настільний» та «портативний» встановлюється S («*Настільний*») = 0,06 й S («*Портативний*») = 0,04) або функціонально.

Функціональний вид зниження порога, як правило, пов'язаний з кількістю підкатегорій або порядковим номером рівня. Наприклад, якщо категорія «Комп'ютери» з рівнем підтримки $S^2 = 0{,}1$ ділиться на дві підкатегорії: «настільний» та «портативний», то в обох під категорій:

$$S^3 = S^2/2 = 0{,}05.$$

Інший варіант функціонального підходу полягає у встановленні порогів мінімальної підтримки винятково залежно від рівня, скільки б там підкатегорій не було $S^k\min = S^1\min / k$. При такому варіанті функціонального обмеження, якщо, наприклад, для всієї бази даних прийнятий мінімальний рівень підтримки $S^1\min$ 0,2, для категорії «комп'ютери» (2 рівень) він буде 0,2/2=0,1; «настільні комп'ютери» – 0,2/3=0,067; «настільні комп'ютери Dell» 0,2/4=0,05 тощо.

Перевага цього підходу очевидна – він дозволяє піти набагато далі по ієрархії в пошуку асоціацій. Однак у випадку з індивідуальним завданням граничного рівня для кожної підкатегорії, левову частину часу займатиме саме ця процедура. У випадку ж встановлення порогу залежно від рівня чи кількості підкатегорій не враховуються індивідуальні переваги й недоліки окремих виробників і моделей.

3. ***Міжрівнева фільтрація*** заснована на так званому рівні проходу (level passage). Відносно високий для верхніх рівнів ієрархії граничний рівень підтримки зберігається на нижніх у першому проході. Потім у кожному проході цей рівень знижується. Ті товари низького рівня, які мають асоціації, таким чином, ідентифікуються раніше, ніж їхні батьківські категорії, які, можливо, і не будуть мати необхідної підтримки. Рекомендується виконувати три–чотири проходи бази даних транзакцій для сполучення рівнів довіри представників різних рівнів.

Послідовні шаблони

Усе викладене раніше не дає відповіді на питання, як пов'язати асоціації із клієнтом, а також встановити часові залежності, тобто не тільки відповісти на запитання, що купує певний клієнт, а що за чим він купує. І якщо є можливість – спрогнозувати, що ж він купить наступного разу. Відповіді на це питання дає використання **послідовних шаблонів** (sequential patterns), заснованих на теорії асоціацій, обов'язковими полями яких є дата/час та ідентифікатор клієнта. Останні можуть бути отримані під час розрахунку кредитними картами, клієнтськими картками або будь-якими іншими ідентифікаторами – законними або не дуже.

Під час розгляду послідовностей транзакцій використовується одне припущення – той самий клієнт у той самий момент не здійснює двох різних транзакцій.

Послідовність – це впорядкований список предметних наборів (товарів або послуг, замовлених або оплачених ідентифікованим клієнтом в ідентифікований момент часу). Послідовність *S* називається *максимальною*, якщо вона не міститься у будь-якій іншій послідовності. Послідовність *S* називається *клієнтською*,

якщо вона, крім набору предметів, дати й часу, містить також ідентифікатор клієнта. Послідовність $S1$ міститься у послідовності $S2$, якщо всі предметні набори $S1$ містяться у предметних наборах $S2$.

Наприклад, послідовність <(3); (4, 5); (8)> міститься в послідовності <(7); (3, 8); (9); (4, 5, 6); (8)>, оскільки (3) $\subseteq$ (3, 8), (4, 5) $\subseteq$ (4, 5, 6) і (8) $\subseteq$ (8). Однак <(3); (5)> $\not\subset$ <(3, 5)>) і навпаки, оскільки в першій послідовності предмети 3 й 5 були куплені один за іншим, а в другій – разом.

Послідовність S називається *підтримуваною клієнтом,* якщо вона міститься в його клієнтській послідовності. Тоді підтримка послідовності визначається як *число клієнтів, що її підтримують,* та, зазвичай, виражається у відсотках від загального числа клієнтів. Таким чином, поняття підтримки для послідовних шаблонів дещо відрізняється від аналогічного поняття для асоціативних правил.

Для бази даних клієнтських транзакцій завдання пошуку послідовних шаблонів полягає у виявленні максимальних послідовностей серед усіх, що мають підтримку вище заданого порога. Кожна така максимальна послідовність і буде послідовним шаблоном. Далі будемо називати послідовності, що задовольняють обмеження мінімальної підтримки, *частими* (за аналогією із наборами, що часто зустрічаються, у теорії асоціативних правил).

Процес пошуку послідовних шаблонів складається з таких етапів:

1. **Сортування**. Транзакції вихідної бази даних сортуються за кодом клієнтів, а транзакції кожного клієнта – за датою, часом і номером візиту. Результат – база даних клієнтських послідовностей.

2. **Пошук частих предметних наборів** F. Частими називають предметні набори, які купувалися кількістю клієнтів, більшою від мінімально-граничного значення. Обрана множина частих предметних наборів переводиться до числового або символьного представлення.

3. **Перетворення**. Необхідно визначити, які з частих послідовностей містяться у клієнтській послідовності. Для цього кожна транзакція клієнтської послідовності заміщується множиною її частих предметних наборів. Якщо в транзакції немає жодного частого набору, вона більше не розглядається. Більш того, якщо в певного клієнта в послідовності немає жодного частого набору, він теж виключається з розгляду. Після перетворення кожна клієнтська послідовність – це впорядкована множина частих наборів.

4. **Пошук частих послідовностей**. На множині частих предметних наборів виконується пошук частих послідовностей. Мінімальний рівень частоти – параметр алгоритму.

5. **Пошук максимальних послідовностей**. Серед частих послідовностей відшукуються максимальні. Іноді цей етап об'єднують із попереднім аби зменшити витрати часу на обчислення немаксимальних послідовностей. Найбільш проблемним етапом пошуку послідовних шаблонів є виявлення частих послідовностей, оскільки велика кількість предметних наборів вимагає розгляду величезної кількості можливих комбінацій і багаторазового проходження за набором трансакцій. Кожне проходження починається з вихідного набору послідовностей, що використовується для генерації нових потенційних частих послідовностей, які мають назву *послідовності-кандидати,* або просто *кандидати.*

Для цього обчислюється їх підтримка та після завершення проходження визначається, чи є виявлені кандидати в дійсності частими. Виявлені часті послідовності кандидатів стануть вихідними для нового проходження.

Питання для самоконтролю до теми 3.4

1. Що означає термін «аффінітивний аналіз» (affinity analysis) – один з поширених методів Data Mining?

2. Що означає типова транзакція при асоціативному аналізі взаємодії з торговельного підприємства із клієнтом?

3. Опишіть поняття асоціативних правил. Призначення, можливі використання в аналізі податкової інформації.

4. Опишіть задачі, для вирішення яких можна застосовувати асоціативні правила в адмініструванні податків.

5. Опишіть основні характеристики, що описують асоціативне правило: *підтримка* (support) і *вірогідність* (confidence).

6. Дайте визначення показника «ліфт», що застосовується при побудові асоціативних правил.

7. Дайте визначення показника «левередж», що застосовується при побудові асоціативних правил.

8. Опишіть стратегії пошуку правил: використання однакового порога мінімальної підтримки; *зниження порога мінімальної підтримки.*

9. Вкажіть правильний варіант. Різниця між спостережуваною частотою, з якою умова й наслідок з'являються спільно, та добутком частот появи умови й наслідку окремо є: левередж; ліфт; поліпшення.

10. Вкажіть правильний варіант. Відношення частоти спостережуваних виконань правила до добутку частот появи умови й наслідку окремо є: левередж; ліфт; поліпшення.

11. Вкажіть правильний варіант. Відношення частоти появи умови в транзакціях, які містять й умову, і наслідок, до частоти появи наслідку в цілому є: левередж; ліфт; поліпшення.

3.5. Застосування методології дерев рішень у класифікаційних завданнях

3.5.1. Основні поняття методології дерев рішень

Дерева рішень (decision trees) належать до найбільш популярних і потужних інструментів Data Mining, що дозволяють ефективно вирішувати завдання класифікації і регресії. На відміну від методів, які використовують статистичний підхід, таких як класифікатор Байєса, лінійна і логістична регресія, дерева рішень засновані на машинному навчанні і здебільшого не вимагають припущень про статистичний розподіл значень ознак.

В основі дерев рішень лежать вирішальні правила виду «якщо ... то ...», які можуть бути сформульовані на природній мові. Тому дерева рішень є найбільш наочними і легко інтерпретуються моделями.

Дерева рішень – ієрархічні деревоподібні структури, що складаються з вирішальних правил виду «якщо ... то ...» і дозволяють виконувати класифікацію об'єктів. У дереві кожному об'єкту відповідає єдиний вузол, що дає рішення. Термін «дерево рішень» використовується не тільки в Data Mining, а й в суміжних областях аналізу й обробки даних, штучного інтелекту. Наприклад, у завданнях алгоритмізації *дерево рішень* – це спосіб представлення процесу прийняття рішення, що має вигляд відповідей на ряд питань, що утворюють деревоподібну структуру, кожна кінцева вершина якої становить елементарне рішення. У менеджменті під деревом рішень розуміють графічне зображення альтернативних дій і їх наслідків. Однак між усіма цими тлумаченнями терміна багато спільного.

Для зручності наведемо базові поняття теорії дерев рішень у таблиці 3.10.

Таблиця 3.10

Назва	Опис
Об’єкт	Приклад, шаблон, спостереження, запис
Атрибут	Ознака, незалежна змінна, властивість, вихідне поле
Мітка класу	Залежна змінна, цільова змінна, вихідне поле
Вузол	Внутрішній вузол дерева
Лист	Кінцевий вузол дерева, вузол рішення
Перевірка	Умова вузла

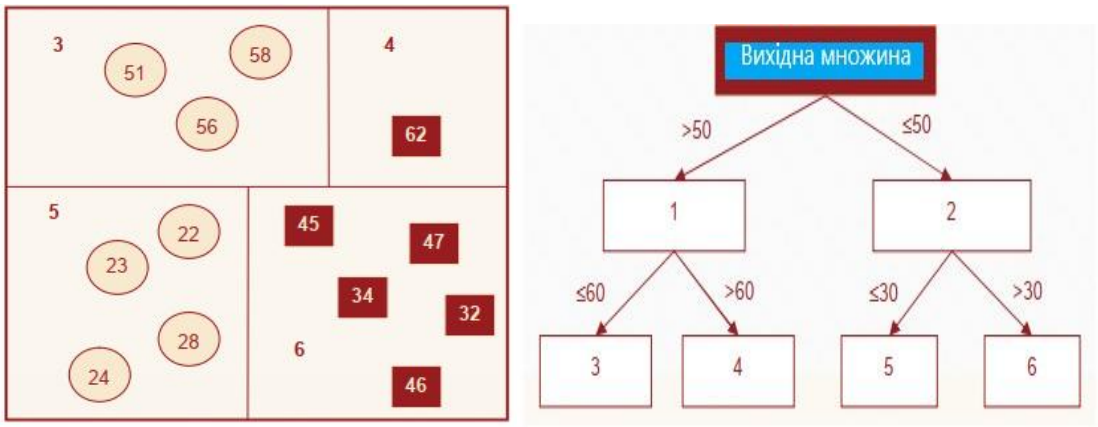
Розглянемо основну ідею алгоритмів побудови дерев рішень на прикладі.

Нехай потрібно передбачити повернення або неповернення кредиту за допомогою набору вирішальних правил на основі єдиного атрибута «Вік клієнта». Для цього будемо використовувати множину спостережень, в кожному з яких вказується вік, а також факт повернення / неповернення позики. Графічно таку множину спостережень подано на рис. 3.25 А.

Умовно приймемо, що об’єкт у формі кола вказує на неповернення, у формі прямокутника – на дефолт по займу, а всередині кожного об’єкта зазначений вік. Необхідно розбити множину об’єктів на підмножини таким чином, щоб в кожне з них потрапили об’єкти тільки одного класу.

Як можна побачити на рис. 3.25 Б, структура дерев рішень проста і в цілому аналогічна деревовидним ієрархічним структурам, що використовуються в інших сферах Data Mining, наприклад деревах асоціативних правил. До складу дерев рішень входять два види об’єктів – вузли (node) і листя (leaf). У вузлах містяться правила, за допомогою яких проводиться перевірка атрибутів і безліч об’єктів у цьому вузлі розбивається на підмножини. Листя – це кінцеві вузли дерева, в яких містяться підмножини, асоційовані з класами. Основною відмінністю листа від вузла є те, що в листі не проводиться перевірка, що розбиває асоційовану з ним підмножину і, відповідно, немає розгалуження. Листом може бути оголошений будь-який вузол, якщо прийнято рішення, що множина у вузлі досить однорідна у плані класової приналежності об’єктів і подальше розбиття не має сенсу, оскільки не приведе до істотного збільшення точності класифікації, а тільки ускладнить дерево.

У дереві рис. 3.26 Б, що класифікує об’єкти, подані на рис. 3.26 А, і об’єкти з номерами 1 і 2 – вузли, а з номерами 3, 4, 5 і 6 – листя. Звернемо увагу на те, що в дереві є по два листи, асоційованих з одним класом. У цьому немає ніякого протиріччя, просто для класифікації об’єктів у них використовувалися різні способи перевірки. Для кожного аркуша в дереві є унікальний шлях. Початковий вузол дерева є вхідним: через нього проходять усі об’єкти. Зазвичай, вхідний вузол називають кореневим вузлом (root node).



А Б

Рис. 3.26. Множина об’єктів і дерево рішень, що їх класифікує

Отже, дерево росте зверху вниз. Вузли і листя, підлеглі вузлу вищого ієрархічного рівня, *називаються нащадками*, або *дочірніми вузлами*, а той вузол відносно них – *предком,* або *батьківським вузлом*.

3.5.2. Алгоритми дерев рішень ID3 і C4.5

Одними з найбільш популярних алгоритмів побудови дерев рішень є алгоритм ID3 і його модифікація C4.5. Алгоритм ID3 починає роботу з усіма навчальними прикладами в кореневому вузлі дерева. Для поділу безлічі прикладів кореневого вузла вибирається один з атрибутів, і для кожного значення, прийнятого цим атрибутом, будується гілка і створюється дочірній вузол. Потім усі приклади розподіляються за дочірніми вузлами відповідно до значення атрибута. Пояснимо це на рис. 3.27. Нехай атрибут X

приймає три значення: A, B і C. Тоді під час розбиття вихідного безлічі T по атрибуту X алгоритм створить три дочірніх вузли T1 (A), T2 (B) і T3 (C), у перший з яких будуть розміщені всі записи зі значенням A, в другій – зі значенням B, а в третій – із значенням C.

Алгоритм повторюється рекурсивно доти, поки у вузлах не залишаться тільки приклади одного класу, після чого вузли будуть оголошені листами і розбиття припиниться. Найбільш проблемним етапом алгоритму є вибір атрибута, за яким буде проводитися розбиття у кожному вузлі.

Для вибору атрибута розбиття ID3 використовує критерій, званий приростом інформації (information gain) або зменшенням ентропії (entropy reduction).

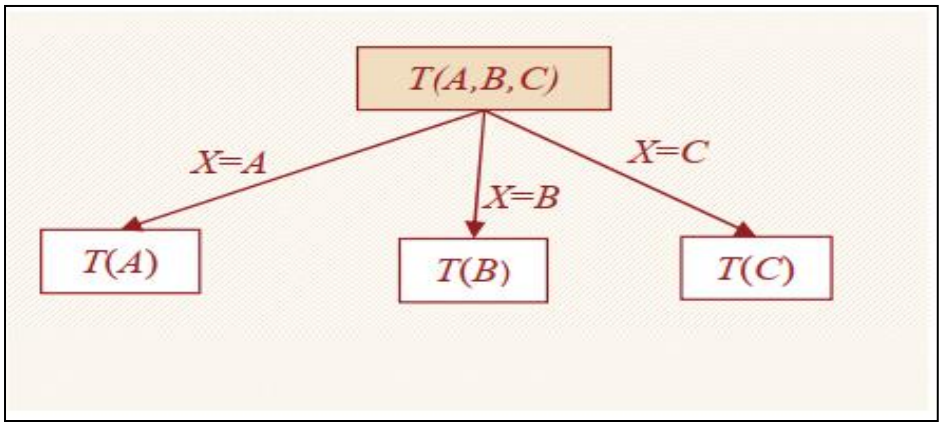


Рис. 3.27. Початкові вузли класифікаційного дерева

Введемо у розгляд *міру приросту інформації*, яка обчислюється як

$$\text{Gain (S)} = \text{Info (T)} - \text{InfoS (T)}, \qquad (3.31)$$

де *Info (T)* – ентропія множини T до розбиття;

InfoS (T) – ентропія після розбиття *S*.

Ця міра є приростом кількості інформації, отриманої в результаті поділу безлічі T на підмножини T1, T2, ..., Tk за допомогою розбиття S. Як найкращого атрибута для використання у разі необхідності розділення S вибирається той атрибут, який забезпечує найбільший приріст інформації Gain (S).

Розглянемо приклад. Нехай дано набір даних T, який містить 14 записів і 3 вхідних атрибута A1, A2, A3, а також вихідний атрибут (мітку класу) C, який приймає два значення – C1 і C2 (табл. 3.11).

Таблиця 3.11

№ з/п	A_1	A_2	A_3	C
1.	A	70	Да	C1
2.	A	90	Да	C2
3.	A	85	Ні	C2
4.	A	95	Ні	C2
5.	A	70	Ні	C1
6.	B	90	Да	C1
7.	B	78	Ні	C1
8.	B	65	Да	C1
9.	B	75	Ні	C1
10.	C	80	Да	C2
11.	C	70	Да	C2
12.	C	80	Ні	C1
13.	C	80	Ні	C1
14.	C	96	Ні	C1

У цьому прикладі 9 записів належить до класу C1, а 5 – до класу C2. Ентропія множини T буде:

Info (T) = – (9 14) × log2 (9 14) – (5 14) × log2 (5 14) = 0,94 біт.

Після використання атрибута A1 для розбиття вихідної множини на три підмножини: T1 (A), T2 (B) і T3 (C) – загальна ентропія розбиття складе (умовно вважаємо, що log2 (0) = 0):

InfoS1 (T) = 5 14 (– 2 5 × log2 (2 5) – 3 5log2 (3 5)) + 4 14 (– 4 4 × log 2 (4 4) – 0 4 × log2 (0 4)) + 5 14 (– 3 5 × log2 (3 5) – 2 5 × log 2 (2 5)) = 0,694 біт.

Таким чином, приріст інформації при використанні атрибута A1 для розбиття вихідної підмножини становитиме:

Gain (S1) = 0,94 – 0,694 = 0,246 біт.

При використанні розбиття на основі атрибута A3, провівши аналогічні обчислення, отримаємо:

ES3 (T) = 6 14 (– 3 6 × log2 (3 6) – 3 6log2 (3 6)) + 8 14 (– 6 8 × log2 (6 8) – 2 8 × log2 (2 8)) = 0,892 біт.

Тоді приріст інформації, який забезпечується розбивкою вихідної безлічі на основі атрибута A3, буде: *Gain (S3) = 0,94 – 0,892 = 0,048 біт*.

Отже, для початкового поділу множини T алгоритм вибере розбиття S1 з використанням атрибута A1, що забезпечує найбільший приріст інформації. Для знаходження оптимальної перевірки необхідно також розглянути атрибут A2. Раніше розглядали розбиття за категоріальними атрибутами. У цьому прикладі зіткнулися з необхідністю перевірки за числовим атрибутом A2. Перевірки за такими атрибутами складніші, оскільки потрібна установка порогів для розбиття, яка в загальному випадку може бути довільною. Існує алгоритм для визначення оптимального значення порога z. Спочатку навчальні приклади упорядковано в тому порядку, в якому будуть розглядатися значення атрибута: (v_1, v_2, ... v_m). Тоді існує тільки m – 1 порогів розбиття так, щоб кожне з них було перевірено на предмет максимального приросту інформації. При цьому як поріг можуть вибиратися серединні точки між сусідніми значеннями (v_i, v_{i+1})/2 або, як варіант, найменші значення з кожного інтервалу (v_i, v_{i+1}).

Для ілюстрації цього способу знаходження порога візьмемо числовий атрибут A2. Після сортування прикладів за зростанням значень атрибута отримаємо такий порядок {65, 70, 75, 78, 80, 85, 90, 95, 96} і безліч можливих порогових значень z = {65, 70, 75, 78, 80, 85, 90, 95}. З дев'яти значень оптимальним буде те, яке забезпечує найбільший приріст інформації. Якщо для кожного потенційного порога визначити приріст інформації подібно до того, як раніше це робилося для значень категоріальної змінної, то можна виявити, що кращим значенням порога буде z = 80. Обчислення приросту інформації для розбиття S2 ($A_2 \leq 80$ або $A_2 \geq 80$) проводиться таким чином:

InfoS2 (T) = 9 14 (– 79 × log 2 (79) – 2 9log 2 (29)) + 5 14 (– 2 5 × log 2 (2 5) – 35 × log 2 (35)) = 0,837 біт;

Gain (X 3) = 0,94 – 0,837 = 0,103 біт.

Тепер можна порівняти приріст інформації для всіх трьох атрибутів ілюстративного прикладу. Бачимо, що атрибут A_1 як і раніше забезпечує найбільший приріст інформації – 0,246, і отже, він повинен бути обраний для використання в першому розбитті

дерева рішень. Кореневий вузол перевірить для атрибута A_1, в результаті чого будуть створені три гілки – по одній для кожного значення атрибута (рис. 3.28).

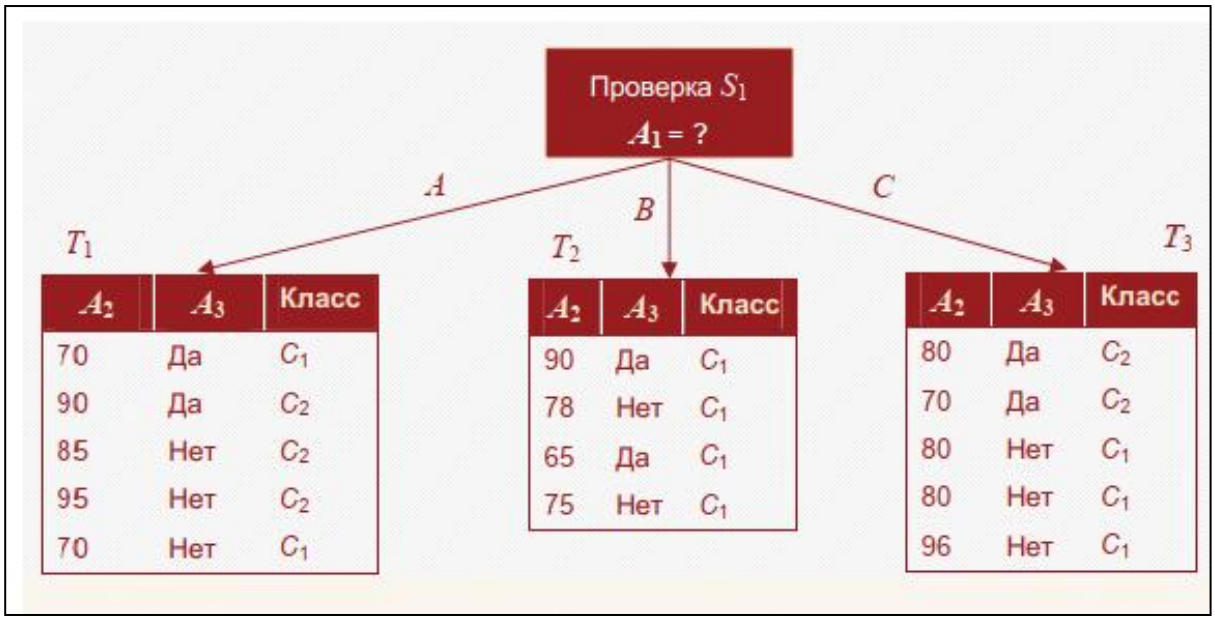


Рис. 3.28. Три гілки класифікаційного дерева (дані – табл. 3.11)

Після початкового розбиття всі вузли-нащадки будуть містити по кілька спостережень з вихідної множини, і для кожного вузла потім буде повторений процес вибору атрибута розбиття і оптимізації порогу (якщо для розбиття використовується числовий атрибут).

Оскільки вузол-нащадок T_2, отриманий з гілки для значення B, містить 4 спостереження, які належать до одного класу, то ентропія дорівнює 0, вузол оголошується листом, і подальше розгалуження для нього не проводиться.

Для вузла T_1, що включає 5 спостережень, може бути зроблена перевірка за знайденими атрибутами. Опускаючи проміжні розрахунки, констатуємо, що оптимальне розбиття буде досягнуто за допомогою атрибута A_2 для альтернативних варіантів $A_2 \leq 70$ або $A_2 > 70$.

Оскільки безліч T_1 містить 5 спостережень, з яких два належать до класу C_1, а три – до класу C_2, то, відповідно, до формули ентропії можна записати:

InfoS2 (T1) = - 2 5 × log 2 (2 5) - 3 5log2 (3 5) = 0,97 біт.

Під час використання атрибута A2 для розбиття T1 з порогом 70 отримаємо чисті вузли, тому

InfoS2 (T1) = 0.

Приріст інформації, забезпечений цим розбиттям, становитиме:

Gain (X 3) = 0,94 – 0 = 0,94 біта і буде максимальним.

Аналогічним чином можуть бути виконані розрахунки і для підмножини T_3. У ньому оптимальна перевірка за значеннями атрибута A3. Отже, будуть створені дві нові гілки, що відповідають значенням A3 = Та й A3 = Немає. У результаті отримаємо дві однорідних підмножини, що містять спостереження одного класу. Результуюче дерево подано на рис. 3.29.

Практичне застосування класичного алгоритму ID3 пов'язане з рядом проблем, характерних для моделей, заснованих на навчанні взагалі і дерев рішень зокрема. Основними з них є перенавчання та наявність пропусків у даних. Для їх ефективного подолання ID3 був доопрацьований, в результаті чого з'явилося його розширення, назване C4.5.

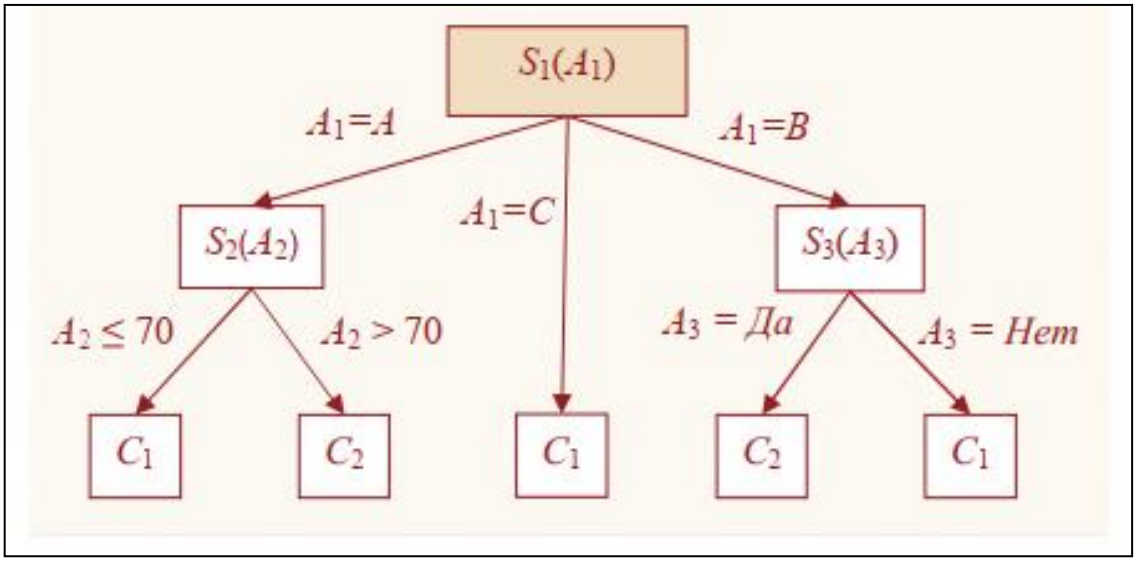


Рис. 3.29. Результуюче дерево класифікації

Детальніше метод C4.5 розглядається на лабораторному практикумі.

3.5.3. Застосування класифікаційних методів Oracle Data Mining

Як уже зазначалося Oracle Data Miner є опцією Oracle Advanced Analytics, дозволяє легко створювати і розгортати додатки для інтелектуального аналізу, які допомагають отримувати від інформаційних масивів нові знання для будь-якої галузі. Використовуючи можливості Database 11g/12/18, Oracle пропонує середовище Oracle R Enterprise для корпоративного класу додатків, розроблених за допомогою мови статистичних досліджень і графічної візуалізації R, а також на кілька порядків збільшує обсяги даних, з якими можуть працювати ці додатки. Oracle Advanced Analytics в поєднанні з оптимізованими програмно-апаратними комплексами Oracle Big Data Appliance, Exadata Database Machine і Exalytics In-Memory Machine надає найповнішу й інтегровану в галузі платформу для аналізу «великих даних». З огляду на функціонування інформаційної системи ДФС України на платформі Oracle важливим аспектом концепції Oracle Advanced Analytics є переміщення алгоритмів аналітики до даних, що зберігаються в Oracle Database 11/12g і Oracle Exadata.

Hadoop – це проект з відкритим вихідним кодом, що знаходиться під управлінням *Apache Software Foundation* (http://hadoop.apache.org/). Hadoop використовується для надійних, масштабних і розподілених обчислень, але може також застосовуватися і як сховище файлів загального призначення, здатних вміщувати петабайти даних. Багато компаній використовують Hadoop в дослідних і виробничих цілях. Hadoop складається з двох ключових компонентів: распределена файлова система Hadoop (англ. HDFS), яка відповідає за зберігання даних на кластері Hadoop; система *MapReduce* призначена для обчислення і обробки великих обсягів даних на кластері. На основі цих ключових компонентів створено кілька підпроектів, таких як Pig, Hive, HBase тощо.

Традиційний підхід, як відомо, заснований на переміщенні даних до алгоритмів, вимагає додаткових витрат на вилучення даних зі сховищ, їх завантаження в ПЕОМ користувача-аналітика або на спеціалізовані сервери. На основі Oracle Advanced Analytics можна створювати аналітичні додатки в режимі реального часу, які допомагають отримувати відповіді на ключові питання бізнесу.

На рис. 3.30 показано SQL Developer як клієнтський додаток, який виконує SQL і PL / SQL на базі даних Oracle. Oracle Data Miner – один з інтегрованих компонентів в Oracle SQL Developer. Репозиторій ODMRSYS зберігає метадані та API в БД, які підтримують операції та адміністрування Oracle Data Miner. Репозиторій Oracle Data Miner підтримує декілька користувачів, використовує служби Oracle Data Mining, Oracle XML DB, Oracle Scheduler та Oracle Text в Oracle Database Enterprise Edition. Oracle Exadata має інтегровані засоби Oracle Big Data SQL, Big Data Connectors для використання потужних можливостей Oracle Big Data Appliance, HDFS, HIVE, NoSQL.

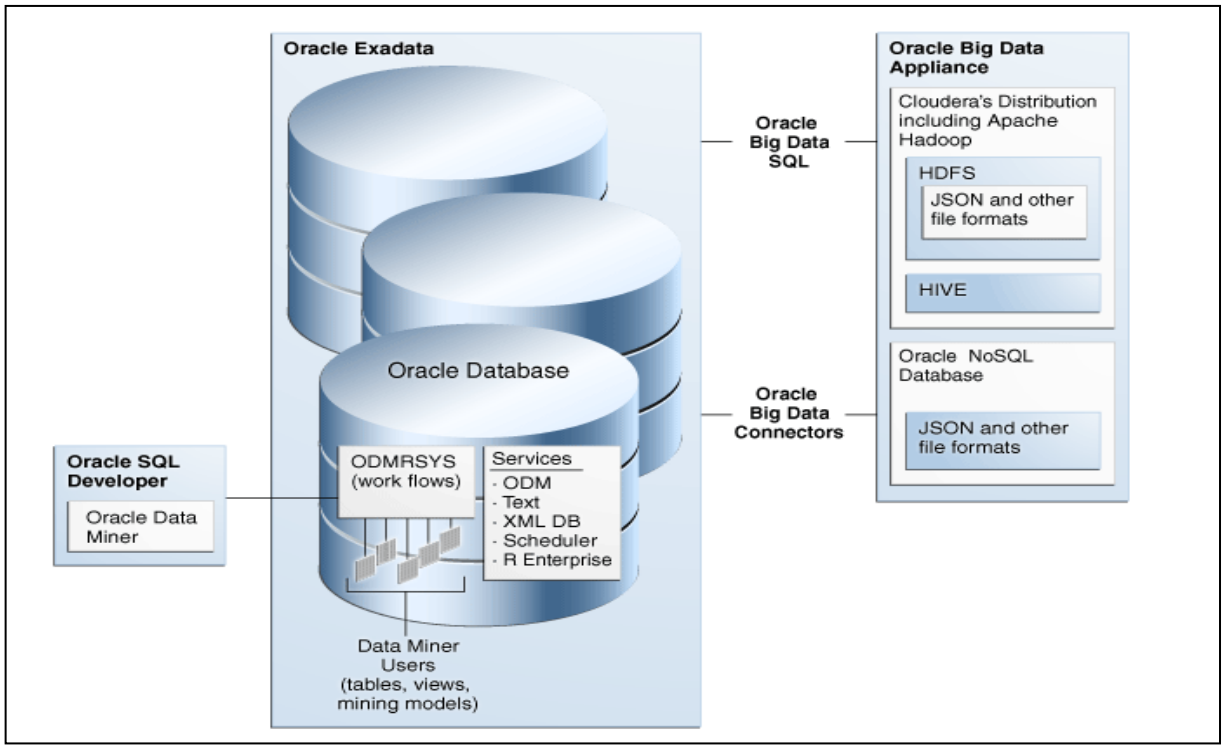


Рис. 3.30. Oracle Data Miner як компонент платформи Oracle Database

Джерело: http://www.oracle.com

JSON (англ. JavaScript Object Notation) – текстовий формат обміну даними, заснований на JavaScript. Як і багато інших текстових форм, JSON легко читається людьми. Формат JSON був розроблений Дугласом Крокфордом.

Oracle Data Miner підтримує наступні алгоритми (рис. 3.30):

1. **Класифікація**: Naïve Bayes, Decision Trees, Generalized Linear Models (GLM), Support Vector Machine; k-Means.
2. **Регресія:** Support Vector Machine, Generalized Linear Models (GLM).
3. **Пошук основних атрибутів:** Minimal Descriptor Length.
4. **Кластеризація:** Enhanced K-means, O-cluster, Orthogonal Partitioning Clustering.
5. **Пошук асоціацій:** Apriory, Association.
6. **Виділення ознак:** Non-Negative Matrix Factorization.



Рис. 3.31. Набір класів модулів інтелектуального аналізу Oracle Data Miner

Важлива особливість алгоритмів полягає у тому, що всі вони працюють безпосередньо в реляційній базі даних Oracle і не вимагають вивантаження і збереження даних у різних форматах. Зокрема, актуальними для інформаційно-аналітичної системи ДФС України є класифікаційні алгоритми, параметри яких можна динамічно змінювати, легко переналагоджувати.

(ODM) – працює безпосередньо на базі даних, уникаючи проблем, пов'язаних з переміщенням, дублюванням і захищенням даних, а також масштабністю інструментів. Більше того, ODM спочатку був розроблений таким чином, щоб добре узгоджувався з SQL. Інші рішення потребують використання складних, громіздких інструментів і операцій під час пошуку закономірностей у схемі типу «зірка». Необхідно використовувати інструменти інтелектуального аналізу даних, необхідних для переведення даних з транзакційної форми у плоску, денормалізовану. ODM, розроблений таким чином, щоб вільно виконувати пошук закономірностей у схемі типу «зірка». Такий підхід робить його роботу простішою, дає в руки потужні інструменти і підвищує продуктивність.

Розглянемо приклад застосування опції Oracle Data Mining як надбудови Oracle SQL Developer для класифікації підприємств за ознаками збитковості (файл даних – Лабораторна 2.3) і порівняємо саме методи **класифікації**: Naïve Bayes, Decision Trees, Generalized Linear Models (GLM), Support Vector Machine.

Як навчальний масив були використані знеособлені дані фінансового стану 152 підприємств, що були класифіковані за станом платоспроможності (збикове/незбикове підприємство) через 12 місяців після розрахунку фінансових коефіцієнтів.

Таблиця 3.12

Статистика вибірки 152 підприємств для побудови класифікаційних моделей

№	Показники	На кінець року
1.	**Підприємства, які будуть збитковими.** Кількість підприємств, що були і залишилися збитковими, або були прибутковими і стали збитковими (у масиві даних – **категорія «2»)**	69
2.	**Підприємства, що будуть прибутковими**. Кількість підприємств, що були збитковими і стали прибутковими, або були і залишилися прибутковими (у масиві даних – **категорія «4»)**	83

Фінансові коефіцієнти, які використувалися для побудови класифікаційних моделей: скоригований валовий дохід (SVD); коеф. рентабельності витрат (KOVV); коеф. рентабельності діяльності (KOVD); коеф. економічної ефективності (KR3); коеф. покриття (KP3); коеф. загальної ліквідності (KZL); коеф. абсолютної ліквідності (KAL); коеф. оборотності обігових коштів (KOOK); коеф. забезпечення власними оборотними коштами (KZVOK); коеф. оборотності дебіторської заборгованості (KODZ); коеф. оборотності кредиторської заборгованості (KOKZ); коеф. фінансової незалежності (KFN); коеф. фінансової стабільності (KFS).

На рис. 3.32 показано робочий простір моделювання у середовищі Oracle SQL Developer . Моделі створюються за допомогою конструктора, де відбираються потрібні дані, підключаються потрібні модулі і проводиться навчання методів.

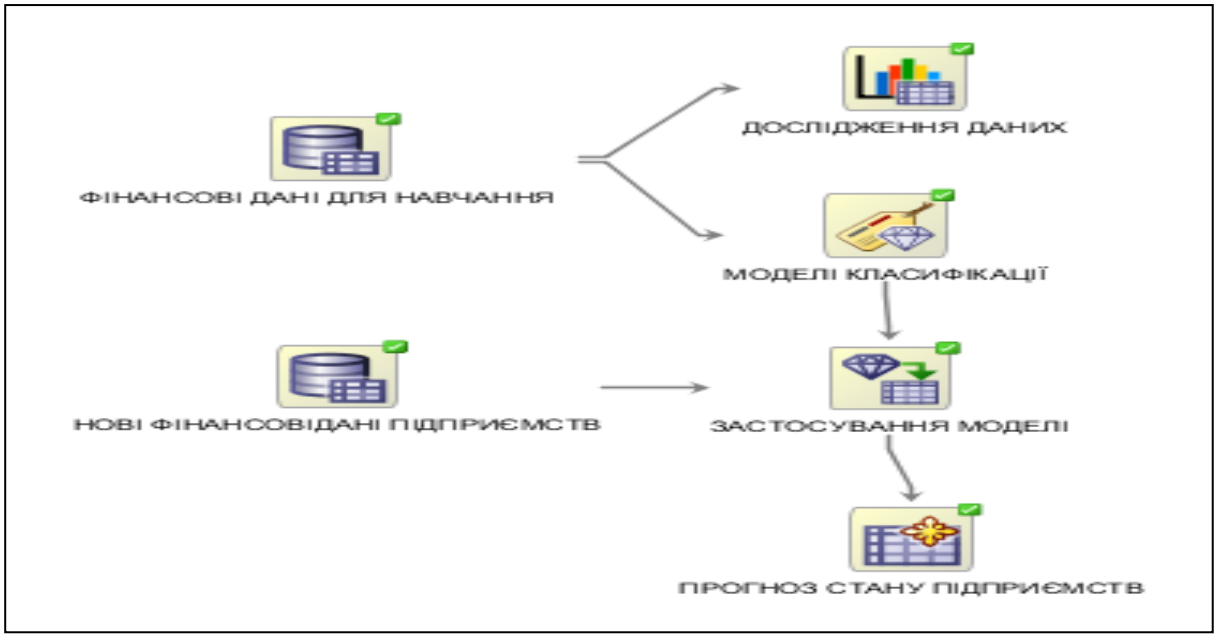
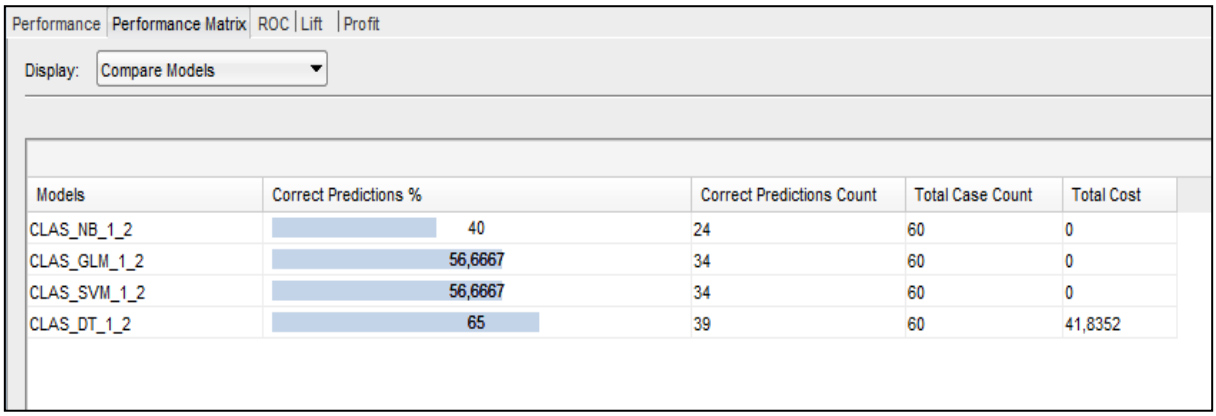


Рис. 3.32. Робочий простір Oracle Data Miner із сформованими класифікаційними моделями у середовищі Oracle SQL Developer

Порівняльний аналіз якості моделей у середовищі Oracle Data Miner показаний на рис. 3.33. Кращу якість класифікації показує алгоритм *дерева рішень* «CLAS_DTY_1_2». Для прикладу наведено одне із отриманих правил, яке свідчить про те, що підприємства, у яких коефіцієнт загальної ліквідності має значення <= 0,57, через рік можуть стати збитковими або залишитися збитковими із ступенем довіри 0,75.



Models	Correct Predictions %	Correct Predictions Count	Total Case Count	Total Cost
CLAS_NB_1_2	40	24	60	0
CLAS_GLM_1_2	56,6667	34	60	0
CLAS_SVM_1_2	56,6667	34	60	0
CLAS_DT_1_2	65	39	60	41,8352

Рис. 3.33. Порівняльний аналіз якості класифікаційних моделей у середовищі Oracle Data Miner

Враховуючи інтеграцію Oracle Data Miner із мовою R, вбудованість у інтерфейс Oracle SQL Developer, можливість генерації скриптів на SQL/PL SQL, цей компонент може бути ефективним засобом розширення функцій інформаційно-аналітичної системи ДФС України.

```
If   KZL3 <= ,568155

Then 2
```

Confidence	0.75
Support	0.4444444444444444

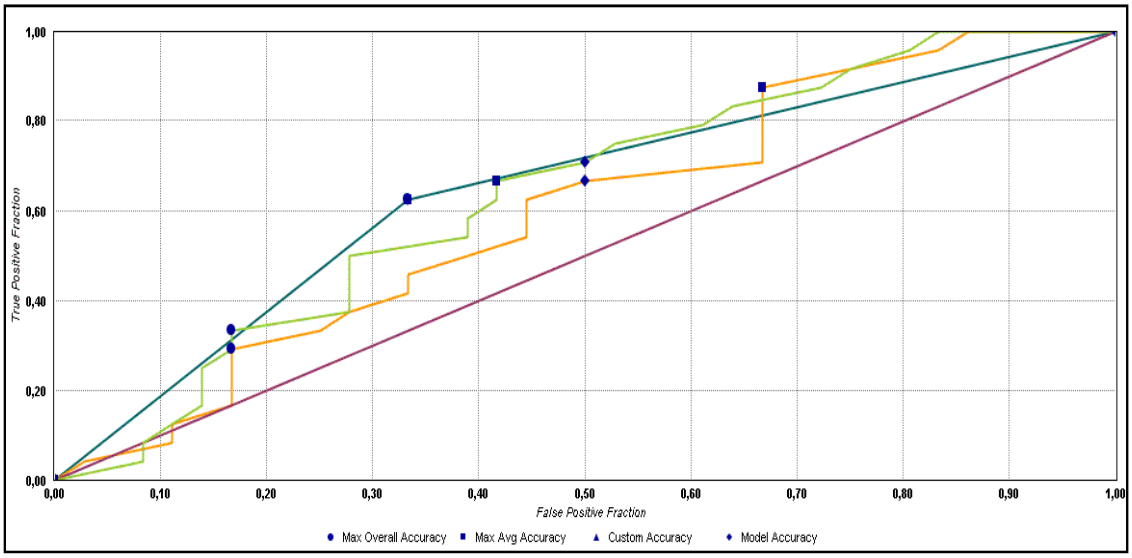


Рис. 3.34. Отримане класифікаційне правило за допомогою алгоритму дерева рішень та ROC крива

У цілому, Oracle Data Miner є простим та логічним у використанні. Наявність вбудованого робочого простору (Workflow) дає можливість виконувати досить складну аналітичну обробку даних фактично без програмування. Недоліками є відсутність гнучкого інструментарію для перетворення вхідних даних.

Питання для самоконтролю до теми 3.5

1. Які задачі вирішують методи дерев рішень. У чому їх відмінність від статистичних методів?

2. Дайте визначення методу дерева рішень.

3. Приведіть у прикладі основні поняття теорії дерев рішень.

4. Опишіть суть алгоритмів дерев рішень ID3 и C4.5.

5. Що означає приріст інформації під час використання атрибута?

6. Особливості застосування класифікаційних методів дерев рішень в Oracle Data Mining.

7. Наведіть приклад обчислення міри приросту інформації.

8. Опишіть особливості і призначення технології Hadoop.

9. Які класифікаційні алгоритми підтримує Oracle Data Miner?

10. Які класи моделей інтелектуального аналізу вміщує Oracle Data Miner?

11. Опишіть основні компоненти Oracle Dig Data Applience.

12. Опишіть компоненти Cloudera’s Distribution (HDFS, HIVE).

13. Опишіть особливості використання NoSQL, JSON.

14. Запропонуйте ваші варіанти застосування методів класифікації в оподаткуванні, адмініструванні податків.

15. Наведіть приклади використання класифікаційних методів та дерев рішень у фінансовій сфері.

3.6. Застосування методології нечітких множин в аналізі ризиковості платників податків

3.6.1. Основні відомості про нечіткі множини

Нечіткою множиною Ã на універсальній множині U називається сукупність пар (μAu), u), де μA(u) – ступінь належності елемента u ∈ U до нечіткої множини Ã. Ступінь належності перебуває в діапазоні [0, 1]. Чим вище ступінь належності, тим більшою мірою елемент універсальної множини відповідає властивостям нечіткої множини.

Функцією належності називається така функція, яка дозволяє обчислити ступінь належності довільного елементу універсальної множини до нечіткої множини. Якщо універсальна множина складається з кінцевого числа елементів U = {u1, u2, …,un}, тоді нечітка множина Ã записується у вигляді:

$$\tilde{A} = \sum_{i=1}^{n} \mu_A(u_i) / u_i. \qquad (3.32)$$

Приклад 1. Представити в вигляді нечіткої множини висловлювання «*людина середнього зросту*».

Рішення:

Ã = 0.1/160 + 0.3/165 + 0.8/170 +1/175 +1/180 + 0.5/185.

Лінгвістичною змінною називається така змінна, значення якої є слова та словосполучення деякої природної чи штучної мови.

Терм–множиною називається множина усіх можливих значень лінгвістичної змінної.

Термом називається елемент терм–множини. У теорії нечітких множин терм задається *функцією належності.*

Приклад 2. Розглянемо змінну «*ставка оподаткування*», яка оцінюється за шкалою «*низька*», «*середня*», «*висока*» та «*дуже висока*». У цьому прикладі лінгвістичною змінною є «*ставка оподаткування*», термами є оцінки «*низька*», «*середня*», «*висока*» та «*дуже висока*», які і складають терм–множину.

Нечіткою базою знань про вплив факторів X = {x1, x2,…,xn} на значення параметра *y* називається сукупність логічних висловлювань типу:

$$\begin{aligned}
&\text{Якщо } x_1 = a_1^{j1} \text{ та } x_2 = a_2^{j1} \text{ та } \ldots \text{ та } x_n = a_n^{j1}, \quad \text{або}\\
&\qquad x_1 = a_1^{j2} \text{ та } x_2 = a_2^{j2} \text{ та } \ldots \text{ та } x_n = a_n^{j2}, \quad \text{або}\\
&\qquad x_1 = a_1^{jk_j} \text{ та } x_2 = a_2^{jk_j} \text{ та } \ldots \text{ та } x_n = a_n^{jk_j},\\
&\text{то} \quad y = b_j \qquad j = \overline{1,m},
\end{aligned}$$

де a_i^{jp} – лінгвістичний терм, що оцінює значення фактора xi в p-ій диз'юнкції j-го логічного висловлювання (j = 1, m, p = 1, k j, i = 1, n);

k_j – число диз'юнкції (або) в j-му логічному висловлюванні.

Нечітким логічним висновком називається апроксимація залежності y = f(x1, x2, K, xn) за допомогою нечіткої бази знань та операцій над нечіткими множинами. Для виконання нечітких логічних висновків необхідно знати нечіткі відношення.

Нечітким відношенням R на множині U називається підмножина декартового добутку U × U, яка характеризується функцією належності μR: U × U→[0, 1]. Значення μR(x, y) цієї функції є суб'єктивною мірою виконання відношення xRy (x, y ∈U).

Якщо множина U – дискретна і кінцева, тобто U ={u1, u2, …,uk}, то функція належності μR може бути задана матрицею R розміром к × к. Рівність μ елемента rij матриці R означає, що ступінь виконання відношення xiRxj дорівнює μ.

Приклад 3. Задати відношення «x набагато більше за y» на множині U = {0, 10, 20, 30} (x, y ∈ U).

Рішення: нечітке відношення «x набагато більше за y» може бути задано такою матрицею:

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 \\ 0.5 & 0.2 & 0 & 0 \\ 1 & 0.8 & 0.1 & 0 \end{bmatrix}$$

Дефаззифікацією (від англ. defuzzification) називається процедура перетворення нечіткої множини в чітке число. У теорії нечітких множин процедура дефаззифікації аналогічна знаходженню характеристик положення (математичного очікування, моди, медіани) випадкових величин у теорії ймовірності.

У пакеті MatLab запрограмовані такі методи дефаззифікації (рис. 3.35):

– *Centroid* – центр тяжіння;
– *Bisector* – медіана;
– *Lom* – найбільший з максимумів;
– *Som* – найменший з максимумів;
– *Mom* – середній з максимумів.

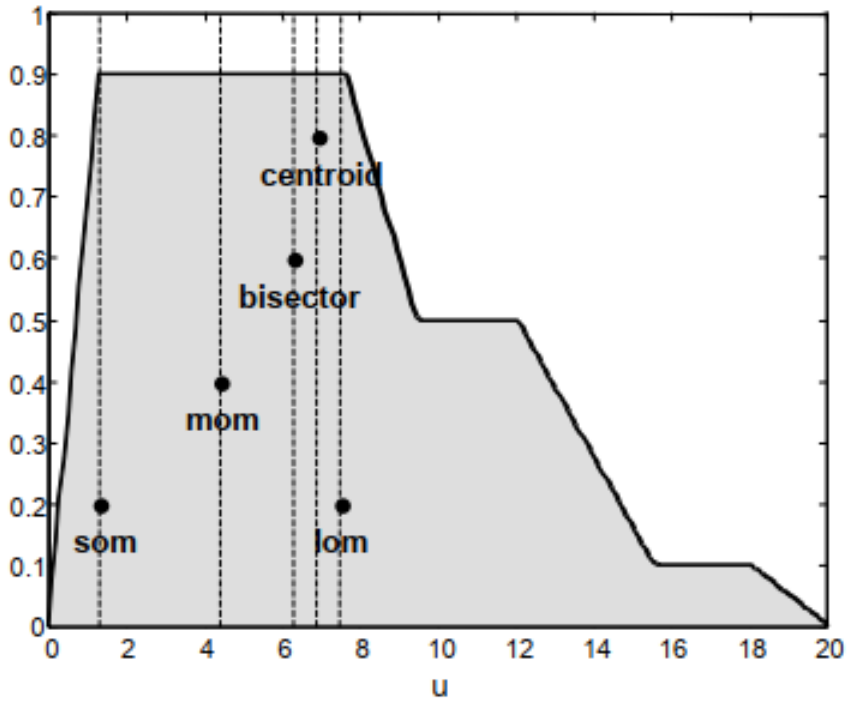


Рис. 3.35. Дефаззифікація за різними методами

Дефаззифікацією нечіткої множини $\tilde{A}$ за методом *Centroid називається обчислення величини* A = Centroid ($\tilde{A}$) за формулою:

$$A = \frac{\sum_{i=1}^{n} u_i \bullet \mu_A(u_i)}{\sum_{i=1}^{n} \mu_A(u_i)}. \qquad (3.33)$$

Приклад 4. Провести дефаззифікацію нечіткої множини «*людина середнього зросту*» з прикладу 1 за методом центра тяжіння (за методом *centroid*).

Рішення: за формулою (3.33) отримуємо:

$$A = - A = \frac{0,1*160 + 0,3*165 + 0,8*170 + 1*180 + 0,5*185}{0,1 + 0,3 + 0,8 + 1 + 1 + 0,5}.$$

Застосування різних методів дефаззифікації проілюстровано на рис. 3.35.

3.6.2. Система нечіткого логічного висновку

Системою нечіткого логічного висновку називається програмне забезпечення, яке моделює залежність Y = f(X) за допомогою нечіткої логіки, де Y(X) – вектор вихідних (вхідних) змінних. Структура *системи нечіткого логічного висновку* показана на рис. 3.36.

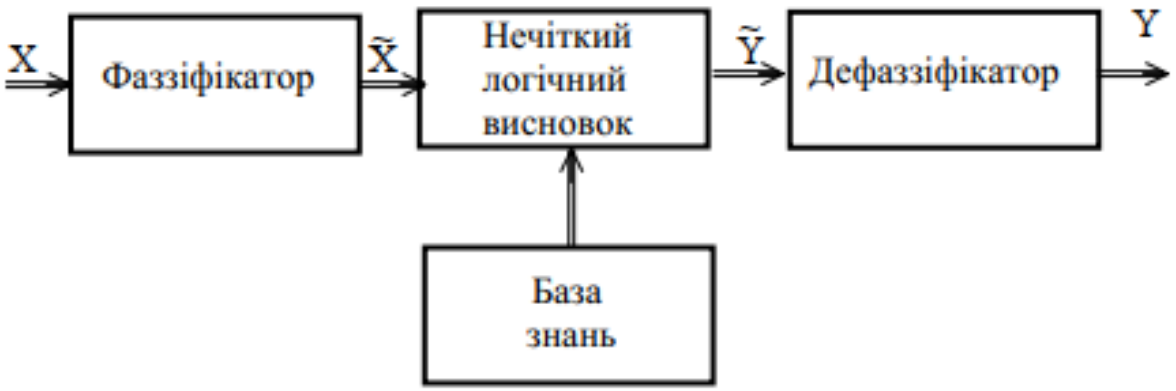


Рис. 3.36. Узагальнена стуктурна схема системи нечіткого висновку

Елементи системи виконують такі функції:

– *фаззифікато*р перетворює фіксований вектор вхідних змінних (вектор факторів впливу) X у вектор нечітких множин X~ за допомогою функцій належностей;

– *база знань* зберігає інформацію про залежність Y = f (X) у вигляді нечітких правил типу «якщо – то»;

– *блок нечіткого логічного висновку* прогнозує вектор нечітких значень вихідних змінних ($\tilde{Y}$), який відповідає вектору нечітких значень вхідних змінних ($\tilde{X}$);

– *дефаззифікатор* перетворює вектор нечітких множин $\tilde{Y}$ у звичайний числовий вектор Y.

Модуль Fuzzy Logic Toolbox дозволяє будувати нечіткі системи двох типів – Мамдані та Сугено. У системах типу Мамдані база знань складається з правил виду *«Якщо* x_1*=низький та* x_2*=середній, то y=високий»*. У системах типу Сугено – з правил виду *«Якщо* x_1*=низький та* x_2*=середній, то* $y=a_0+a_1x_1+a_2x_2$*»*. Таким чином, основна відмінність між системами Мамдані та Сугено полягає у різних способах завдання значень вихідної змінної у правилах, що утворять базу знань.

У системах типу Мамдані значення вихідної змінної задаються нечіткими термами, у системах типу Сугено – як лінійна комбінація вхідних змінних.

3.6.3. Використання методології нечітких множин в оцінюванні податкових ризиків

Для оцінювання ризику несплати податків та розподілу платників податків за категоріями уваги у ДФС України використовуються методи нечіткої логіки. Вони, на відміну від інших методів класифікації, не допускають обмежень виду вхідної інформації (вхідні фактори можуть мати числову, лінгвістичну, логічну форми тощо), дозволяють встановлювати набір правил для здійснення оцінювання рівня ризику та розподіл на категорії, надають можливість провести налагодження параметрів моделі із урахуванням наявного статистичного матеріалу.

Для оцінювання ризиків несплати податків та розподілу платників податків за категоріями уваги (Y) використовується набір факторів X_i, $i=\overline{1,n}$, на основі яких можна обчислити імовірність невиконання суб'єктами господарювання своїх податкових зобов'язань:

$$Y = f_Y(X_1,...,X_n).$$

Алгоритм розрахунків такий: 1) визначається можливий діапазон змінювання контрольованих параметрів X_i, $i=\overline{1,n}$ та вихідної змінної Y; 2) проводиться розрахунок числових значень кожного параметра для здійснення переходу до лінгвістичної форми.

Як вихідна змінна Y автори методики запропонували використовувати значення категорії уваги, до якої було віднесено даного суб'єкта господарювання в результаті проведення виїзної перевірки або сум коштів, донарахованих за результатами перевірки, приведених до добровільно сплачених даних суб'єктом господарювання податків.

Для оцінювання показників X_i, $i=\overline{1,n}$, і подальшого використання в розрахунках на основі нечіткої моделі формується єдина шкала із п'яти якісних термів: *ДН* – дуже низький рівень показника, *Н* – низький рівень показника, *С* – середній рівень показника, *В* – високий рівень показника, *ДВ* – дуже високий рівень показника. Деякі показники, залежно від їх природи, можуть обмежитися і меншою кількістю властивих їм лінгвістичних термів.

Множину значень вихідної лінгвістичної змінної Y, що визначає належність до певної категорії уваги, складатимуть терми: *К*1 – перша категорія уваги, яку характеризує низький ризик несплати податків цим суб'єктом господарювання, *К*2 – друга категорія уваги (ризик несплати податків помірний), *К*3 – третя категорія уваги (високий рівень ризику несплати податків). Із урахуванням належності до певної категорії уваги цей суб'єкт господарювання буде включений до плану-графіка документальних перевірок суб'єктів господарювання.

Задається вигляд функцій належності нечітких термів для різних контрольованих параметрів. Для побудови функцій належності п'яти нечітких термів вхідної змінної {*ДН, Н, С, В, ДВ*} можна відобразити діапазони зміни параметрів X_i, $i=\overline{1,n}$, на єдину універсальну множину X. Функції належності вказаних п'яти нечітких підмножин схематично зображені на рис. 3.37.

Квазідзвоноподібні функції належності для нечітких термів {*К*1, *К*2, *К*3} вихідної змінної Y подані у графічному вигляді на рис. 3.37.

Усі квазідзвоноподібні функції належності нечітких термів як вхідних, так і вихідної змінної зображені на рис. 3.37 та 3.38.

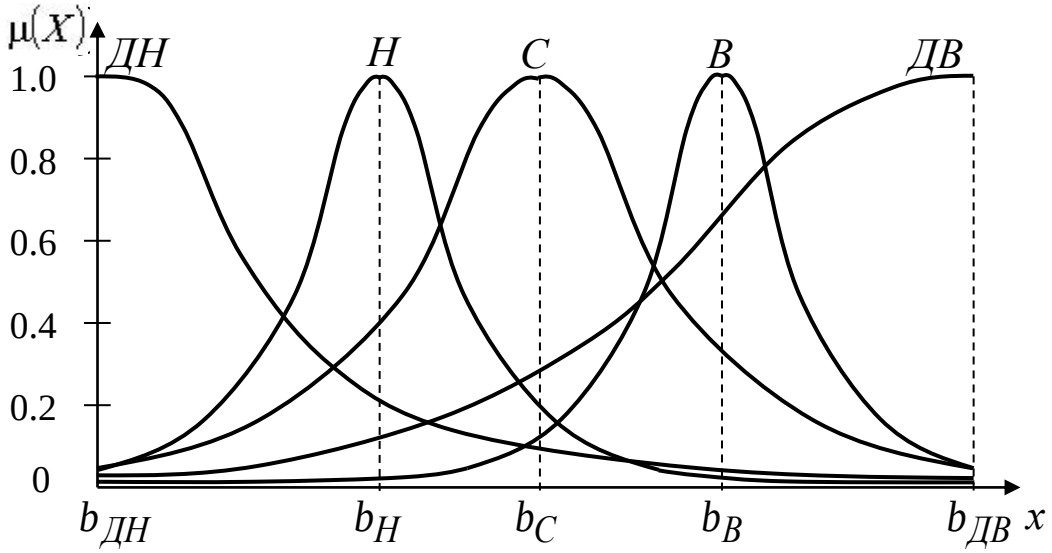


Рис. 3.37. Нечітка зміна X із квазідзвоноподібною функцією належності

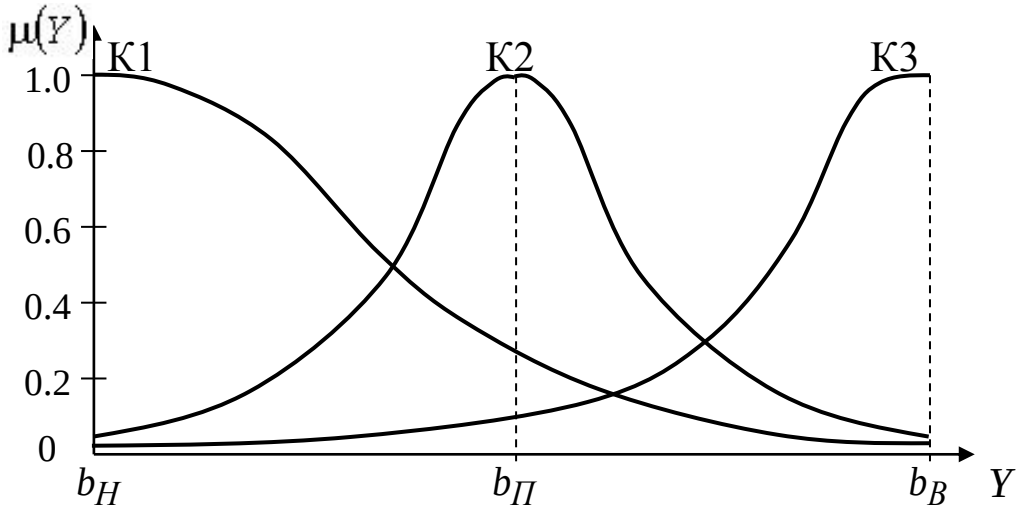


Рис. 3.38. Вихідна зміна Y із квазідзвоноподібною функцією належності

Значення функцій належності бічних термів усіх змінних за межами своїх максимумів *b* прирівнюються до одиниці.

Для математично обґрунтованих суджень про рівень ризику несплати податків підприємством, розподіл суб'єктів господарювання за категоріями уваги із урахуванням відібраних для аналізу показників формується база правил із реалізацією механізму нечіткого логічного висновку.

Значення функції належності з кожного правила розраховуються шляхом множення функцій належності всіх вхідних змінних. А для знаходження функцій належності вихідної змінної з кожного її терму застосовується операція максимізації. Аналітична форма запису вирішального правила для віднесення суб'єкта господарювання до *j*-ї категорії уваги (належність до терму d_j із множини {*К*1, *К*2, *К*3}) із застосуванням вагових коефіцієнтів та функцій належності всіх змінних має вигляд:

$$\mu^j(X_1,...,X_n)= \max_{p=\overline{1,k_j}}\left\{w_{jp}\prod_{i=1}^{n}\mu^{jp}(X_i)\right\}, \qquad j=\overline{1,m},$$

де $\mu^j(X_1,...,X_n)$ – функція належності вихідної змінної *Y* до класу d_j, що визначається на основі вхідних змінних X_i, $i=\overline{1,n}$;

w_{jp} – ваговий коефіцієнт *j*-го, $j=\overline{1,k_j}$, правила для терму d_j вихідної змінної *Y*.

При побудові математичної моделі необхідно задати всі ключові правила. Чим більше система містить логічних правил віднесення підприємств до певної категорії уваги, тим точніше буде проведений аналіз. Якщо в базі знань відсутнє правило, що відповідає поточному стану конкретного підприємства, система видасть рішення, що найбільше підходить цій ситуації. Тобто таке рішення, для якого функція належності вихідної змінної *Y* буде найбільшою серед інших для конкретних значень вхідних змінних X_i, $i=\overline{1,n}$.

У кожній категорії уваги здійснюється ранжування суб'єктів господарювання на основі кількісних значень вихідної змінної *Y*, розрахованих у результаті проведення операції дефаззифікації.

Після побудови моделі та бази знань проводиться її налагодження на існуючому статистичному матеріалі. У такому випадку вдається суттєво підвищити якість логічного висновку. Для налагодження параметрів нечіткої моделі необхідна наявність статистичних даних для вхідних змінних та відповідні до них значення вихідної змінної. Вихідна змінна може бути визначена як на основі експертних суджень, так і розрахована кількісно, а потім

переведена в лінгвістичну форму. Кількісний еквівалент будь-якому лінгвістичному терму вихідної змінної може бути визначений на основі таких показників, як, наприклад, обсяг донарахованих за результатами перевірки коштів, приведений до деякої бази. Як база, може бути використаний обсяг сплачених податків, величина прибутку, обсяг оборотних коштів підприємства чи інші.

Важливо зазначити, що процес розпізнання, виявлення та ідентифікації ризиків необхідно розпочинати із основних цілей, розділяючи їх на підпорядковані цілі (підцілі) на різних функціональних напрямах та структурних рівнях організації.

Питання для самоконтролю до теми 3.6

1. Основні поняття методології нечітких множин. Поняття нечіткої множини.

2. Дайте визначення та наведіть приклад функції належності.

3. Дайте визначення та наведіть приклади: лінгвістичної змінної; терм-множини; терму; функції належності.

4. Дайте визначення нечіткого логічного висновку.

5. Запишіть поняття фаззифікації, дефаззифікації.

6. Дайте визначення нечіткої бази знань.

7. Дана змінна «податкові порушення», яка оцінюється за шкалою: «незначні», «помірні», «значні» та «великі». В цьому прикладі що є: лінгвістичною змінною, термами, терм-множиною?

8. Перерахуйте основні методи дефаззифікації. Наведіть один із прикладів.

9. Наведіть приклад дефаззифікації нечіткої множини за методом Centroid .

10. Опишіть структури та призначення елементів системи нечіткого логічного висновку.

11. У чому відмінність системи нечіткого висновку Мамдані та Сугено?

12. Опишіть особливості застосування нечітких множин в оцінюванні податкових ризиків.

13. Які функції належності використовуються під час оцінювання ризиковості платників податків в податковому відомстві держави?

14. Вкажіть правильний варіант. Для математично обґрунтованих суджень про рівень ризику несплати податків підприємством, розподіл суб'єктів господарювання за категоріями уваги із урахуванням відібраних для аналізу показників формується: реляційна база даних: нечітка база знань; база правил; база фактів; база нечітких висновків.

15. Вкажіть правильний варіант. Значення функцій належності бічних термів усіх змінних за межами своїх максимумів b прирівнюються до: одиниці; до нуля; до відносного значення термів; до їх максимумів.

Лабораторний практикум 3

Лабораторна робота 3.1
Кластерний аналіз регіонів за показниками податкових надходжень (пакет Statistica)

Мета: формування навичок з використання методів кластерного аналізу у середовищі Statistica для класифікаційних задач в аналізі податкової інформації.

Завдання роботи: провести дослідження на наявність кластерів регіонів України за показниками податкового навантаження та валового регіонального продукту у розрахунку на одного звітуючого платника.

Порядок виконання лабораторної роботи

1) ознайомтесь із базовими функціями кластерного аналізу у середовищі Statistica;

2) проведіть кластерний аналіз показників регіонів за наданими у варіанті даними;

3) сформуйте висновки та пропозиції щодо використання цих методів в аналітичній роботі органів податкового відомства України;

4) підготуйте звіт про виконання роботи.

1. Базові функції кластерного аналізу Statistica

Завданням кластерного аналізу є організація спостережуваних даних у наочні структури. Для вирішення цього завдання в кластерному аналізі використовуються такі методи: Joining (tree clustering) (ієрархічні агломеративні методи або деревоподібна кластеризація), K – means clustering (метод До середніх), Two-way joining (двувходове об'єднання). Розберемо принцип проведення кластерного аналізу на основі даних, поданих у таблиці 1 (файл **date_2.sta)**. У файлі містяться дані показника рівня життя населення та показники-аргументи, які беруть участь у класифікації. Розглянемо процес формування вибірок у системі Statistica.

1.1. Відкрийте модуль STATISTIC Cluster Analysis (Кластерний аналіз). Висвітиться назва модуля і далі натисніть кнопку Switch to (Переключитися в) або просто двічі клацніть по назві модуля Cluster Analysis. На екрані з'явиться стартова панель модуля Clustering Method (методи кластерного аналізу): Joining (tree clustering) (ієрархічні агломеративні методи або деревоподібна кластеризація), K – means clustering (метод До середніх), Two-way joining (двувходове об'єднання). Розберемо кожен із цих методів.

1.2. Відкрийте файл (Open Data) date_1.sta. Після вибору Joining (tree clustering) і натискання кн. ОК з'являється вікно Cluster Analysis: Joing (Tree Clustering) (вікно введення режимів роботи для ієрархічних агломеративних методів) (рис. 3.39), в якому кн. Variables дозволяє вибрати змінні, які беруть участь у класифікації. Натисніть на кн. Variables і виберемо всі змінні Select All. Після відповідного вибору натиснемо кн. ОК.

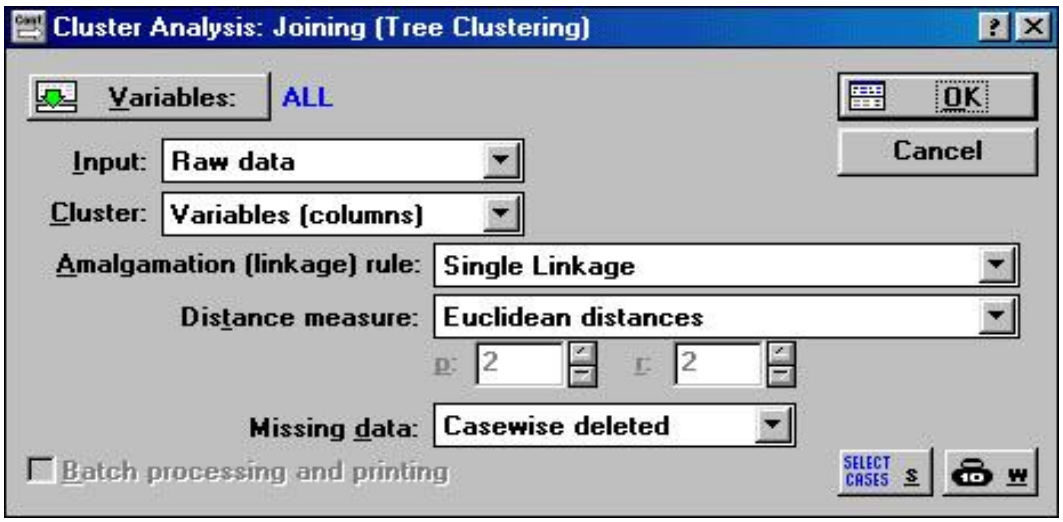


Рис. 3.39. Cluster Analysis: Joing (Tree Clustering) (вікно введення режимів роботи для ієрархічних агломеративних методів)

1.3. Проставте значення як показано на рис. 3.39.

1.4. Після завдання всіх необхідних параметрів і натискання кн. ОК будуть проведені обчислення, а на екрані з'явиться вікно, що містить результати кластерного аналізу Joining Results рис. 3.40.

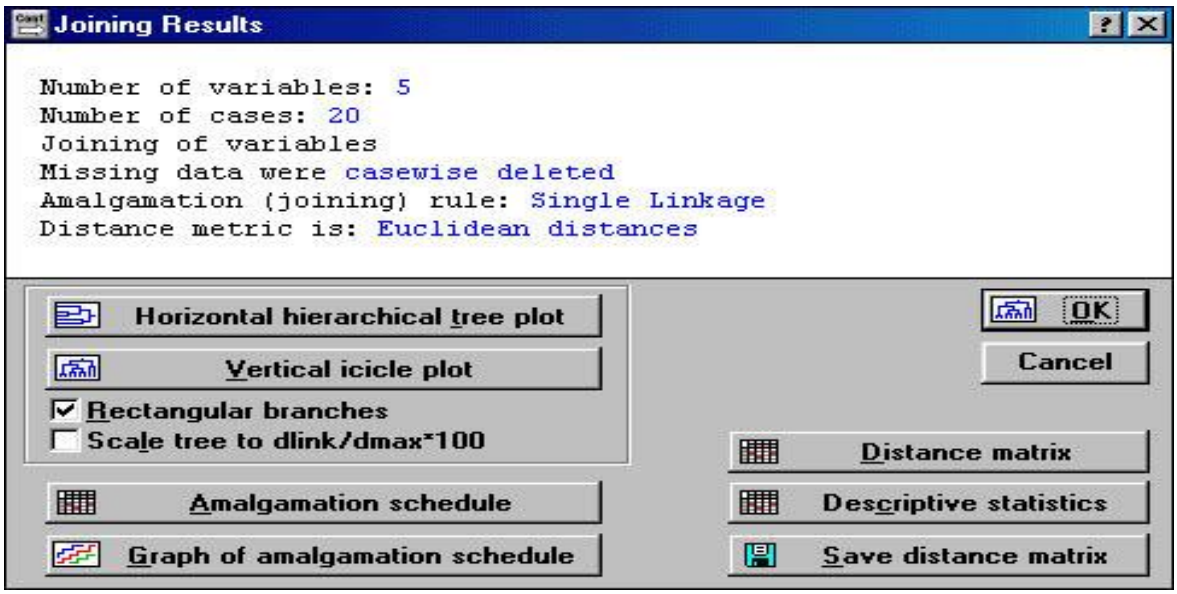


Рис. 3.40. Вікно, що містить результати кластерного аналізу Joining Results

Аналіз результатів

Інформаційна частина діалогового вікна Joining Results Discriminant Function Analisis Results (результати аналізу кластерних функцій) повідомляє, що:

– Number of variables – число змінних;

– Number of cases – число спостережень;

– Missing data were casewise deleted – здійснено класифікацію спостережень або змінних (залежить від рівня параметра в рядку Cluster в попередньому вікні настройки);

– Amalgation (joing) rule – правило об'єднання кластерів (назва ієрархічного агломеративного методу, заданого в рядку Amalgation rules, а в попередньому вікні настройки);

– Distanse.metric is – Метрика відстані (залежить від установки в рядку Distance measure в попередньому вікні настройки.

1.5. Виведіть на екран горизонтальну і вертикальну діаграму (Horizontal hierachical plot або Vertical icicle plot). Найбільш традиційне – вертикальне уявлення (рис. 3.41).

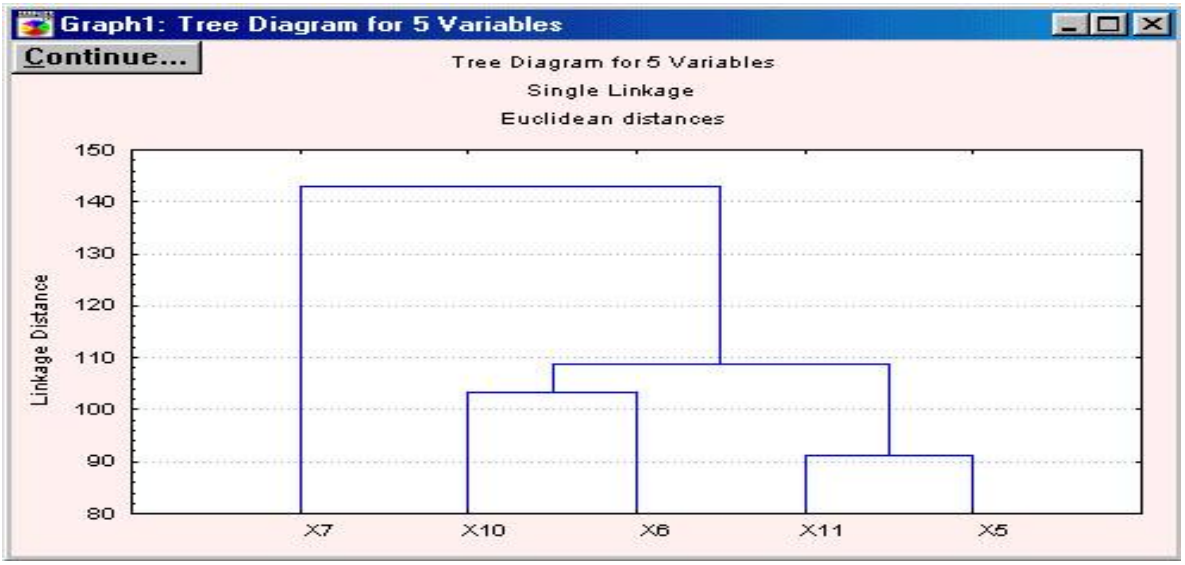


Рис. 3.41. Вертикальна діаграма кластерного аналізу

1.6. Щоб повернутися у вікно, що містить інші результати кластерного аналізу, необхідно клацнути по Continue. Клацнувши мишу, можна розкрити рядок Amalgamation schedule, що містить протокол об'єднання кластерів (рис. 3.42).

Amalgamation Schedule (date_22.sta)

Continue...

Single Linkage
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5
91,07861	X5	X11			
103,1825	X6	X10			
108,6525	X5	X11	X6	X10	
142,8405	X5	X11	X6	X10	X7

Рис. 3.42. Протокол об'єднання кластерів

1.7. Наступною у вікні результатів йде кнопка Graph of amalgamation schedule. Після клацання відкривається вікно, що містить ступеневе, графічне зображення змін відстаней при об'єднанні кластерів (рис. 3.43).

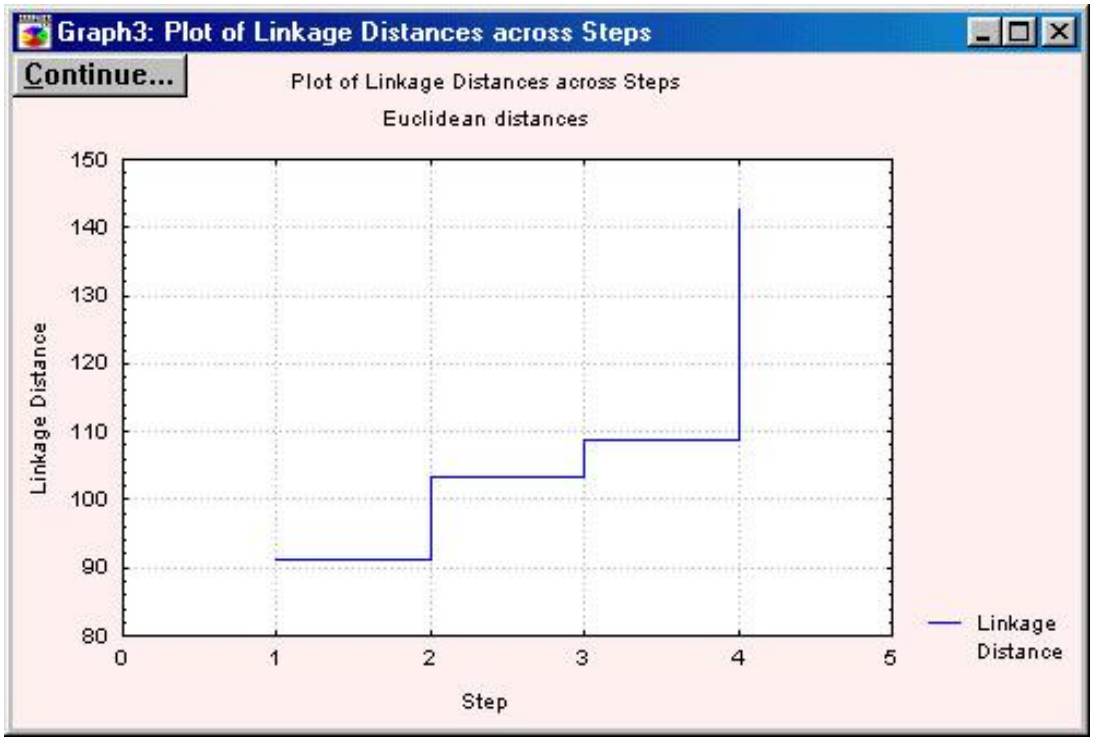


Рис. 3.43. Зображення змін відстаней при об'єднанні кластерів

Повертаючись в основне вікно результатів і класифікації, для перегляду матриці відстаней необхідно клацнути на рядку Distance matrix (рис. 3.44).

Euclidean distances (date_22.sta)

Continue...	X5	X11	X6	X7	X10
X5	,	91,	148,	143,	109,
X11	91,	,	197,	203,	156,
X6	148,	197,	,	210,	103,
X7	143,	203,	210,	,	149,
X10	109,	156,	103,	149,	,

Рис. 3.44. Матриця відстаней між кластерами

В основному вікні результатів класифікації є рядок Save distance matrix as: (Зберегти матрицю відстаней як:) вказати ім'я файлу, в якому буде збережена матриця відстаней, яка в подальшому буде оброблена.

Рядок *Discriptive statistics* містить такі найважливіші описові статистики, як середнє (means) і середньоквадратичне відхилення (standart deviations) для кожного спостереження.

1.8. Проведіть кластерний аналіз показників регіонів за наданими у варіанті даними з податкових надходжень (сплата) та ВРП у розрахунку на одного звітуючого платника. Орієнтовний результат, що повинен бути отриманий, показано на рис. 3.45.

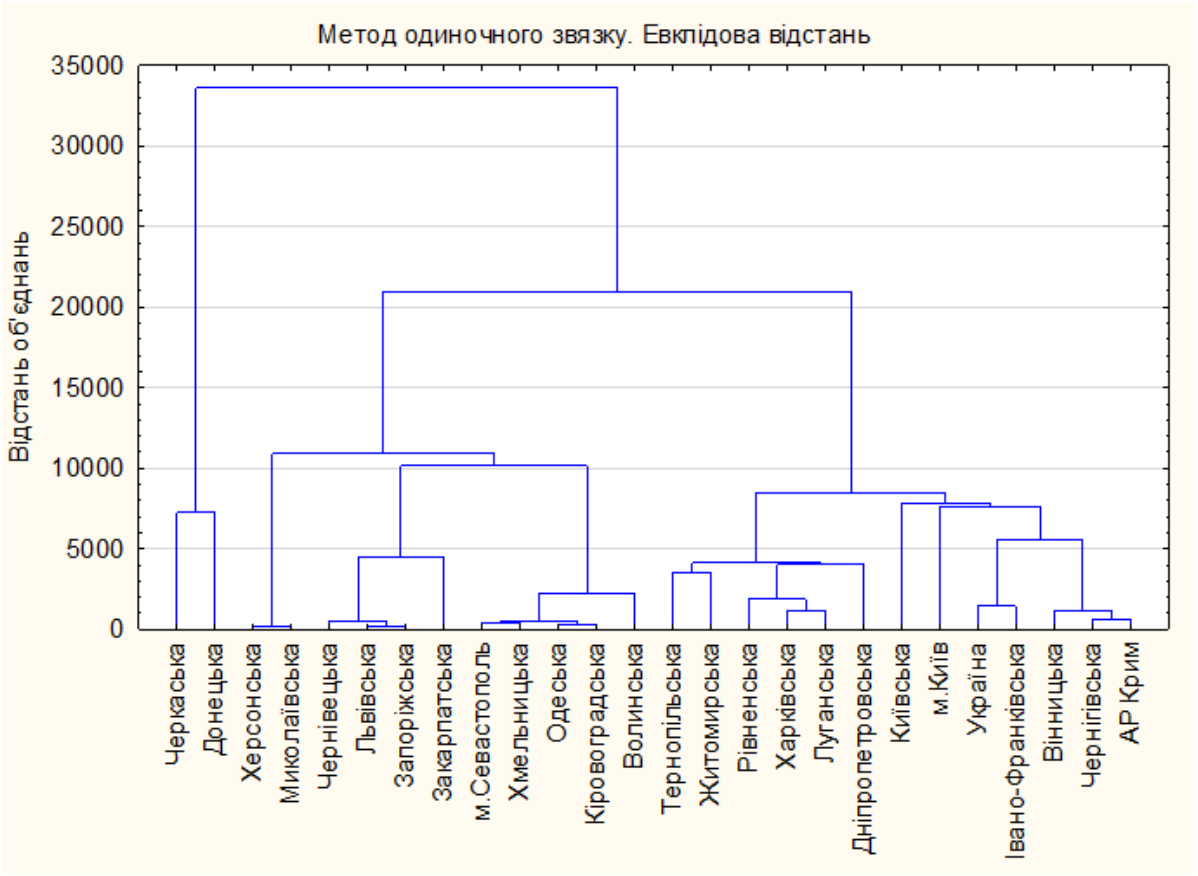


Рис. 3.45. Кластерна дендрограма регіонів за даними з податкових надходжень (сплата) та ВРП у розрахунку на одного звітуючого платника

1.9. Сформуйте висновки та пропозиції щодо використання цих методів у аналітичній роботі органів ДФС України.

1.10. Підготуйте звіт про виконаня лабораторної роботи і розмістіть його для оцінювання в СДН Moodle.

Питання для самоконтролю

1.Поясніть призначення кластерного аналізу.

2. Поясніть суть метрик кластерного аналізу, їх обчислення.

3. Поясніть суть дендрограми кластерного аналізу.

4. Що пояснює матриця відстаней між кластерами?

Лабораторна робота 3.2
Дискримінантний аналіз фінансових показників підприємств за допомогою пакета Statistica

Мета: придбання навичок у використанні пакета Statistica для побудови дискримінантних функцій на основі фінансових показників підприємств та їх ознак загальної збитковості.

Завдання роботи: використовуючи методи дискримінантного аналізу, визначити фінансові показники підприємств, на основі яких можна буде з відповідною ймовірністю прогнозувати збитковість підприємств на 3–4 квартали вперед.

Порядок виконання завдання

1. Ознайомтеся із файлом даних, наданих викладчем для аналізу. Файл даних – excel файл з фінансовими коефіцієнтами даними підприємств: покриття (kp); загальної ліквідності (kzl); абсолютної ліквідності (kal); обіговості оборотних коштів (kook), забезпечення власними оборотними коштами (kzvok); оборотності дебіторської заборгованості (kodz); оборотності кредиторської заборгованості (kokz); фінансової незалежності (kfn); фінансової стабільності (kfs) та ознаками збитковості через календарний рік після розрахунку вказаних коефіцієнтів. Збитковим тут для прикладу розглядається підприємство, що,

А саме, у колонці GR4:

– 1 – ознака підприємств, що були і залишилися збитковими (ЗЗ);

– 2 – підприємства, що були прибутковими і стали збитковими (СЗ);

– 3 – підприємства, що були збитковими і стали прибутковими (СП);

– 4 – підприємства, що були прибутковими і залишилися прибутковими (ЗП).

У колонці GR2:

– цифра 2 – ознака підприємств, що були і залишилися збитковими, об’єднані із підприємствами, що були прибутковими і стали збитковими (ЗЗПЗ);

– цифра 4 – підприємства, що були збитковими і стали прибутковими, об’єднані з тими, що були прибутковими і залишилися прибутковими (СПЗП). Кількісний розподіл підприємств на групи наведено у табл. 3.13

Таблиця 3.13

Класифікація підприємств дослідної вибірки

№	Групи підприємств	Дві групи	Кіль-кість	Аналіз 4-х груп	Кіль-кість
1.	Підприємства, що були і залишилися збитковими	2-ЗЗСЗ	69	1-ЗЗ	46
2.	Підприємства, що були прибутковими і стали збитковими			2-СЗ	23
3.	Підприємства, що були збитковими і стали прибутковими	4-СПЗП	83	3-СП	29
4.	Підприємства, що були прибутковими і залишилися прибутковими			4-ЗП	54
	Усього		152		152

2. 2. Завантажити файл у середовище Statistica.

Орієнтовний вигляд завантаженого файлу показано на рис. 3.46.

	F:\00_DATA\00_УДФСУ\00_КАФЕДРА-ПРЕДМЕТИ-КУРСИ\007_ІАТ-ПМО\04_ПРАКТИКУМ\00_Т									
	1 EDRPC	2 NAME	3 KTP	4 KFV	5 KOPF	6 svd3	7 w3	**8 po303**	9 pz3	10 kow3
1	131363	ДП ЧЕРН	34	31	140	52477,8	45707,7	6009,7	1802,9	0,131481
2	132606	ДП "ЧЕР	275	20	330	137,6	192,1	-99,8	0	-0,51952
3	176348	ДВАТ "В.	190	31	330	102	132	-51	0	-0,38636
4	191916	ВАТ "ВА	285	20	231	7712	6092,3	1105,5	99,5	0,181459
5	193163	ЧЕРКАС	680	20	231	2737,7	2217,9	274,6	82,4	0,123811
6	203826	АЗОТ В.	680	20	231	189153,5	156021,2	25482,1	5118,42	0,163325
7	204033	ЧЕРКАС	680	20	231	56168,2	53857,4	542	162,6	0,010064
8	205104	ЗАВОД Х	190	31	140	4808	2728,3	1721,8	0	0,631089
9	222456	ВІДКРИТ	405	20	231	1001,6	1083,1	-214	0	-0,19758
10	223941	ВАТ "ДА	405	20	231	3058,1	3182,1	-546,5	0	-0,17174
11	255289	ВАТ" ТЕН	680	20	231	1821,3	1613,7	-145,4	0	-0,0901

Рис. 3.46. Файл даних підприємств із вказаними ознаками збитковості (GR2, GR4)

3. Змінними групуваннями можемо вибирати класифікаційну ознаку GR4, у якій група підприємств розділена на чотири групи, або GR2, у якій група підприємств розділена на дві групи за показником збитковості. Як незалежними змінними – вищезазначені коефіцієнти kp, kzl, kal, kook, kzvok, kodz, kokz, kfn, kfs (число при коефіцієнтах вказує на варіант).

4. Аналіз будемо здійснювати поетапно. Кількість кроків відповідає кількості змінних – 9. Поетапний аналіз з *включенням / виключенням* на кожну ітерацію за допомогою статистики Фішера визначає, чи варто включати у модель відповідну змінну. Виберіть опцію *Ststistics/Multivariate Exploratory Technique/Discriminant Analysis* (рис. 3.47).

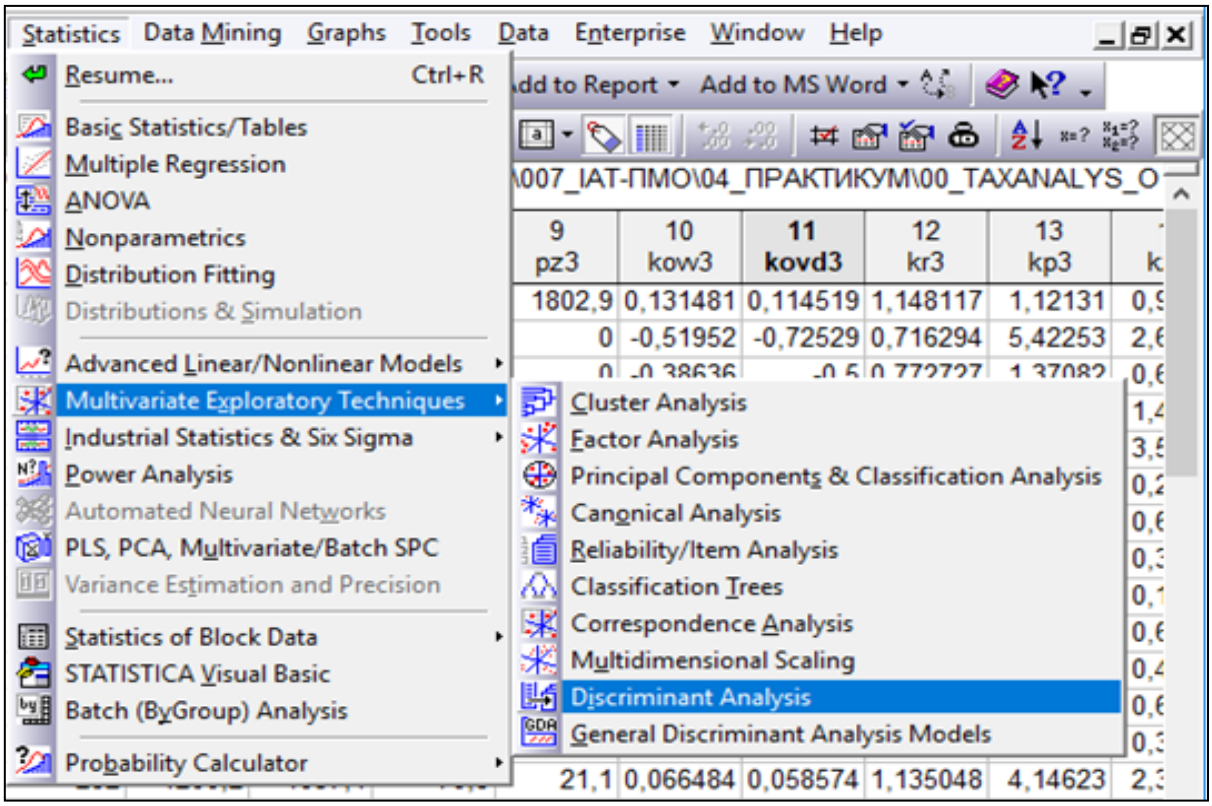


Рис. 3.47. Вибір опції *Discriminant Analysis*

5. Виберіть як змінну групування показник GR4, а як незалежні змінні – вищезазначені коефіцієнти (усього 9) . Натисніть «Ок» (рис. 3.48).

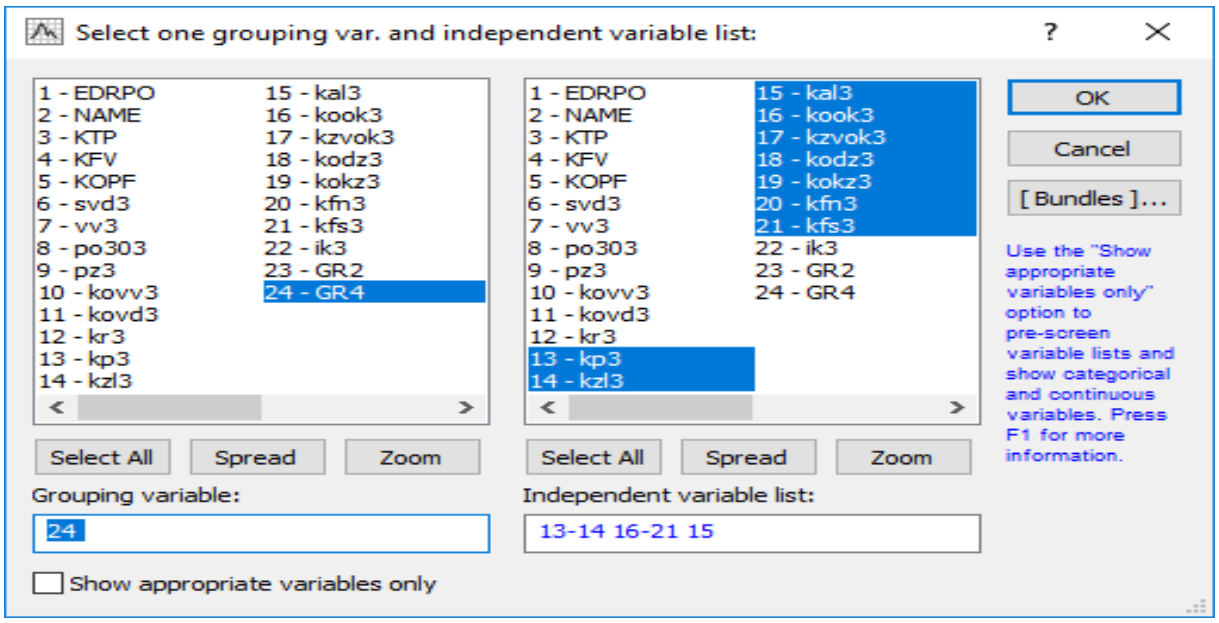


Рис. 3.48. Вікно вибору змінних

Зазвичай у поетапному аналізі дискримінантної функції змінні включають у модель, якщо відповідне їм значення F більше ніж *значення F-включити*, змінні видаляють з моделі, якщо відповідне їм значення F менше ніж значення *F-виключити*.

6. Установка значень *F-включити/виключити* змінні.

Значення F-включити завжди має бути більше ніж значення F-виключити. Якщо при проведенні поетапного аналізу з включенням будемо включати всі змінні, встановлюємо в полі *F-включити* значення, рівне дуже маленькому числу (наприклад, 0.0001), а в полі *F-виключити* значення 0.0 (рис. 3.49).

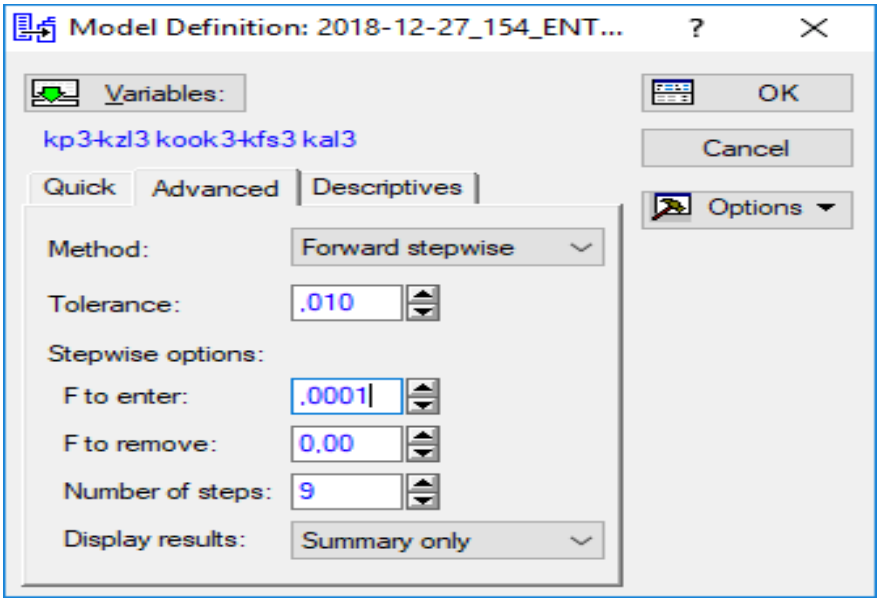


Рис. 3.49. Вікно установи параментрів аналізу

Якщо при проведенні поетапного аналізу *з виключенням* будемо виключити всі змінні з моделі, встановлюємо в полі *F-включити* значення, рівне дуже великому числу (наприклад, 0.9999), а в полі *F-виключити* трохи менше значення того ж порядку (наприклад 0.9998).

7. Для отримання дискримінантних функцій натисніть Summary.

Лямбда Уілкса (Wilks' Lambda) – значення визначає значення L при включенні зміної на даному етапі.

Частинна лямбда *Лямбда Уїлкса.* Ця статистика для одноразового вкладу відповідної змінної в дискримінацію між сукупностями є аналогом частинної кореляції.

F-включити(F-remove) і p-значення (p-value). Використовується як і F-статистика для всієї моделі тільки замість лямбда Уілкса підставляється частинна лямбда Уїлкса.

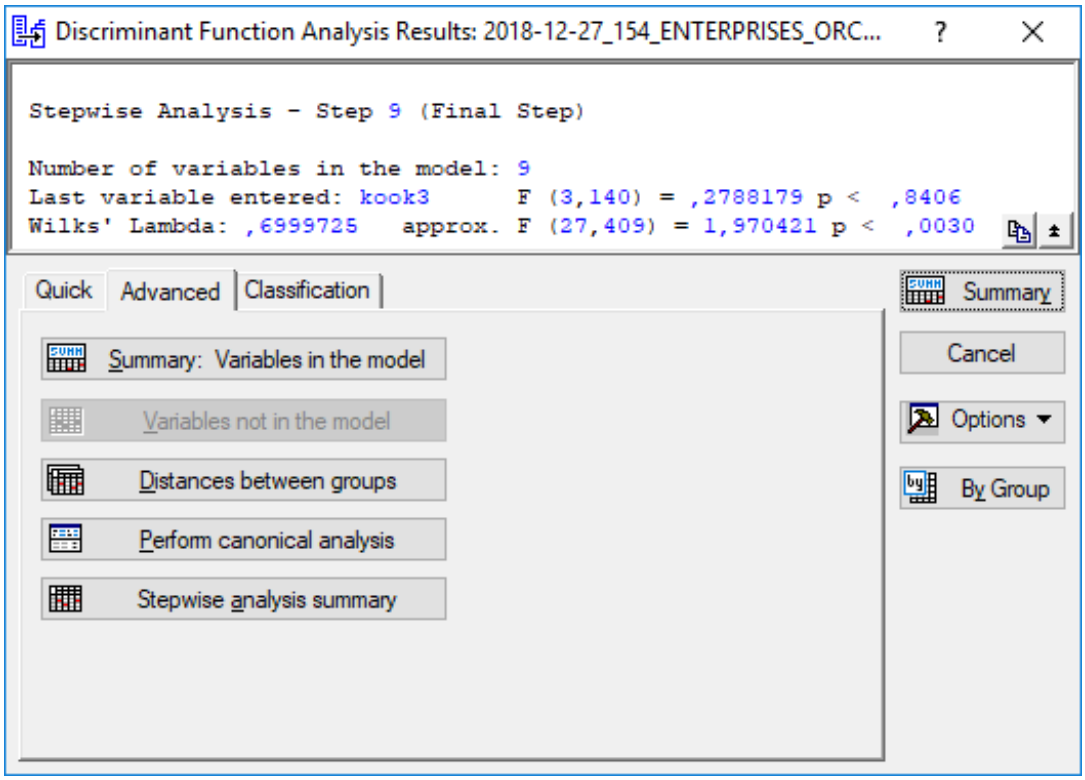


Рис. 3.50. Вікно відображення результатів аналізу

8. Алгоритм дискримінантного аналізу зупиняється, якщо на наступному етапі значення *F-вкл. в модель* виявляється менше заданого значення (у нашому прикладі F-вкл. = 9) або якщо на наступному етапі вже всі змінні будуть у моделі. Проаналізуйте отриману дискримінантну функцію та результуючу таблицю (рис. 3.51).

N=152	Discriminant Function Analysis Summary (2018-12-27_154_ENTI Step 9, N of vars in model: 9; Grouping: GR4 (4 grps) Wilks' Lambda: ,69997 approx. F (27,409)=1,9704 p< ,0030 **Wilks' Lambda**	Partial Lambda	F-remove (3,140)	p-value	Toler.	1-Toler. (R-Sqr.)
kzvok3	0,780716	0,896577	5,383134	0,001550	0,709277	0,290723
kokz3	0,715398	0,978438	1,028395	0,382085	0,247281	0,752719
kodz3	0,739840	0,946113	2,657938	0,050727	0,885666	0,114334
kfs3	0,731920	0,956351	2,129944	0,099168	0,744189	0,255811
kal3	0,714440	0,979750	0,964525	0,411421	0,634007	0,365993
kfn3	0,710682	0,984930	0,714029	0,545180	0,578847	0,421153
kp3	0,716752	0,976590	1,118678	0,343733	0,238291	0,761709
kzl3	0,707613	0,989203	0,509374	0,676455	0,323436	0,676564
kook3	0,704155	0,994061	0,278817	0,840615	0,467020	0,532980

Рис. 3.51. Отримана дискримінантна функція

Як видно із рис. 3.51, найбільші значення величини *F-включити* дає змінна kzvok3 – *забезпечення власними оборотними коштами*. Змінна з максимальним значенням F-включити буде включена в модель на першому етапі (тобто вносить найбільший вклад у модель).

9. Проведіть класифікацію підприємств на основі отриманої дискримінантної функції, для чого виберіть закладку *Classification.*

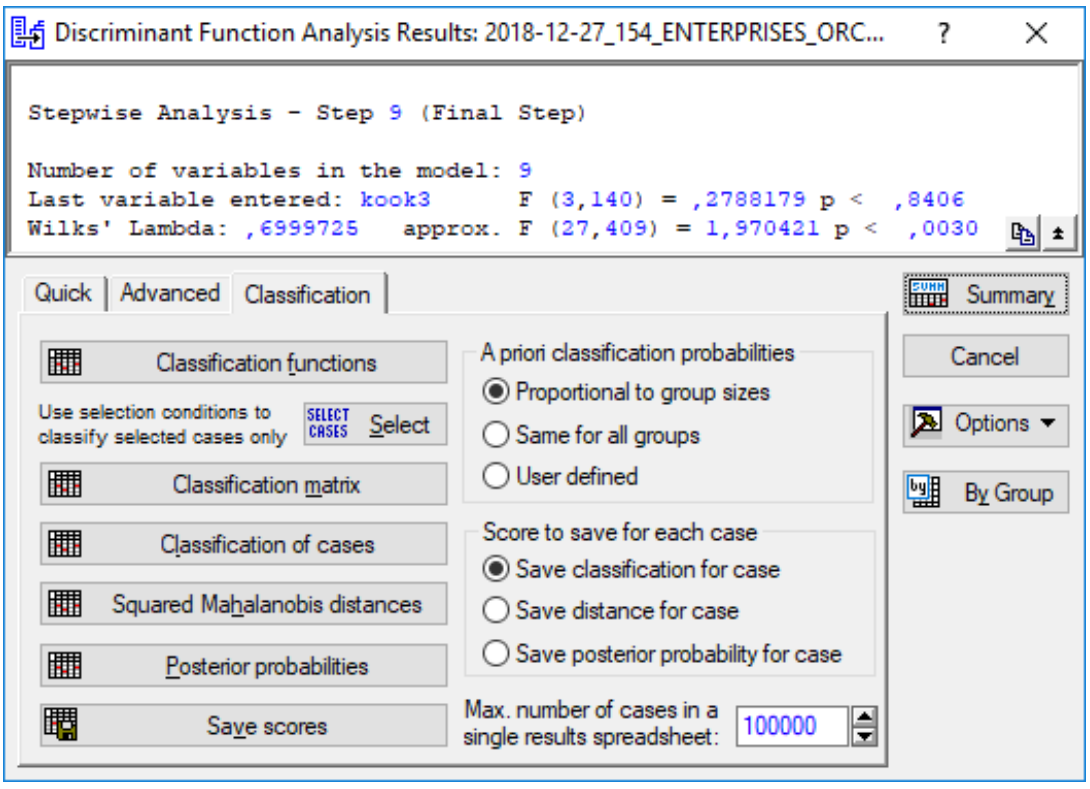


Рис. 3.52. Вікно розширених параметрів дискримінантного аналізу

Для виведення коефіцієнтів функцій класифікації кожної групи виберіть *Classification functions.* Результат показано на рис. 3.53. У кожному стовпці знаходяться коефіцієнти дискримінантної функції для відповідного класу (варто зазначити, що мається на увазі лінійна функція).

	Classification Functions; grouping: GR4 (20			
Variable	**G_1:1 p=,30263**	G_2:2 p=,15132	G_3:3 p=,19079	G_4:4 p=,35526
kzvok3	-2,12841	-1,41294	-1,46250	-1,37524
kokz3	-0,39623	-0,26669	-0,36330	-0,29318
kodz3	-0,00203	0,01955	0,00034	-0,00292
kfs3	-0,05579	-0,06332	-0,06539	-0,03097
kal3	-0,59746	-1,36015	-0,62015	0,38541
kfn3	14,43720	14,12087	14,69399	13,10327
kp3	0,67002	0,15320	0,48433	0,25235
kzl3	0,98028	1,27887	1,00339	1,45480
kook3	1,04111	0,89367	0,94482	1,02815
Constant	-7,72925	-7,71491	-7,61236	-6,82049

Рис. 3.53. Класифікаційні функції для кожної із груп ознаки GR4

10. Виведіть матрицю класифікації. Для цього виберіть *Classification Matrix (рис. 3.54).*

	Classification Matrix (2018-12-27_154_ENTERPRISES Rows: Observed classifications Columns: Predicted classifications				
Group	Percent Correct	G_1:1 p=,30263	**G_2:2 p=,15132**	G_3:3 p=,19079	G_4:4 p=,35526
G_1:1	47,82609	22	0	11	13
G_2:2	8,69565	6	2	4	11
G_3:3	24,13793	7	1	7	14
G_4:4	83,33334	4	1	4	45
Total	50,00000	39	4	26	83

Рис. 3.54. Матриця класифікації на основі отриманих функцій

Як видно із рис. 3.53, отримані дискримінантні функції найбільш якісно класифікували групу підприємств із ознакою «4» – підприємства, що були прибутковими і залишилися прибутковими, 83 % коректно класифікованих підприємства. Найнижча якість класифікації підприємств групи «3» – підприємств, що були прибутковими і стали збитковими.

11. Повторіть пункти 5–9 для змінної групування GR2. Результат повинен бути орієнтовно таким, як показано на рис. 3.55.

Group	Classification Matrix (2018-12-27_15 Rows: Observed classifications Columns: Predicted classifications Percent Correct	G_1:2 p=,45395	G_2:4 p=,54605
G_1:2	39,13044	27	42
G_2:4	87,95181	10	73
Total	65,78947	37	115

Рис. 3.55. Результат класифікації для змінної GR2

У першому стовпчику таблиці показано відсоток спостережень, які були правильно класифіковані для кожної сукупності отриманими функціями класифікації. Решта стовпчиків дають число випадків правильної і неправильної класифікації для кожної іншої сукупності.

Отже, коли класифікуються спостереження, за якими вже була знайдена дискримінуюча функція, для деяких груп отримується порівняно непогана дискримінація. Однак цю класифікацію розглядають тільки як діагностичний засіб ідентифікації сильних і слабких сторін отриманих дискримінантних функцій, оскільки ці класифікації не є апріорними прогнозами, а швидше апостеріорними. Лише за умови класифікації нових зразків цю таблицю можна інтерпретувати в термінах як дискримінація.

Питання для самоконтролю

1. Які статистики використовують для оцінювання спроможності дискримінантної функції розпізнавати класи у багатовимірному ознаковому просторі?
2. Що характеризує канонічні корені дискримінантних функцій?
3. Як характеризується якість дискримінації?
4. До якого класу методів Data Mining належить метод дискримінації?
5. Підготуйте звіт про виконання лабораторної роботи. Розмітьте для перевірки у системі СДН Moodle.

Лабораторна робота 3.3
Оцінювання параметрів кривої Лаффера у середовищі Statistica

Мета: придбання навичок у використанні Statistica для оцінювання параметрів нелінійних моделей.

Завдання роботи

1. Ознайомитися із теоретичним матеріалом теми (пп. 3.2.1).

2. Завантажити у середовище статистики вказаний у варіанті файл для аналізу.

3. Здійснити оцінювання параметрів нелінійної моделі кривої Лаффера для групи визначених викладачем регіонів за вказаними показниками.

4. Визначити стан здійснення податкової політики на основі отриманих параметрів функцій та сформувати відповідні висновки.

5. Підготувати звіт про виконання роботи.

Порядок виконання лабораторної роботи

1. Оцінювання параметрів нелінійної моделі кривої Лаффера у середовищі Statistica.

1.1. Завантажте файл 99_1GR.STA в середовище Statistica. Пропуски у даних вказують на наявність у файлі даних однієї із сформованих кластерним аналізом груп. У цьому випадку у файлі дані по групі 1. Це Вінницька, Житомирська, Закарпатська, Тернопільська, Херсонська, Хмельницька, Чернівецька. Другу групу утворюють відповідно: АР Крим (АРК), Івано-Франківська, Луганська, Львівська, Миколаївська, Одеська, Сумська, Черкаська області.

Відповідно до теорії будемо аналізувати наявність кривої Лаффера для таких комбінацій даних:

– фіскальна крива: податкове збов’язання у розрахунку на одного фактичного платника податків (показник 41 - `tz1p)`;

– виробнича крива: ВДВ у розрахунку на одного фактичного платника податків (показник 42- `vdv1p)`;

– ефективна ставка оподаткування ВРП: (показник 23- `est)`.

Відповідні формули виробничої та фіскальної кривих для нелінійного оцінювання параметрів показані нижче:

Формула виробничої кривої:
vdv1p=b*est+a*est**2

Формула фіскальної кривої
tz1p=b*est**2+a*est**3

Відповідно, для визначення точок першого і другого порядку необхідно оцінити параметри *а* і *в* кожної формули.

1.2. Відкрийте вікно введення параметрів методу «Нелінійне оцінювання» («Statistucs/Advanced Leanear / Nonleaner Models / Nonleaner Estinations»)/.

1.3. Виберіть «Оцінювання параметрів функції заданої користувачем» («User-specified regression, list squares»).

1.4. Виберіть закладку «Function to be estimated» і введіть у вікні «Estimated Function» формулу виробничої кривої, натисніть «ОК» (рис. 3.56).

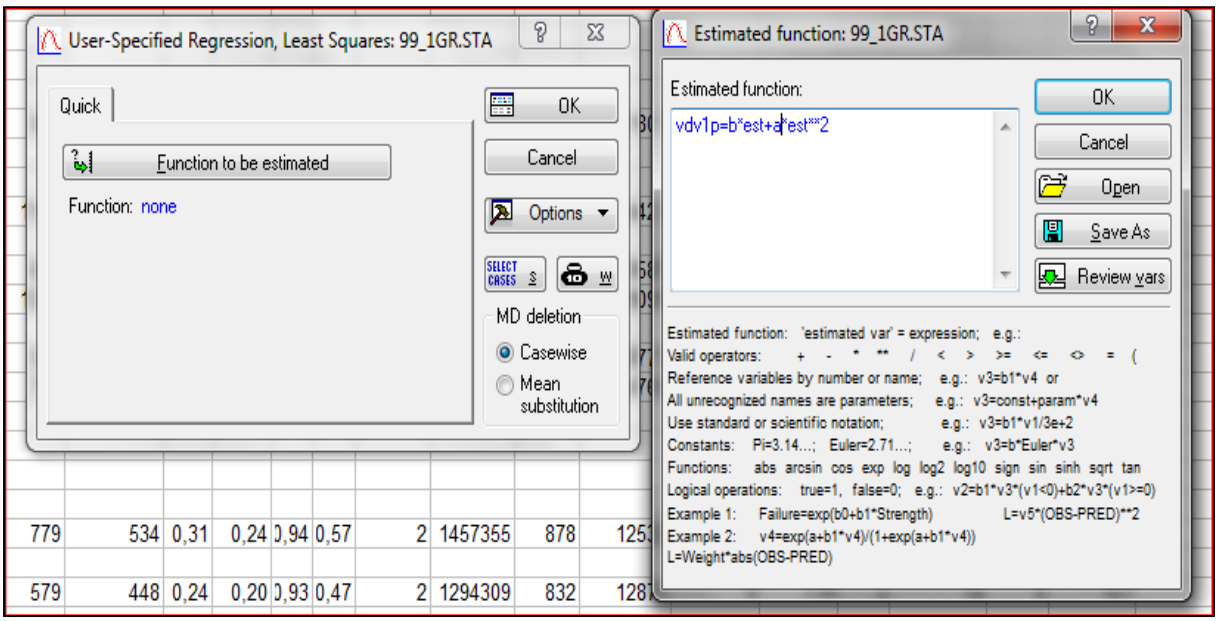


Рис. 3.56. Вікно «Estimated Function»

1.5. Вибиріть у закладці «Quick» метод оцінювання «Gauss-Newton». У закладці «Advanced» залишаємо параметри оцінювання, натисніть «ОК» (рис. 3.57):

Maximum number of iterations: 50 = кількість ітерацій;

Convergence criterion: 1,0 *E -6 – рівень збіжності.

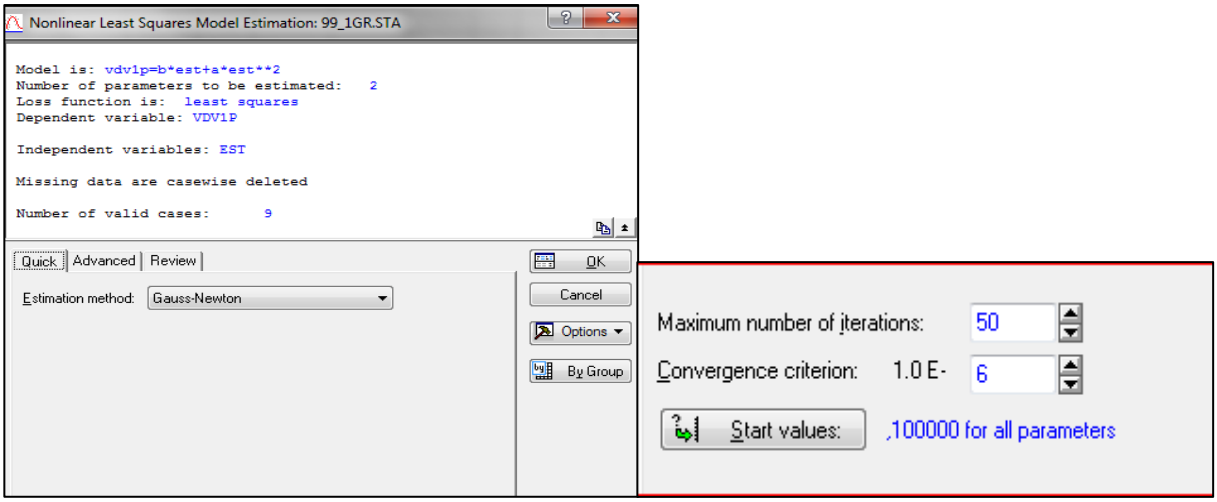


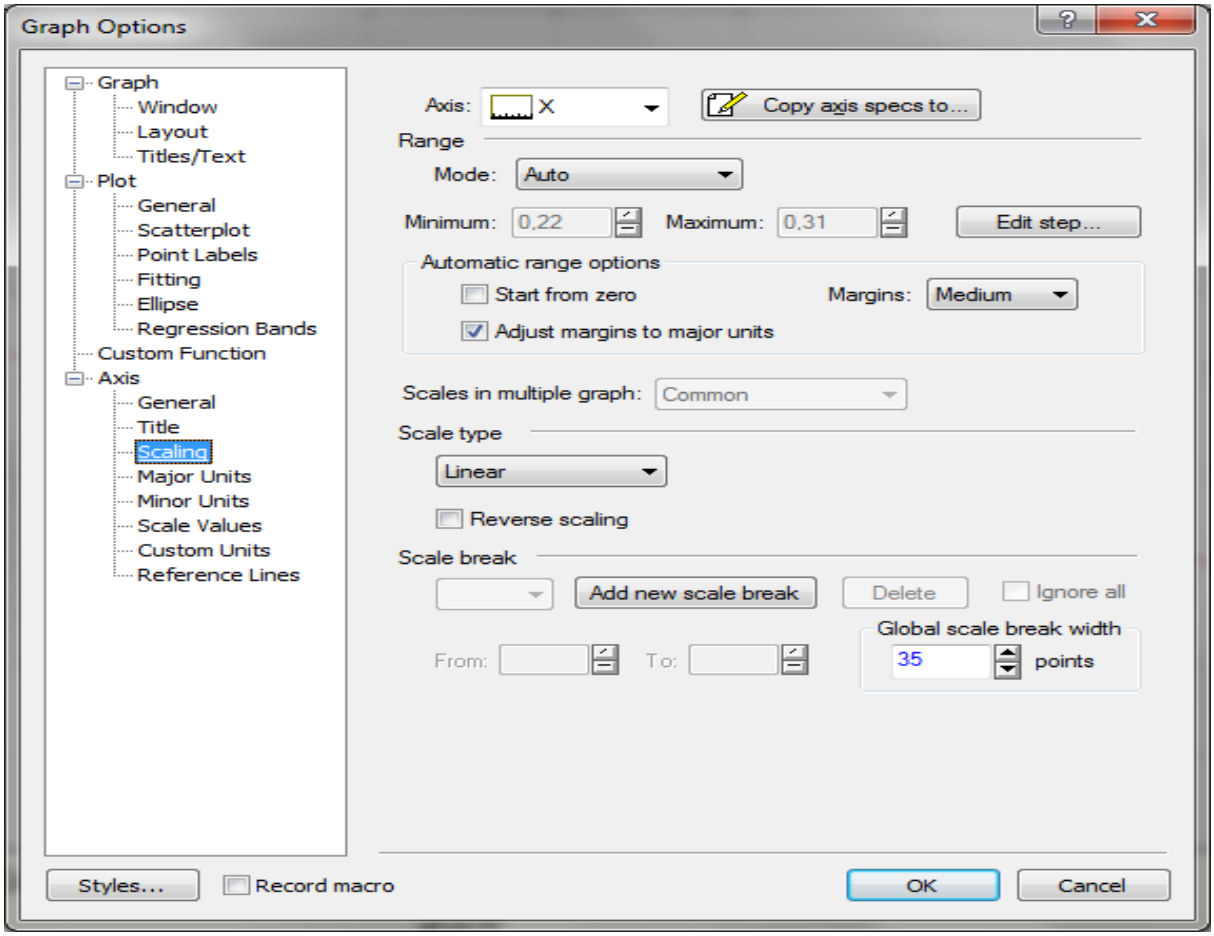
Рис. 3.57. Закладка бистрого оцінювання

Отримуємо модель виробничої кривої із оціненими параметрами:

y=(22311e2)*x+(-5365E3)*x**2або

vdv1p=(22311e2)*est+(-5365E3)*est**2

1.6. Для виведення графіка повністю у вікні модуля редагуємо параметри «Scaling» як показано на рис. 3.58. Результат – рис. 3.59.



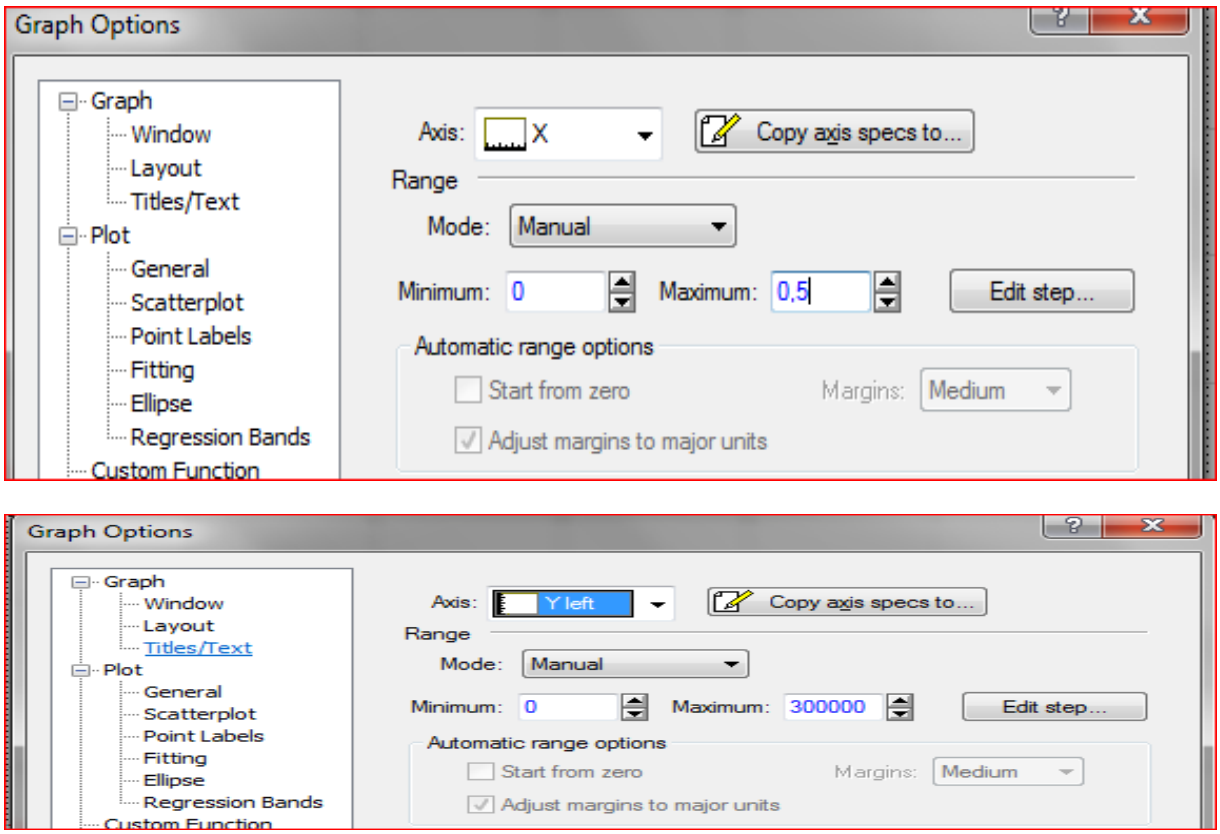


Рис. 3.58. Редагування параметрів для виведення графіка

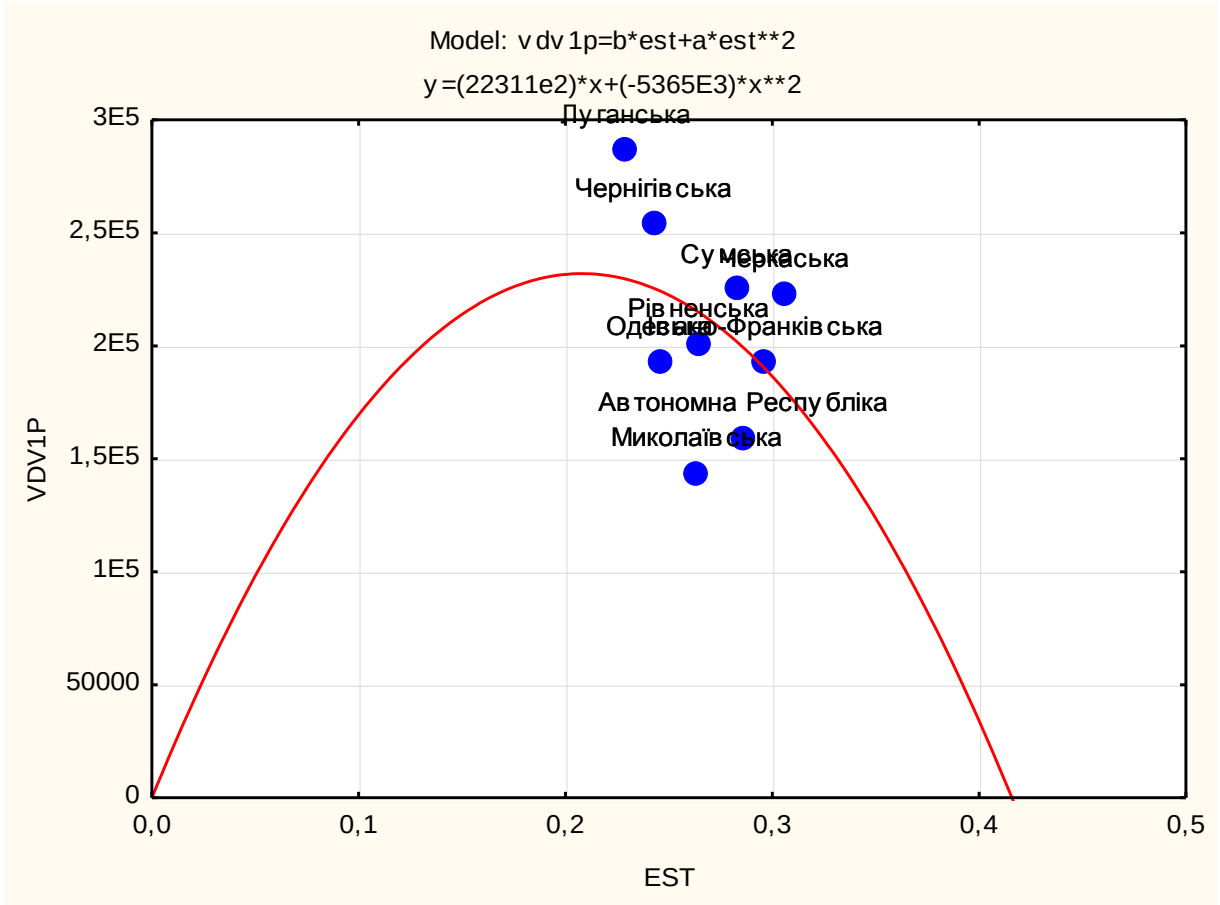


Рис. 3.59. Результат побудови графіка виробничої функції

1.7. Вводимо формули. Формули можна вводити як копіюванням – вставкою із буфера, так і в ручному режимі, використовуючи закладку Estimtion function/Review vars.

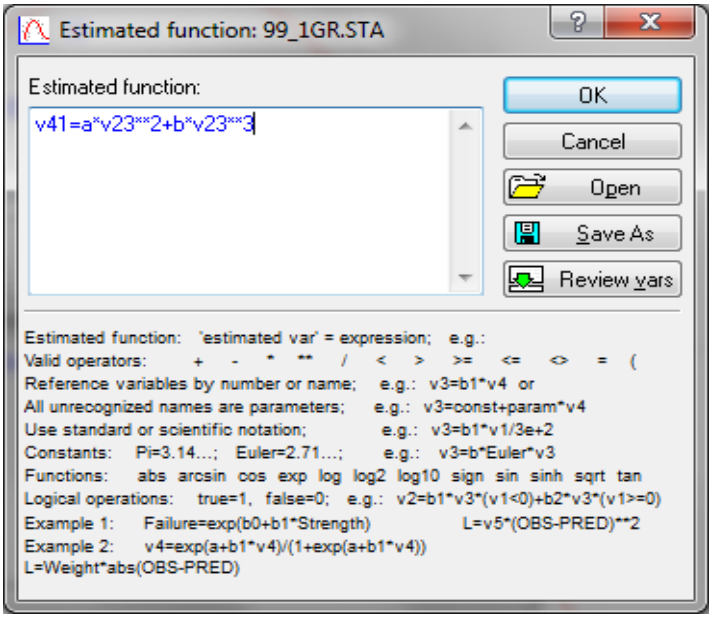


Рис. 3.60. Вікно введення формули для нелінійного оцінювання

1.8. Таким же чином отримуємо графік із оціненими параметрами фіскальної кривої (рис. 3.61):

Формула фіскальної кривої
tz1p=b*est**2+a*est**3,

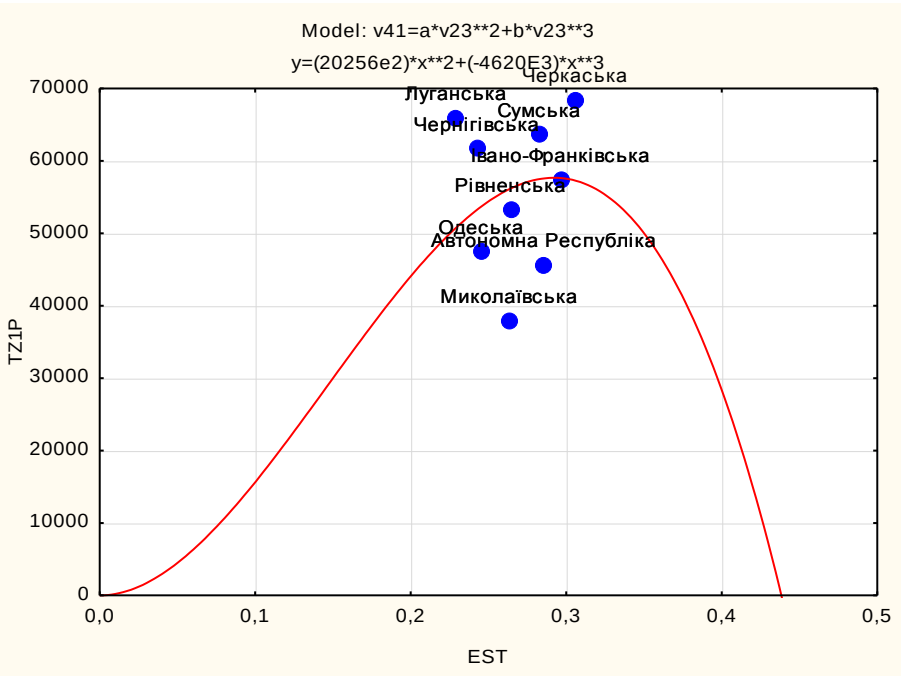


Рис. 3.61. Графік із оціненими параметрами фіскальної кривої

Отриманий результат оцінювання параметрів нелінійної функції : `tz1p` =(20256e2)*`est` **2+(-4620E3)*`est` **3

1.9. На основі отриманих параметрів виробничої та фіскальної функцій за допомогою формул (1.11) можна розрахувати значення точок Лаффера (першого та другого типу та індекс фіскальності (формула 1.12), що характеризує рівень фіскальності податкової політики для цієї групи регіонів. Розрахуйте ці показники і зведіть їх у таблицю (дані заповнюються у рядку залежно від наданих викладачем даних), сформуйте висновки.

1.10. Для зведення двох графіків в один графік виконайте такі процедури:

1. Виберіть у головному меню: Graphs / Scatterplots.

2. У закладці 2D Scatterplots виберіть змінні (Variables), Grahp type і Fit як вказано на рис. 3.62. При натисканні «Ок» у виведеному графіку будуть присутні значення для обох функцій, позначені різними кольорами.

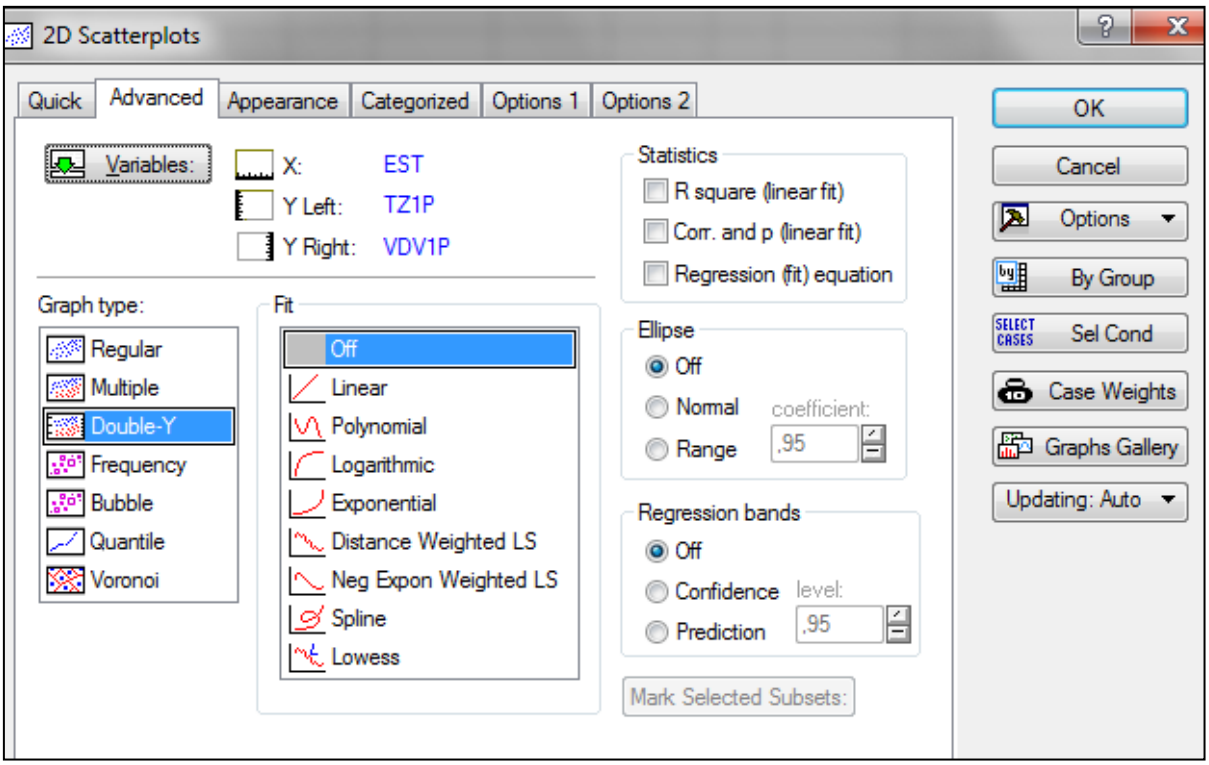


Рис. 3.62. Вікно «2D Scatterplots»

1.11. Далі на одному екрані будуємо два графіки:

1.tz1p =(20256e2)*est **2+(-4620E3)*est **3
2. vdv1p =(22311e2)*est+(-5365E3)*est**2

Для введення замінимо імена змінних на y та x:

y =(20256e2)*x **2+(-4620E3)*x**3
y =(22311e2)*x+(-5365E3)*x**2

Натиснувши на полі рисунка правою клавішею, вибираємо Graph Option / Fitting /Add new fit . У полях:

Plot – для першого графіка вибираємо першу залежну змінну VDV1P, відповідно для другого графіка – TZ1P.

У полі Fit type вибираємо Custom Function.

У полі Cotom Function вводимо або копіюємо і вставляємо формули функції як показано на рис. 3.63–3.64.

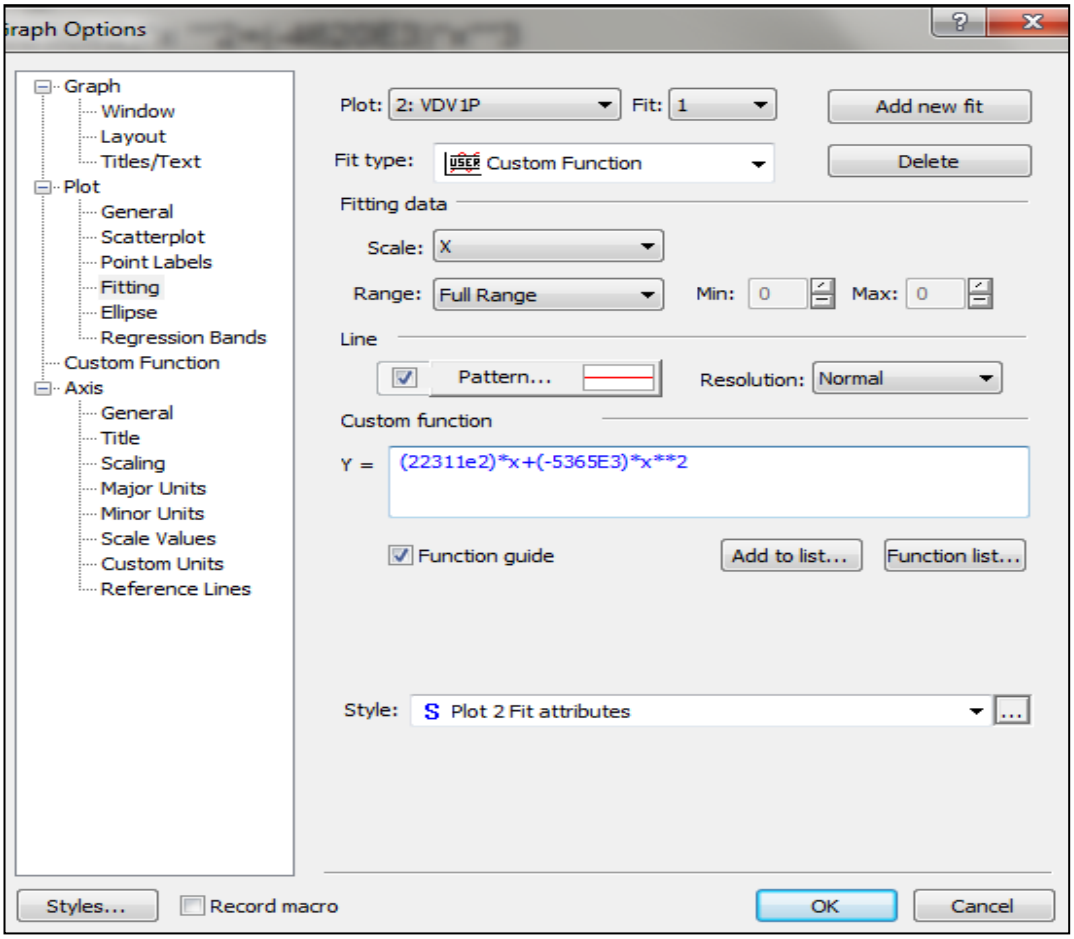


Рис. 3.63. Вікно введення формули виробничої функції

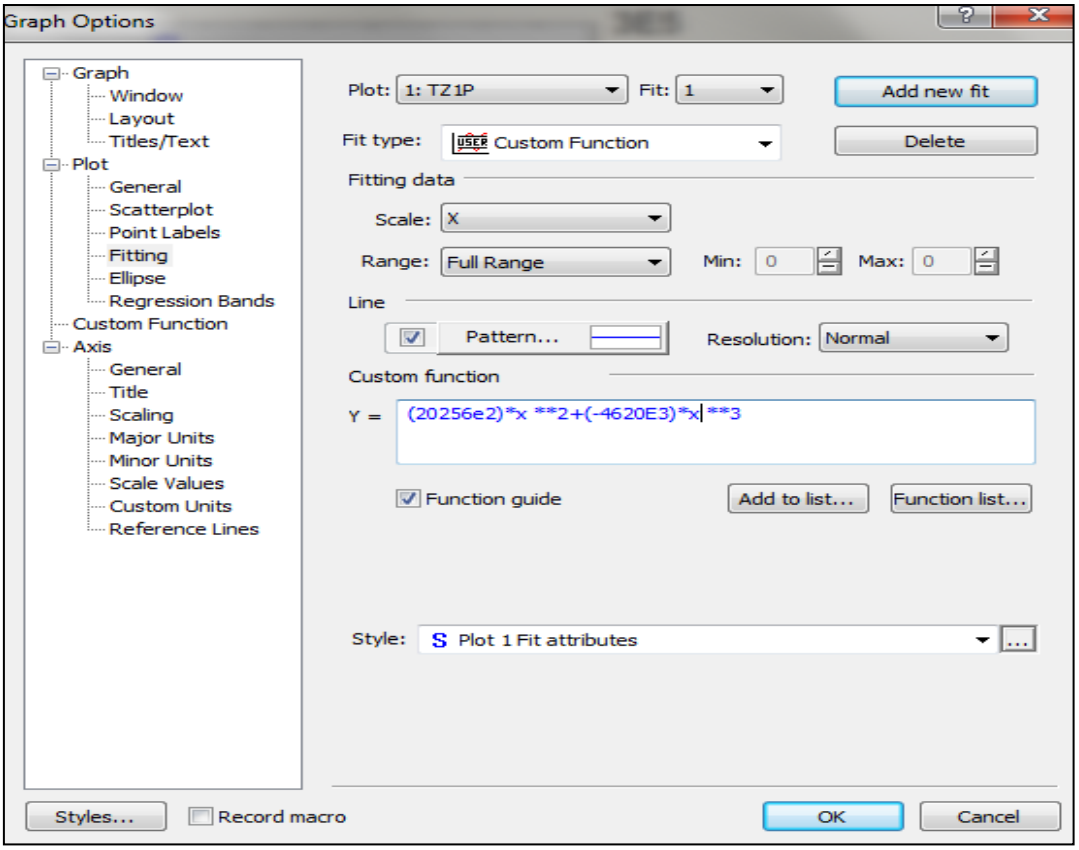


Рис. 3.64. Вікно введення формули фіскальної функції

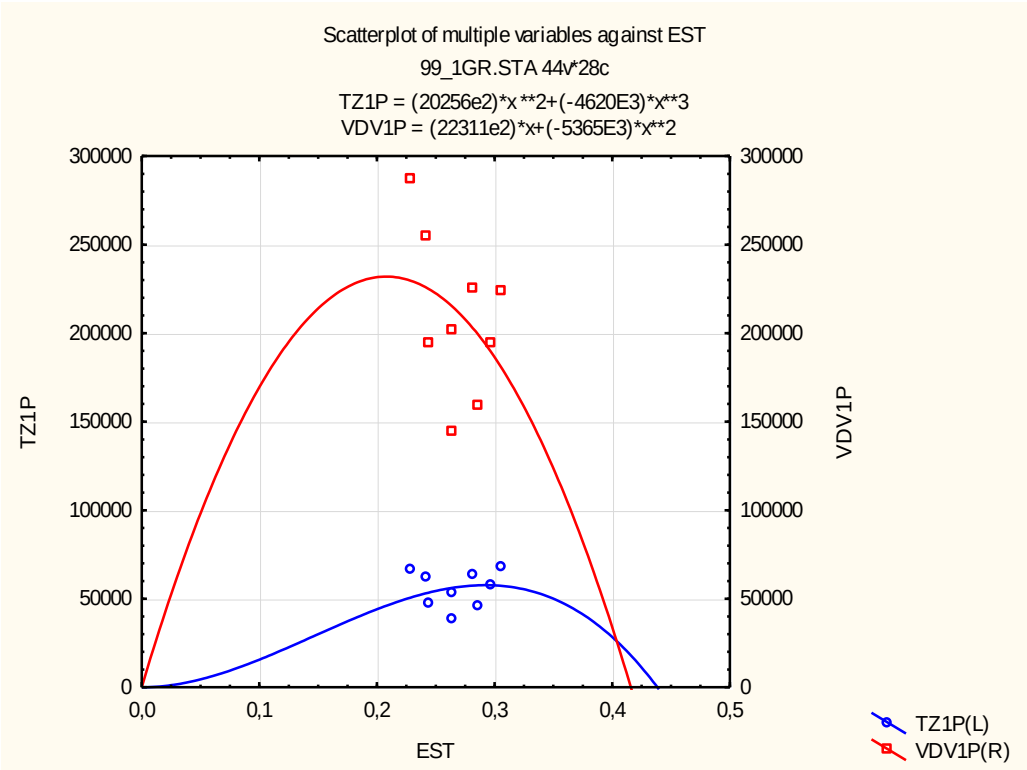


Рис. 3.65. Результат побудови двох графіків після налагодження осей X, Y та вирівнювання шкал осей Y зліва і справа

1.12. Для вирівнювання шкал осей по Y використовується налагодження осей як показано на рис. 3.66.

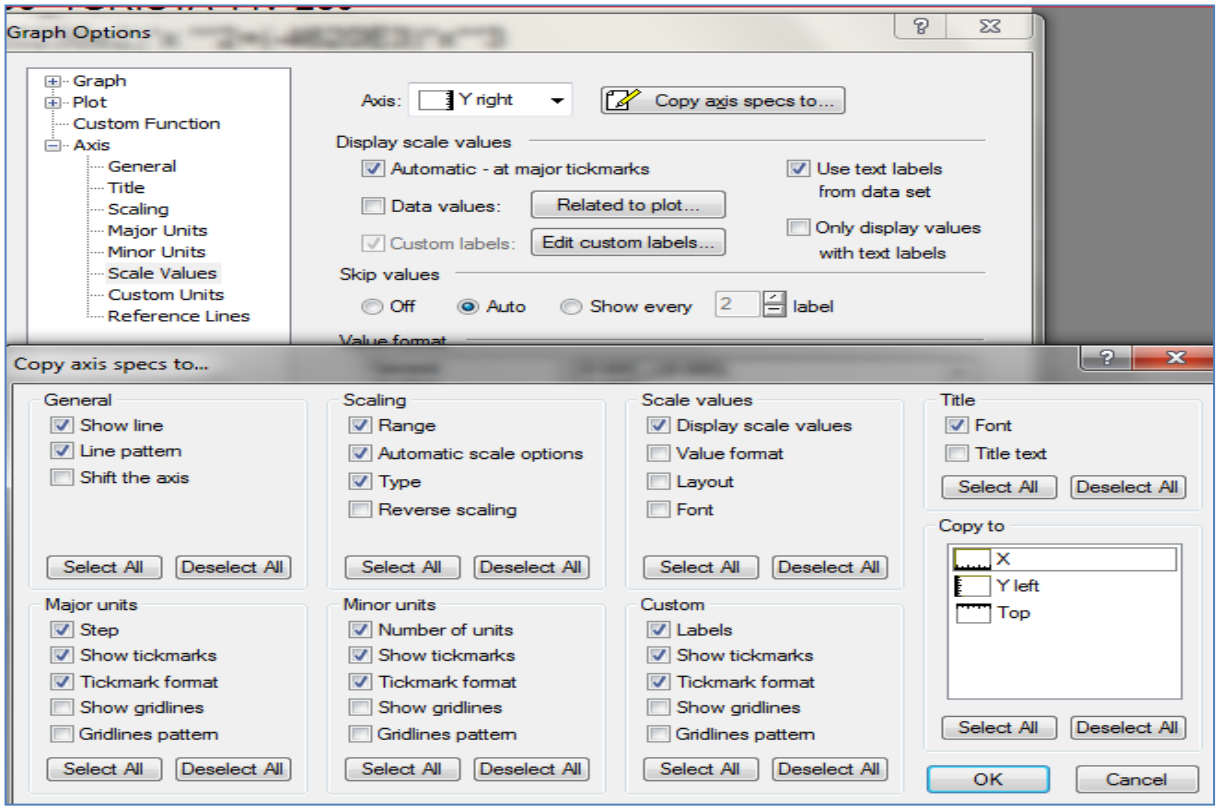


Рис. 3.66. Вирівнювання шкал осей графіків по Y

2. Аналіз результатів оцінювання параметрів точок Лаффера.

2.1. Як було зазначено в пп. 3.2.1, двопараметричний метод на відміну від інших передбачає оцінювання точок Лаффера першого *(T_j*)* та другого типу *(T_j**),* максимуми яких визначаються так:

$$T_j^{*} = -\frac{a}{2b}, \quad T_j^{**} = -\frac{2a}{3b}, \quad j = (1,2,3)$$

На рис. 3.17 для прикладу показаний вигляд виробничої та фіскальної кривих для групи регіонів у 2009 році, отриманий за вказаними процедурами у цій лабораторній роботі.

Одним із основних висновків аналізу точок Лаффера (пп. 3.2.1) роки є те, що для більшості річних регіональних даних $T_{ij}^{*} < Твдв_j < T_{ij}^{**}$, тобто податкове навантаження не стимулювало зростання виробництва.

2.2. Проведіть аналіз отриманих параметів виробничої і фіскальної функцій і зробіть висновки щодо податкового навантаження на економіку регіонів на основі отриманих значень.

Питання для самоконтролю

1. Що характеризують виробнича та фіскальна криві?
2. Що характеризують точки Лаффера першого та другого роду?
3. Які процедури здійснюються для нелінійного оцінювання параметів функцій у середовищі Statistica?
4. Які процедури здійснюються для виведення двох графіків нелінійних функцій на одному зображенні?
5. Які методи оцінювання параметрів використовуються у середовищі Statistica?
6. Поясніть суть податкового навантаження та сукупної ставки оподаткування валового регіонального продукту.
7. Які методи пошуку екстремумів функцій ви знаєте? Які із них використовують у пакетах Excel, Statistica?

Лабораторна робота 3.4
Карти Кохонена (середвище Deduktor Studio)

Мета: оторимання навичок у застосуванні самоорганізуючих карт Кохонена для задач класифікації в середовищі Deduktor Studio.

Завдання роботи: отримати класифікаційні карти Кохонена підприємств за ознаками збитковості на основі їх фінансових показників.

Порядок виконання лабораторної роботи

1. Завантажте вказаний у варіанті файл у Deduktor Studio.

Файл вміщує фінансові дані по 152 підприємствах з ознаками збитковості через 12 місяців після їх оцінювання (лабораторна робота «Дискримінантний аналіз»).

2. Для аналізу у Майстрі обробки виберіть «Карта Кохонена» (рис. 3.67.

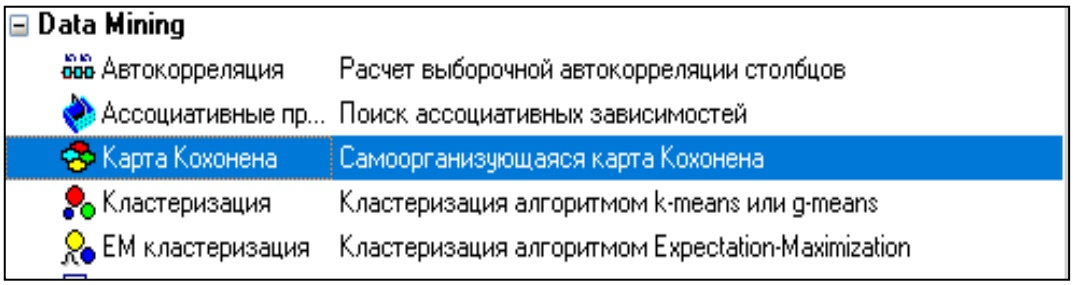


Рис. 3.67. Вибір режиму «Карти Кохонена» в Майстрі обробки

3. Для побудови «Карти Кохонена» виберіть тіж вісім показників, що бралися для Дискримінантного аналізу: покриття (kp); загальної ліквідності (kzl); абсолютної ліквідності (kal); обіговості оборотних коштів (kook), забезпечення власними оборотними коштами (kzvok); оборотності дебіторської заборгованості (kodz); оборотності кредиторської заборгованості (kokz); фінансової незалежності (kfn); фінансової стабільності (kfs) та ознаками збитковості через календарний рік після розрахунку вказаних коефіцієнтів (рис. 3.68).

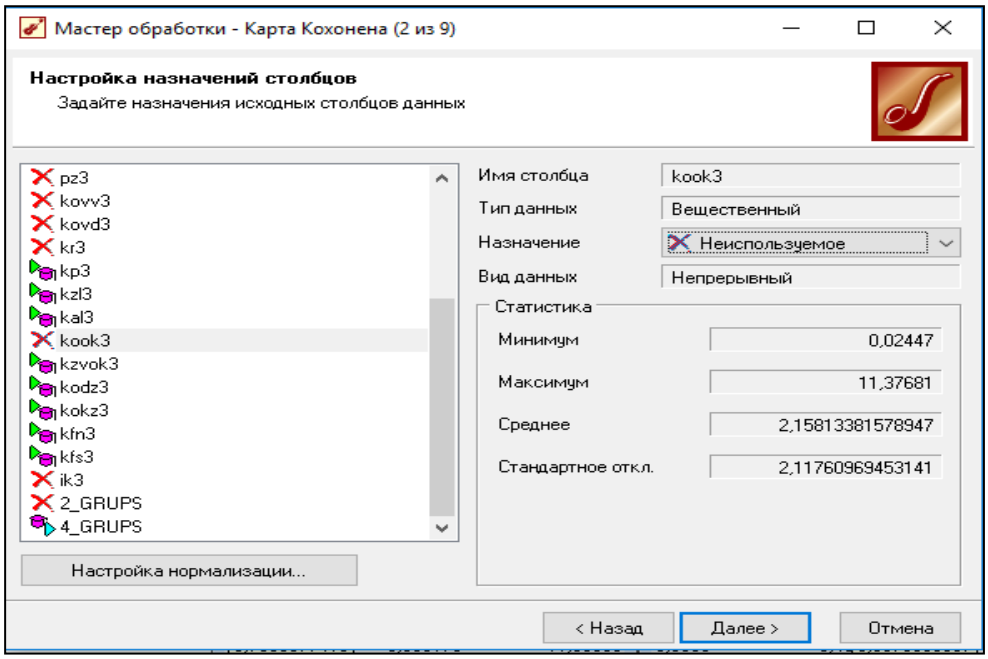


Рис. 3.68. Вибір показників для побудови карти Кохонена

Ознаки збитковості:

1) ознака підприємств, що були і залишилися збитковими (ЗЗ);

2) підприємства, що були прибутковими і стали збитковими (СЗ);

3) підприємства, що були збитковими і стали прибутковими (СП);

4) підприємства, що були прибутковими і залишилися прибутковими (ЗП).

На кроці три сформуйте навчальну та тестову вибірку (рис. 3.69).

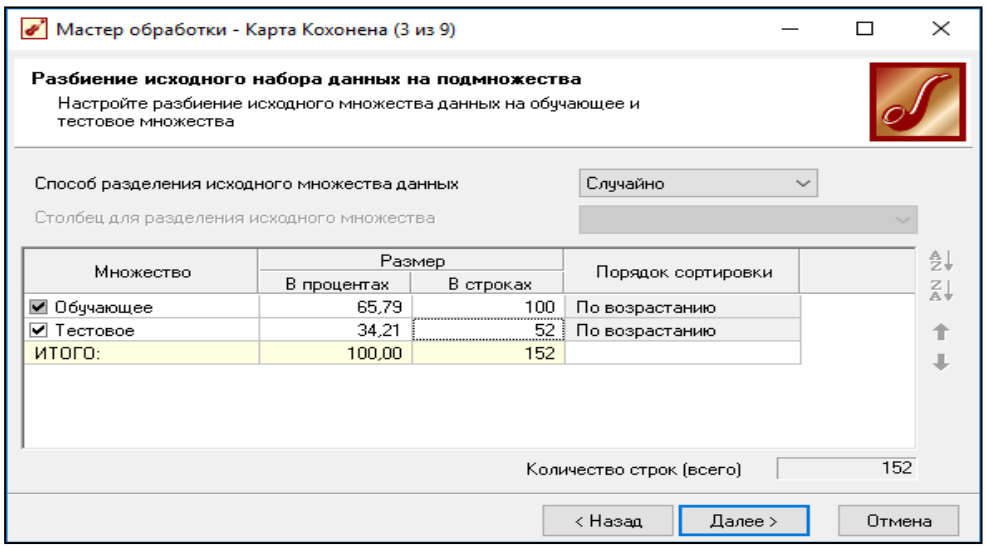


Рис. 3.69. Формування навчальної та тестової вибірки для аналізу

На кроках 5–6 вкажіть параметри зупинки навчання та параметри навчання карт Кохонена. Вкажіть, що у результаті потрібно сформувати 4-и кластери підприємств (рис. 3.70–3.71).

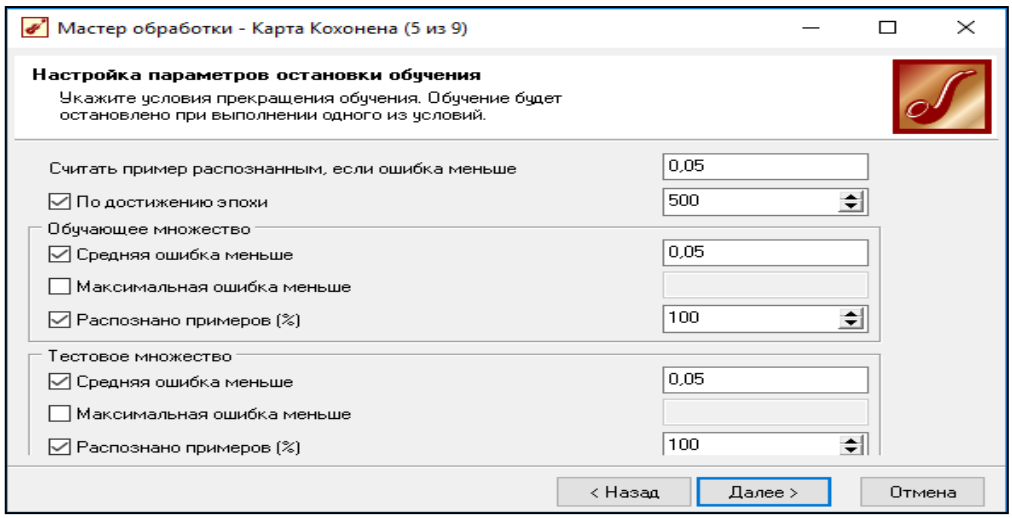


Рис. 3.70. Установка параметрів зупинки навчання

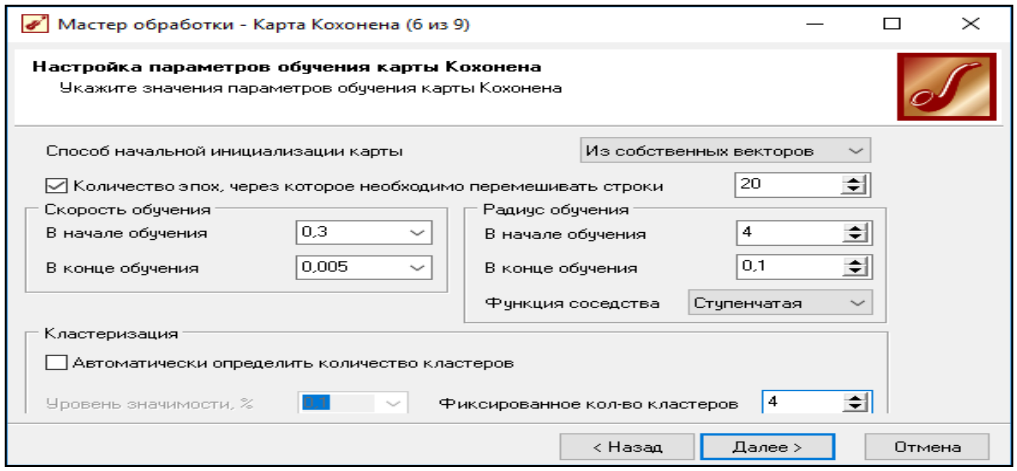


Рис. 3.71. Установка параметрів навчання карт Кохонена

На кроці 7 проведіть навчання мережі. Натисніть клавішу «Пуск» (рис. 3.72).

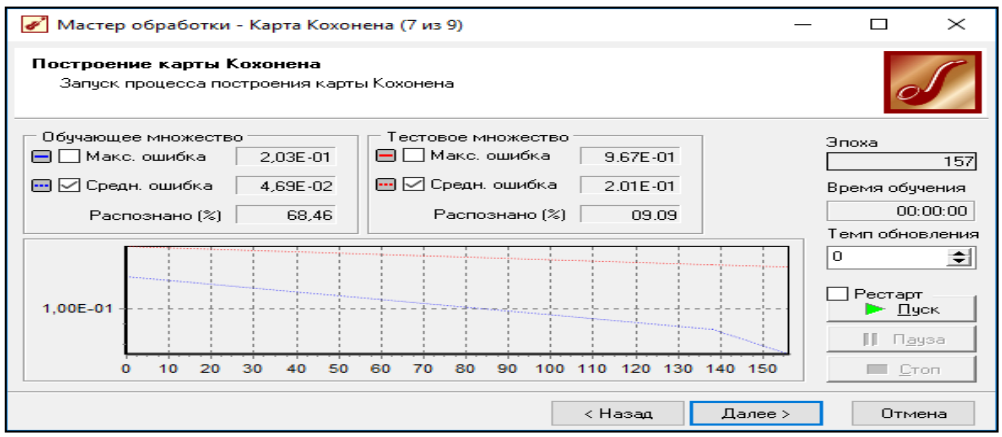


Рис. 3.72. Процес навчання карт Кохонена

На кроці 8 відмітьте способи та формати виведення інформації. У першу чергу відмітьте карти Кохонена та інші, вказані на рис. 3.73.

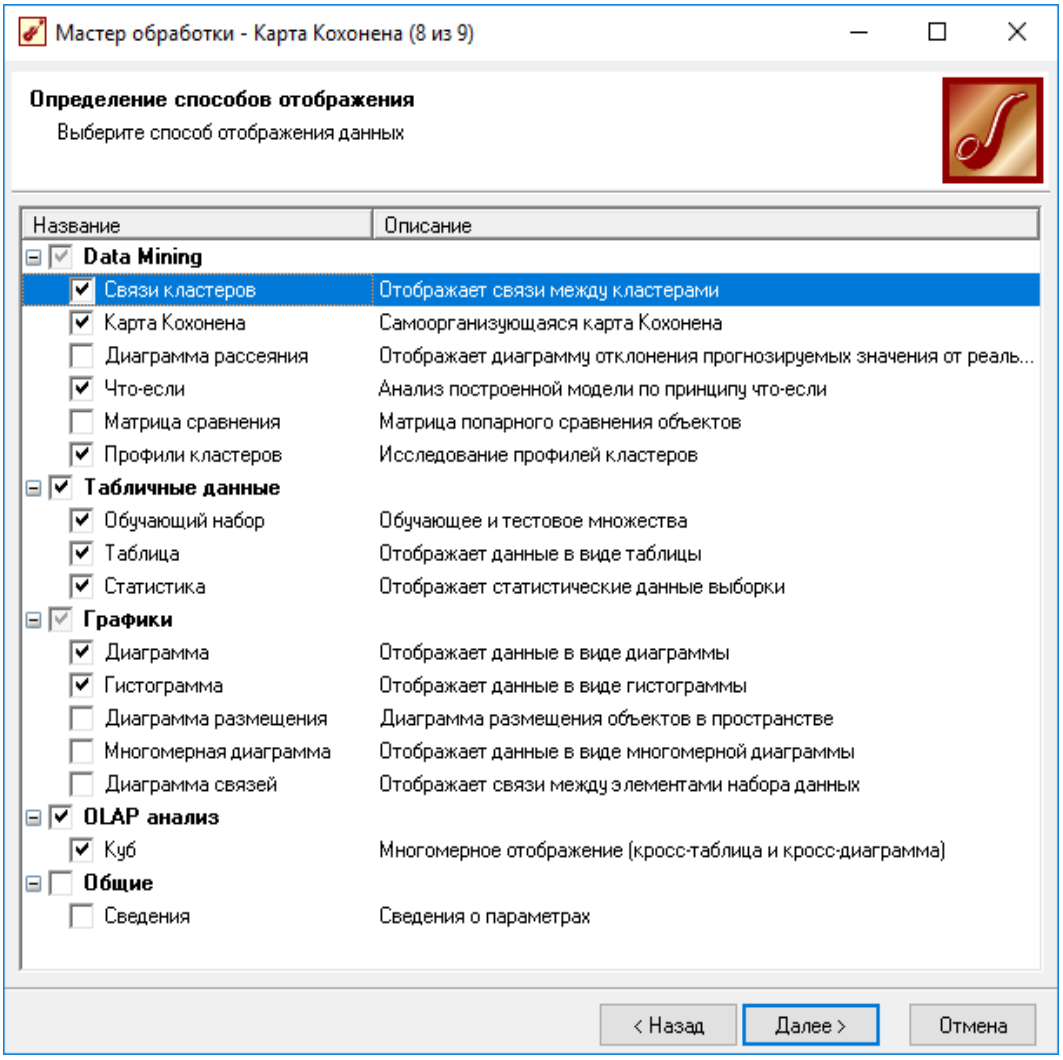
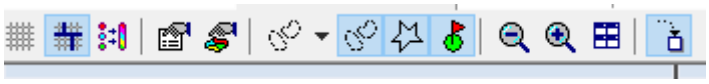


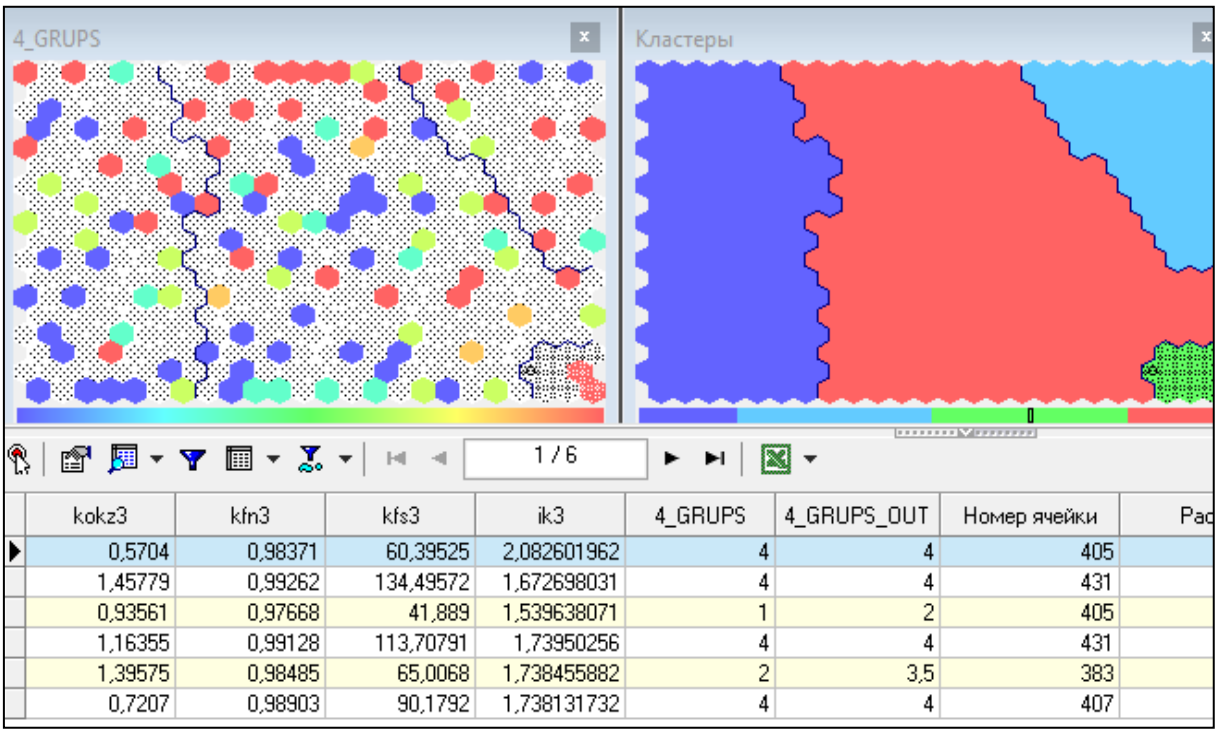
Рис. 3.73. Вибір способів та форматів виведення інформації

Аналіз карт Кохонена: поточна комірка Карти відображається на карті маленькою окружністю чорного кольору. Змінити поточну комірку просто: клацнути мишею в потрібну ділянку карти. Внизу кожного відображення на градієнтній шкалі в жовтому прямокутнику відображається числове значення ознаки, відповідне її кольору.

Під час роботи з картою користуйтеся операціями, що виконуються за допомогою кнопок на панелі інструментів візуалізатора або контекстного меню, що викликається правою кнопкою миші в будь-якому вікні карти.



Виділіть комірки кластера у правому нижньому куті вікна «Кластери» і дослідіть показники підприємств та групи збитковості, для яких вини віднесені (рис. 3.74) (інструкція до базових лабораторних робіт з Deductor Studio).



kokz3	kfn3	kfs3	ik3	4_GRUPS	4_GRUPS_OUT	Номер ячейки	Рас
0,5704	0,98371	60,39525	2,082601962	4	4	405	
1,45779	0,99262	134,49572	1,672698031	4	4	431	
0,93561	0,97668	41,889	1,539638071	1	2	405	
1,16355	0,99128	113,70791	1,73950256	4	4	431	
1,39575	0,98485	65,0068	1,738455882	2	3,5	383	
0,7207	0,98903	90,1792	1,738131732	4	4	407	

Рис. 3.74. Отримані карти Кохонена

На отриманих картах Кохонена (рис. 3.75):

1. Проаналізуйте отримані кластери підприємств за фінансовими показниками.

2. Порівняйте фактичний розподіл підприємств на чотири групи за ознаками збитковості та отримані кластери підприємств (Карти Кохонена: «4_GRUPS», «Кластери»).

4. Сформуйте узагальнений опис підприємств кожного кластера, дайте висновки (для аналізу використайте іншу виведену інформацію за результатом аналізу).

Підготуйте звіт про виконання лабораторної роботи і розмістіть його для перевірки у СДН Moodle.

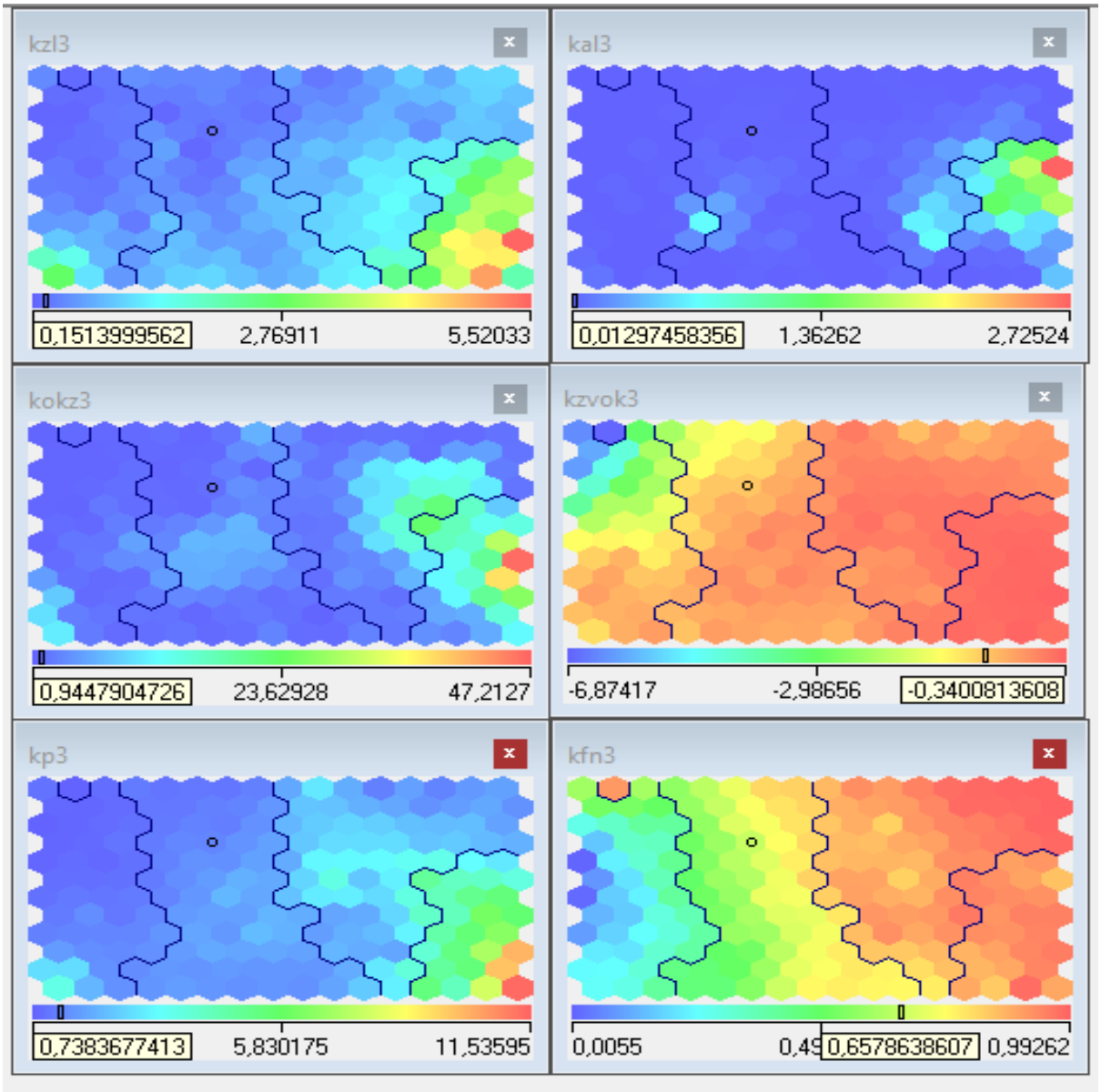


Рис. 3.75. Карти Кохонена із групуванням підприємств у чотири кластери у розрізі фінансових коефіцієнтів

Питання для самоконтролю

1. Поясніть суть методу навчання мереж Кохонена.
2. Чим відрізняються карти Кохонена від інших типів нейромереж?
3. Як правильно у відсотковому співвідношенні сформувати навчальну та тестову вибірки?
4. У чому відмінність класифікації за фактом стану збитковості підприємства і за результатами навчання карт Кохонена?

Лабораторна робота 3.5
Дослідження методів класифікації в середовищі Oracle Data Miner

Мета: отримання навичок у роботі із класифікаційними методами Oracle Data Miner.

Завдання роботи: провести порівняльний аналіз класифікаційних методів та сформувати пропозиції щодо їх використання у адмініструванні податків.

Порядок виконання лабораторної роботи

Попередньо на сервері БД Oracle повинен бути створений користувач із правами доступу до схем Oracle Data Miner.

Користувачу, що працює з Oracle Data Miner, необхідно видати деякі додаткові привілеї. Привілея CREATE MINING MODEL необхідна для тренування моделі. Крім того, він повинен мати можливість створити таблицю та представлення (REATE TABLE, CREATE VIEW), щоб виконати необхідні операції з тренування моделей. У цій роботі будуть створені процедури (CREATE PROCEDURE) для генерації даних. Користувач повинен мати привілеї для підключення до баз (CREATE SESSION) і квоти на простір у табличному просторі (або UNLIMITED TABLESPACE).

1. Підготовка даних для моделювання

1.1. Надамо користувачу відповідні права. Скрипт надання привілеїв такому користувачу (за умовчанням DMUSER) показний нижче:

```
GRANT CREATE SESSION, CREATE TABLE, CREATE VIEW,
CREATE MINING MODEL, CREATE PROCEDURE, UNLIMITED
TABLESPACE TO DMUSER IDENTIFIED BY ORACLE;
```

1.2. При запущеному Oracle SQL Developer виберіть у меню Tools/DataMiner/MakeVisible.

1.3. Далі виберіть опцію «Створити новий проект / NewProject» (рис. 3.76).

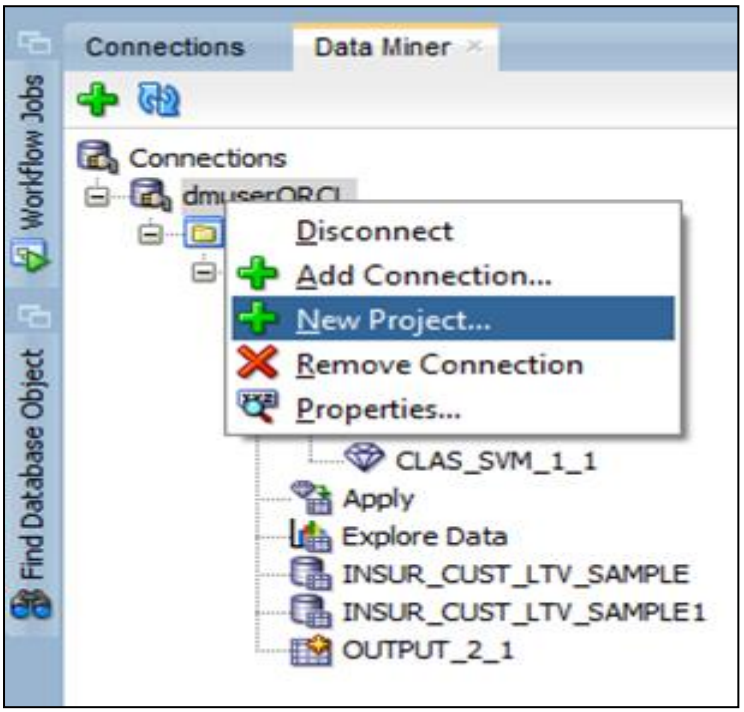


Рис. 3.76. Вікно створення нового проекту

1.4. У вікнах, що з'являться, задайте ім'я проекту Klassification і робочого простору Workflow, наприклад «WorkKlass». Після цього відкриється робочий простір WorkKlass (рис. 3.84). Зліва буде дерево, що являє собою проект. Справа – набір інструментів. У центрі – робоче місце, де будемо створювати класифікаційні моделі.

1.5. Дані будемо брати із баз даних Oracle. Для представлення цих даних додайте на робоче місце «Джерело даних/Data/ DataSourse» і виберіть таблицю **CUST_INSUR_LTV_SAMPLE** з усіма полями.

1.6. Для перегляду даних їх дослідження (розвідувального аналізу) використайте модуль «Data/Explore Data». З'єднайте модулі між собою, натиснувши праву клавішу і вибравши опцію Connect (рис. 3.77).

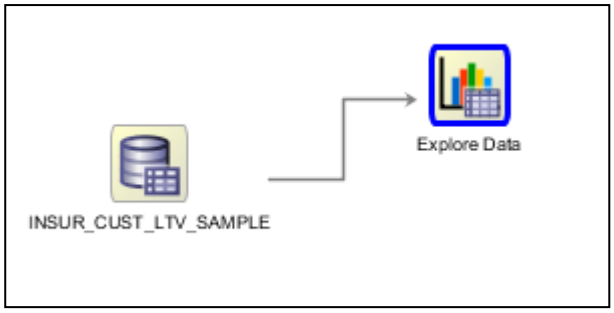


Рис. 3.77. З'єднання модулів між собою

1.7. Припустимо, що ми хочемо завантажити дані BUY_INSURANCE. Для цього два рази натисніть Explore Data і виберіть групування за BUY_INSURANCE (рис. 3.78).

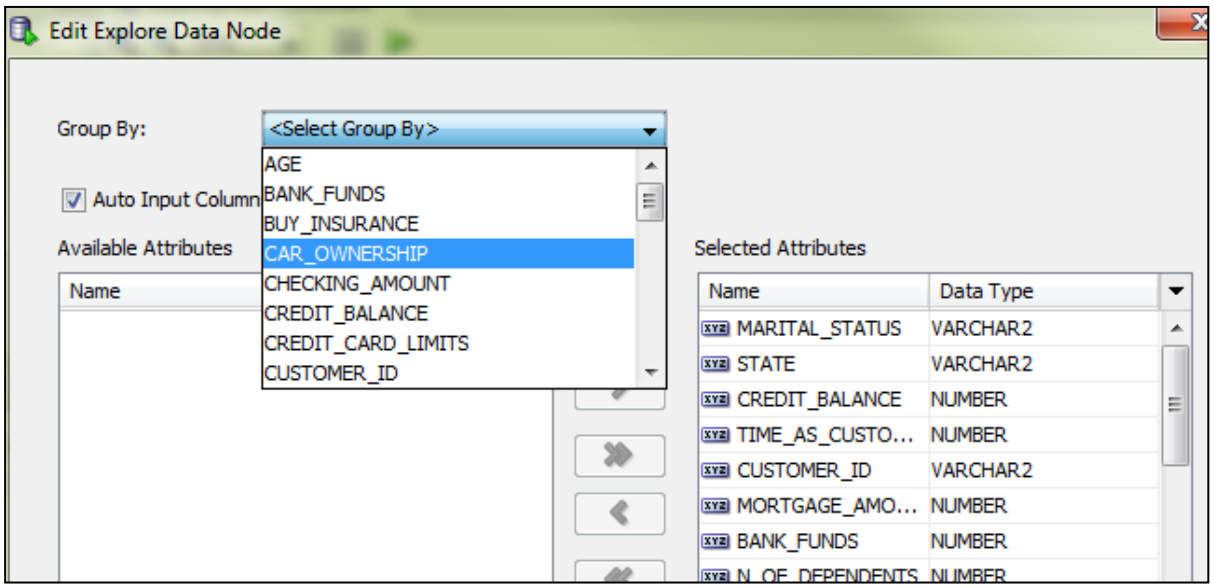


Рис. 3.78. Завантаження даних до робочого простору

1.8. Виберіть правою клавішею Run і після закінчення роботи Workflow у модулі Explore Data виберіть View Data.

1.9. Ознайомтеся із інформацією, яку Data Miner надає по кожному атрибуту з прив'язкою до атрибута, обраного в Group By. Серед цієї інформації є гістограма значень, кількість різних значень, середнє, мінімальне і максимальне значення, відхилення від середнього, дисперсія, коефіцієнти асиметрії та ексцесу (рис. 3.79).

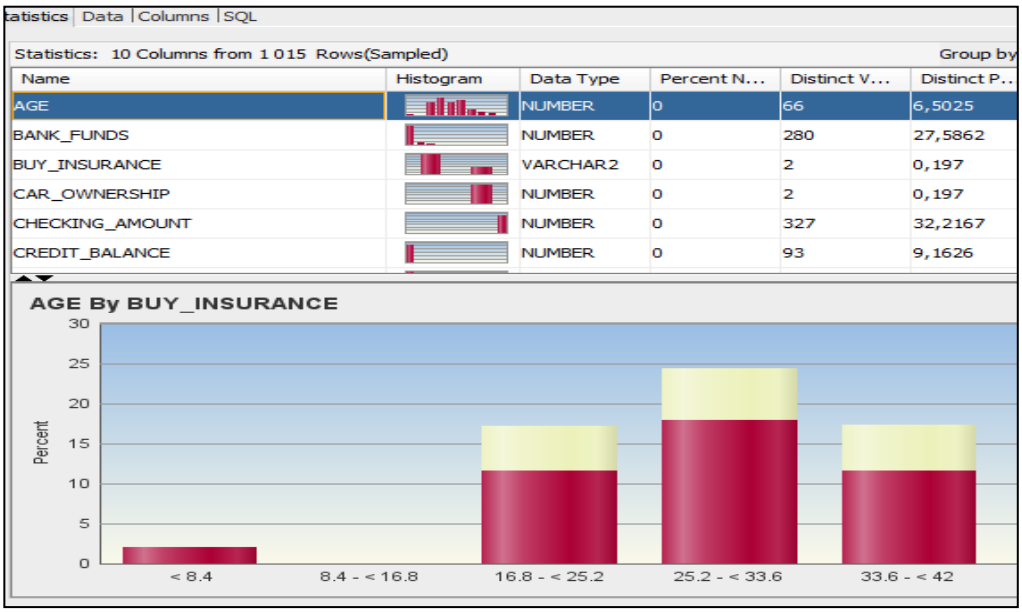


Рис. 3.79. Інформація, що надається при прогляді даних модулі «Explore Data», «View Data»

2. Створення класифікаційної моделі

2.1. На прикладі Oracle за даними страхової компанії щодо покупки страхового полісу клієнтами здійснимо оцінювання ймовірності придбання цієї послуги, сформуємо рекомендації щодо вибору методу класифікації та застосування його у адмініструванні податків.

2.2. Перетягнемо класифікаційний модуль (Classification) з розділу Models палітри інструментів і підключимо її до джерела даних за допомогою інструменту link.

2.3. У діалоговому вікні виберемо як цільовий атрибут BUY_INSURANCE. А як Case Id вибирають первинний ключ, тобто CUSTOMER_ID. Залишимо активними всі чотири доступних алгоритми класифікації (рис. 3.80).

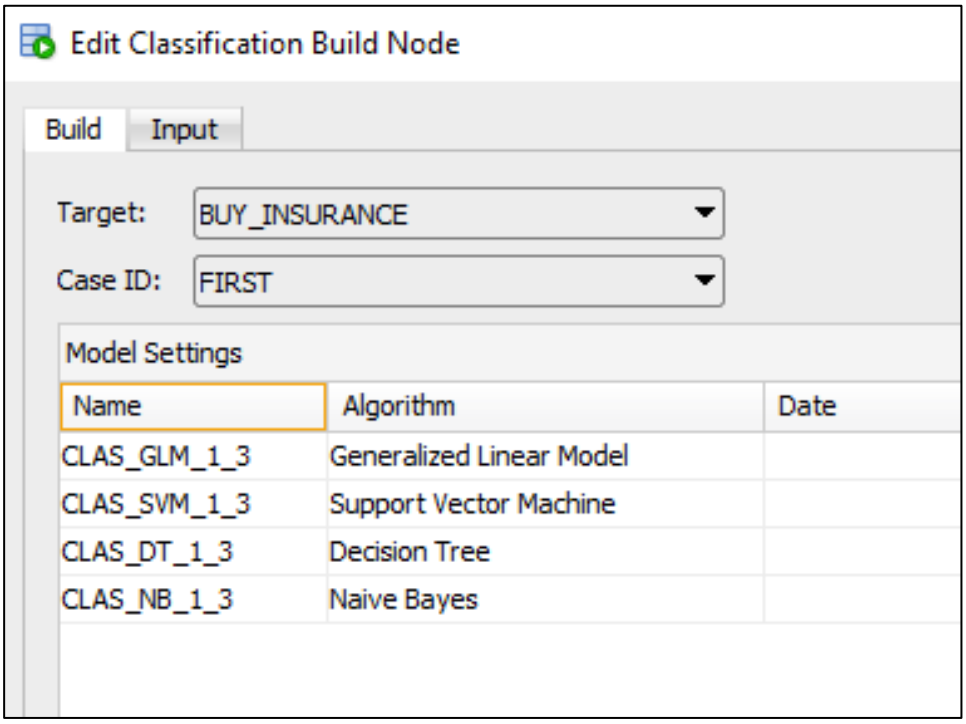


Рис. 3.80. Вибір методів класифікації

2.4. Натиснувши на кнопку Advanced ..., можна зробити розширене налаштування алгоритмів. Виберемо алгоритм Support Vector Machine і закладку Algorithm Settings змінимо на Kernel Function на Linear (рис.3. 81).

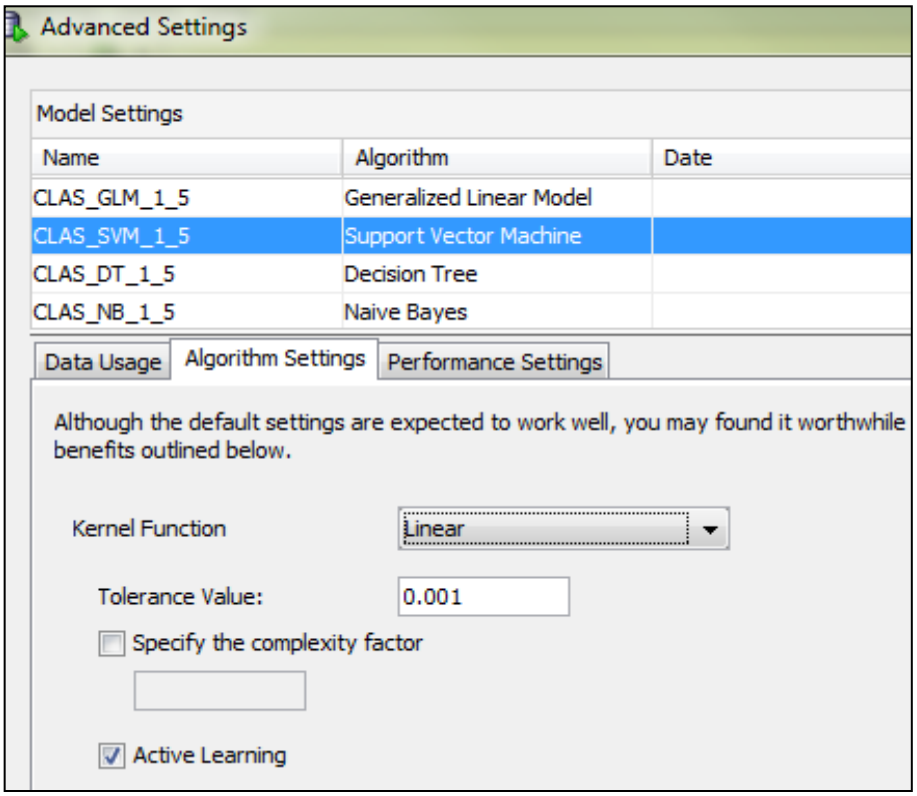


Рис.3. 81. Вікно розширеного налаштування алгоритмів

2.5. Далі редагуємо властивості моделей. Закриємо вікна і виберемо модель на діаграмі. Внизу з'явиться редактор властивостей моделі (рис. 3.82).

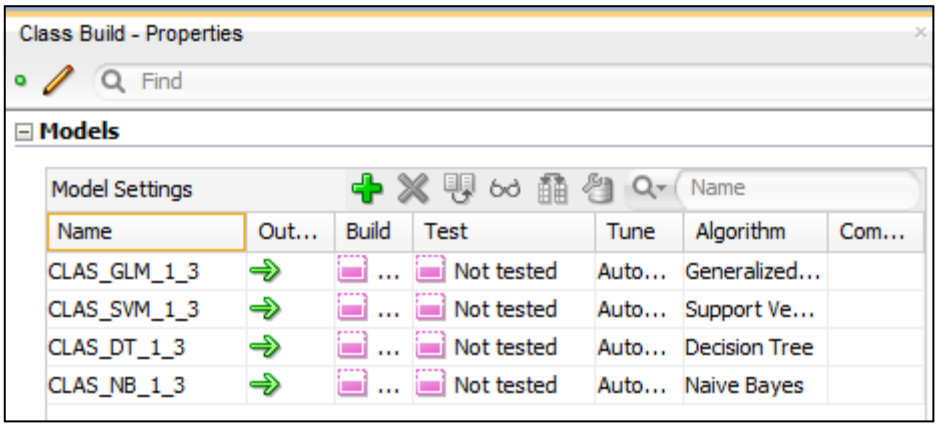


Рис. 3.82. Вікно редагування властивостей моделі

3. Побудова та тренування моделі

3.1. Побудуємо моделі на базі набору даних. Ця операція називається «тренування» або навчання моделі (про модель кажуть, що вона навчається на цих даних). Стандартною практикою є побудова моделі на навчальному наборі даних, а потім перевірка цієї моделі на тих відомих даних, що залишилися. За замовчуванням ODM використовує цей підхід.

3.2. Перед початком побудови моделі перейдемо на закладку Test у властивостях моделі і вкажемо, що потрібно виділити 50 % даних для тесту (рис. 3.83).

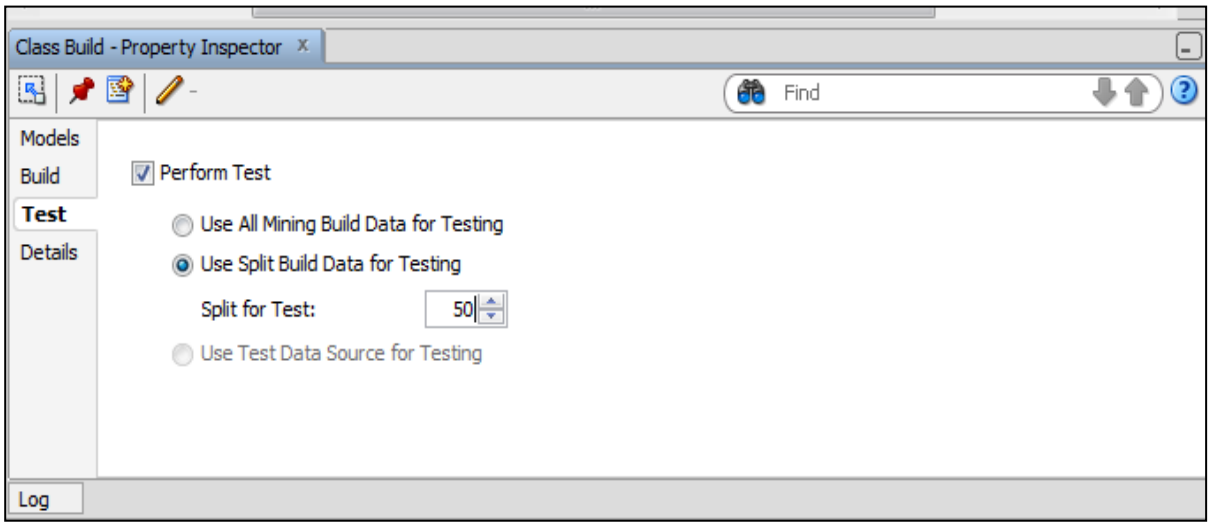


Рис. 3.83. Встановлення параметрів навчального набору даних

3.3. Клікнемо на моделі правою кнопкою і запустимо її побудову (Run). Дочекаємося її завершення. Про це просигналізує поява зеленої галочки біля піктограми Class Build і додаткової інформації в «Інспекторі властивостей» моделі:

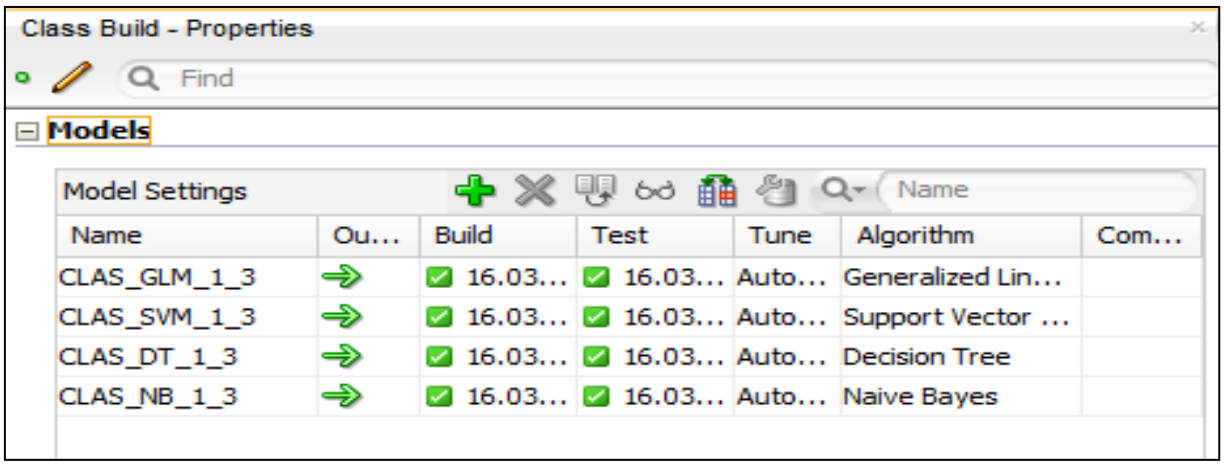


Рис. 3.84. Відображення успішної побудови моделей

3.4. У результаті побудовані 4 моделі. Всі вони будувалися для прогнозування цільового параметра BUY_INSURANCE. Вихідні дані автоматично були розділені для побудови і тестування моделі.

4. Порівняння моделей

4.1. Побудовані моделі відрізняються своєю точністю – проаналізуємо результати їх тестування. Клікнемо на моделі правою клавішею і виберемо Compare Test Results. Відкриється вікно (рис. 3.85).

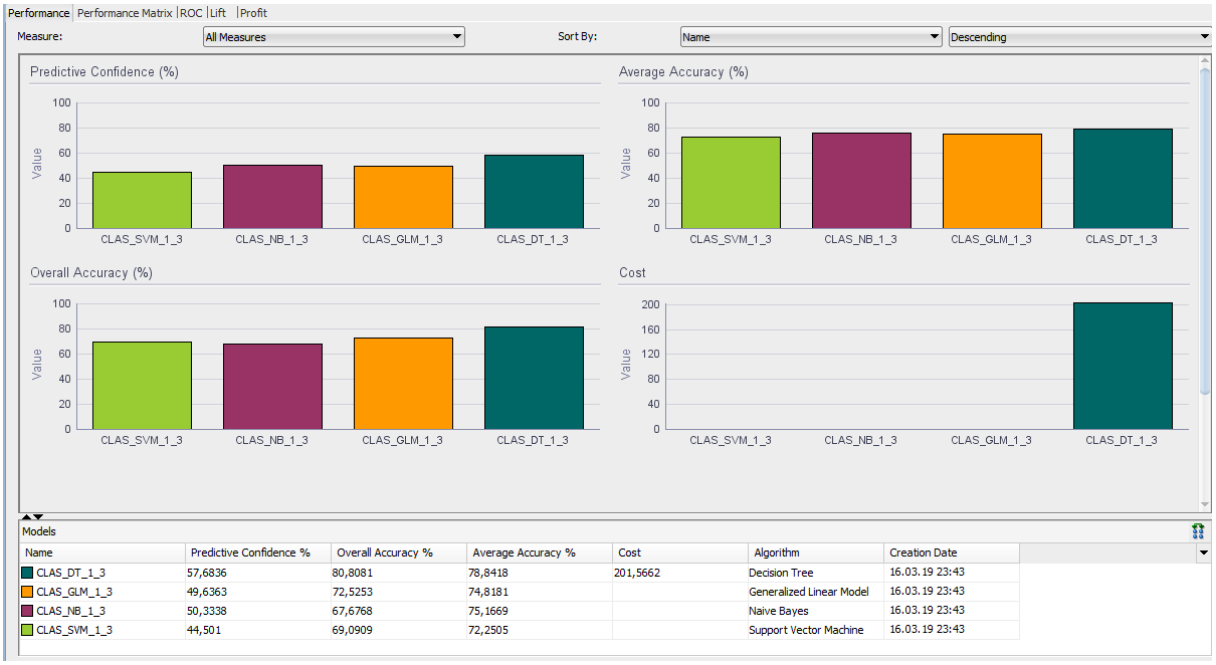


Рис. 3.85. Результати тестування моделей

4.2. Набір даних маленький, тому цифри у вас можуть трохи відрізнятися від зазначених на рис. 3.85. Кольори також можуть відрізнятися. Результати порівняння моделей включає 5 закладок: Performance, Performance matrix, ROC, Lift, and Profit. Закладка Performance показує числову і графічну інформацію за такими показниками: Predictive Confidence, Average Accuracy і Overall Accuracy. Як видно із цієї закладки, Decision Tree (DT) дає найкращі показники за всіма параметрами.

4.3. Для аналізу швидкості пошуку моделлю цільових значень виберемо закладку Lift (рис. 3.86). Закладка Lift дає графічне представлення lift для кожної моделі. Червона лінія показує результати для моделі з випадковим вибором. Вертикальна синя лінія показує поріг.

Lift – це ще один тип тесту моделі, який показує, наскільки «швидко» модель знайде реальні позитивні цільові значення.



Рис. 3.86. Вікно аналізу швидкості моделі

Вікно перегляду Lift порівнює lift кожної моделі. Вікно перегляду Lift відображає Cumulative Positive Cases і Cumulative Lift

У нашому прикладі:

У 0.20-квантилі модель Decision Tree забезпечує найбільшу Cumulative Lift – трохи більше ніж модель GLM. Можна переміщати точку вимірювання квантилі по осі X. Дані будуть автоматично оновлюватися. Переміщаючи повзунок квантилі, можна побачити як Cumulative Lift моделей GLM і SVM будуть злегка змінюватися. У 0.52-квантилі моделі NB і GLM будуть давати приблизно такий Cumulative Lift, що і DT.

4.4. Виберемо закладку Performance Matrix для ознайомлення із точністю класифікації моделями (рис. 3.87).

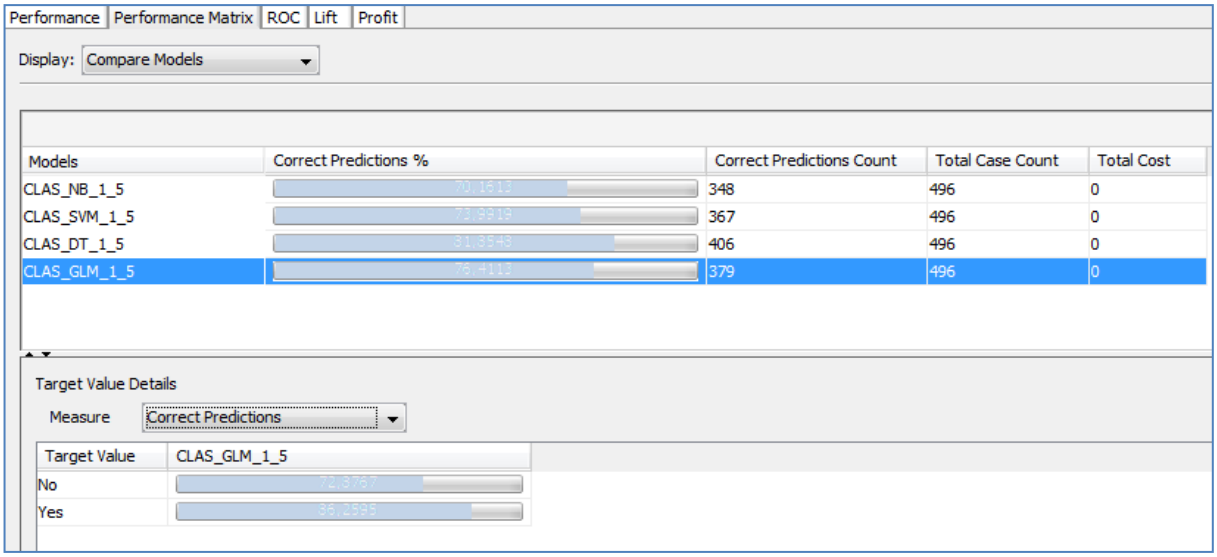


Рис. 3.87. Закладка «Performance Matrix» – точність класифікації

Можна бачити, що моделі GLM і DT дають більшу кількість коректних прогнозів – близько 80 %. Якщо вибрати по черзі GLM і DT, то можна побачити, що DT краще прогнозує відповідь No, а GLM – Yes. Відповідно для класифікації суб'єктів господарювання рекомедується застосовувати метод Decision Trees. В адмініструванні податків цей метод і Oracle Data Miner зокрема можна застосовувати для прогнозування порушення платниками податків законодавства, прогнозування обсягів бази оподаткування, оцінювання податкових ризиків.

4.5. Підготуйте звіт про виконання лабораторної роботи і розмістіь його для оцінювання в СДН Moodle.

Питання для самоконтролю

1. Які моделі інтелектуального аналізу інтегровані в Oracle Data Miner?

2. Які привілеї повинен мати користувач БД Oracle для доступу до репозитарію Oracle Data Miner?

3. Які основні етапи створення нового проекту в Oracle Data Miner?

4. Які моделі класифікації інгегровані в Oracle Data Miner?

5. Яка модель класифікації показує більшу точність? Чому це відбувається на ваш погляд?

6. У якому режимі можна порівняти точність результатів моделей?

7. Що означає поняття Lift?

8. Як можна пояснити графіки Lift кривих?

Лабораторна робота 3.6
Дослідження побудованих моделей Decision Trees і застосування їх для прогнозування поведінки платників податків

Мета: отримання навичок у використанні класифікаційних моделей для прогнозування.

Завдання роботи: використати отриману класифікаційну модель у лабораторній роботі 3.4 і використати її для прогнозування поведінки клієнтів.

Порядок виконання лабораторної роботи

1. Завантаження отриманої моделі у лабораторній роботі 3.4

1.1. Під'єднайтеся до БД Oracle за допомогою Oracle SQL Developer за своїми обліковими даними.

1.2. Виберіть у меню Tools/DataMiner/MakeVisible. Запустіть свій проект, створений на попередній лабораторній роботі.

2. Перегляд конкретної моделі

2.1. У Workflow правою кнопкою кликнемо на Class Buid і виберемо View Models-> CLASS_DT_1_5 (рис. 3.88).

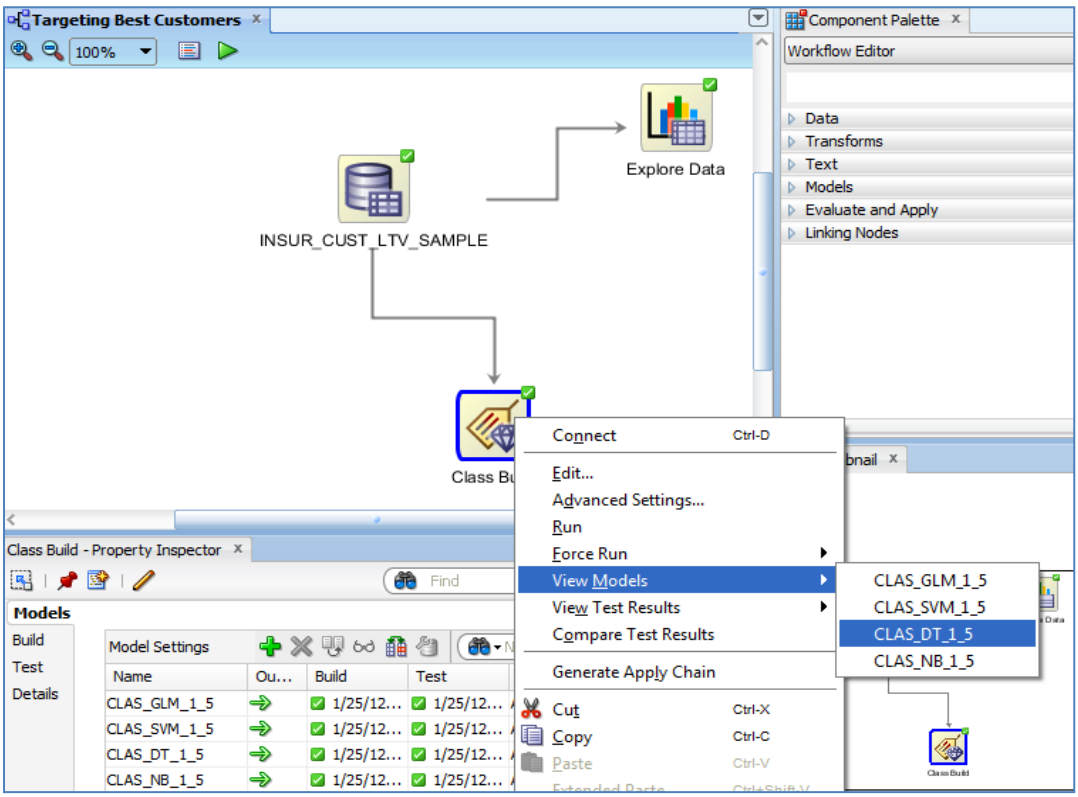


Рис. 3.88. Вікно вибору режиму прогляду параметрів моделі

2.2. Відкриється вікно з деревом рішень. Виберіть вузол Node 2 (рис. 3.89). На кожному рівні в дереві рішень буде представлена фраза IF / THEN, що описує правило. На кожному наступному рівні будуть додавати нові умови в цю фразу IF / THEN. Зазвичай, це дерево рішень буде набагато глибше, тому має сенс згортати деякі вузли.

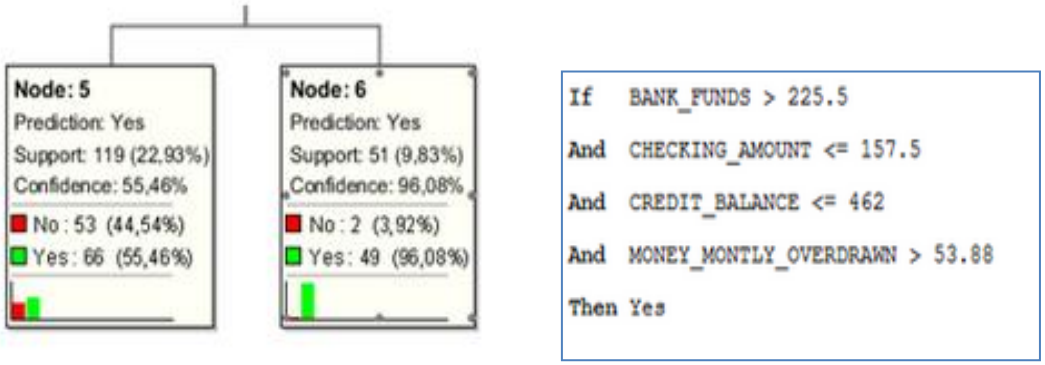


Рис. 3.89. Дерево рішень. Вузли (Node) 2, 8

Для вузла Node 2 можна бачити, що перший поділ засновано на атрибуті BANK_FUNDS, а друге – CHECKING_AMOUNT. Node 2 говорить, що якщо BANK_FUNDS більше ніж 225,5, а CHECKING_AMOUNT менше ніж 157.5, то ймовірність покупки страхування – 61.38 %.

2.3. Виберемо вузол Node 6 (рис. 3.90).

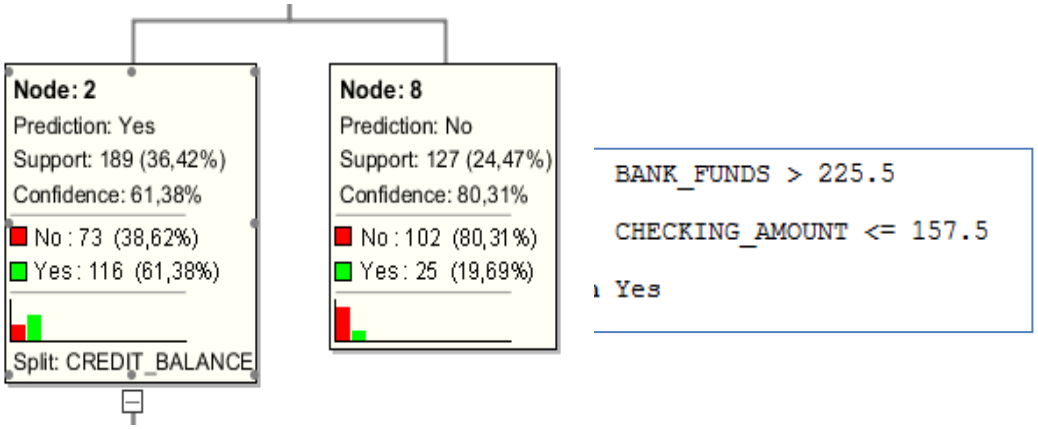


Рис. 3.90. Дерево рішень. Вузли (Node) 5, 6

Для вузла Node 6 можна бачити, що перший поділ засновано на атрибуті CHECKING_AMOUNT, другий на CREDIT_ BALANCE, третій – на MONEY_OVERDRAWN. Node 6 говорить, що якщо CHECKING_AMOUNT <= 157,5, CREDIT_ BALANCE <= 462, MONEY_OVERDRAWN >лісу, то ймовірність покупки страхування – 96,08 %.

3. Застосування моделі

Для застосування моделі виконуються такі етапи:

– вибирається модель (або моделі) в вузлі Class Build;

– доповнюються новим джерелом даних;

– додається вузол Apply в Workflow;

– пов’язується джерело і вузол Apply.

3.1. У Workflow вибираємо вузол Class Build. На закладці Models вибираємо тільки модель DT. Для цього натискаємо на велику зелену стрілку image в стовпці Output моделі, щоб вона стала стрілкою з червоним хрестиком image. І так для всіх моделей, крім DT (рис. 3.91).

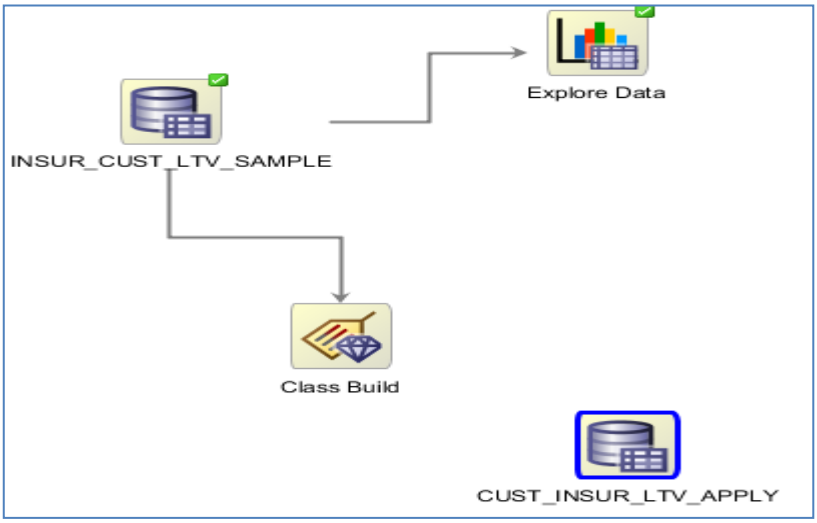


Рис. 3.91. Вибір моделі для застосування

3.2. Створюємо нове джерело даних на базі тієї ж таблиці DMUSER.INSUR_CUST_LTV_SAMPLE. Потім перейменуємо джерело в CUST_INSUR_LTV_APPLY (рис. 3.92).

Model Settings

Name	Output	Build	Test	Tune
CLAS_GLM_1_5		1/25/12 5:18 PM	1/25/12 5:18 PM	Automatic
CLAS_SVM_1_5		1/25/12 5:18 PM	1/25/12 5:18 PM	Automatic
CLAS_DT_1_5		1/25/12 5:18 PM	1/25/12 5:18 PM	Automatic
CLAS_NB_1_5		1/25/12 5:18 PM	1/25/12 5:18 PM	Automatic

Рис. 3.92. Створення нового джерела даних для моделі

3.3. Створимо вузол Apply за допомогою інструменту Apply з закладки Evaluate and Apply (рис. 3.93).

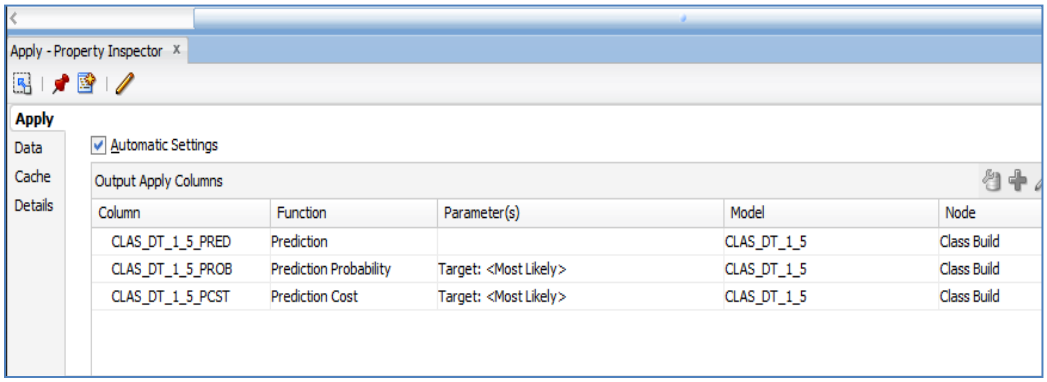


Рис. 3.93. Створення вузла «Застосування моделі» («Apply»)

3.4. Перед запуском вузла apply в workflow варто глянути список стовпців, які з'являться у результаті (рис. 3.94).

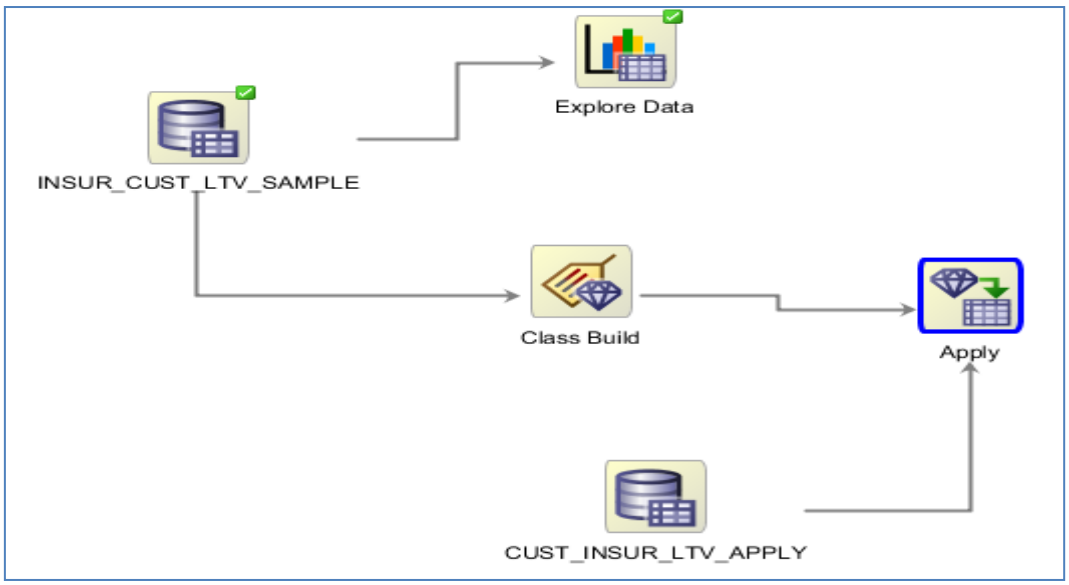


Рис. 3.94. Прогляд стовпців, які з'являться у результаті

Можна бачити, що в результатах буде:

– прогноз (Yes or No);

– імовірність прогнозу;

– вартість.

3.5. Для того, щоб ця інформація була корисна, в результати потрібно включити CUSTOMER_ID. Для цього натискаємо правою кнопкою на Apply Model і вибираємо Edit.

3.6. На закладці Data Columns додамо стовпець CUSTOMER_ID (рис. 3.95).

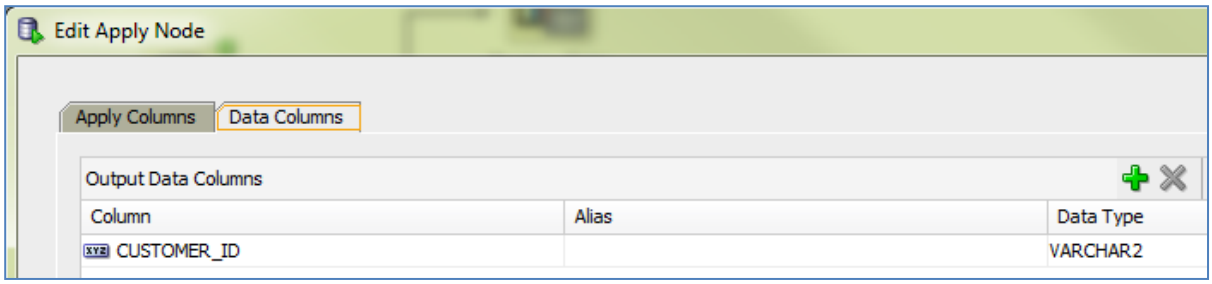


Рис. 3.95. Додавання стовпця CUSTOMER_ID

3.7. Застосуємо модель. Для цього правою кнопкою натиснимо на модель і виберемо Run. Дочекаємося поки з'явиться зелена галочка, що свідчить про успішне застосування моделі.

3.8. Щоб подивитися результат, можна натиснути правою кнопкою і вибрати View Data.

3.9. Тут на закладці SQL можна набрати команду, яка використовується для застосування моделі – дуже корисна функція.

3.10. Результат роботи алгоритму можна помістити в таблицю. Для цього використовуємо інструмент Create Table or View. З'єднаємо його з вузлом Apply (рис. 3.96).

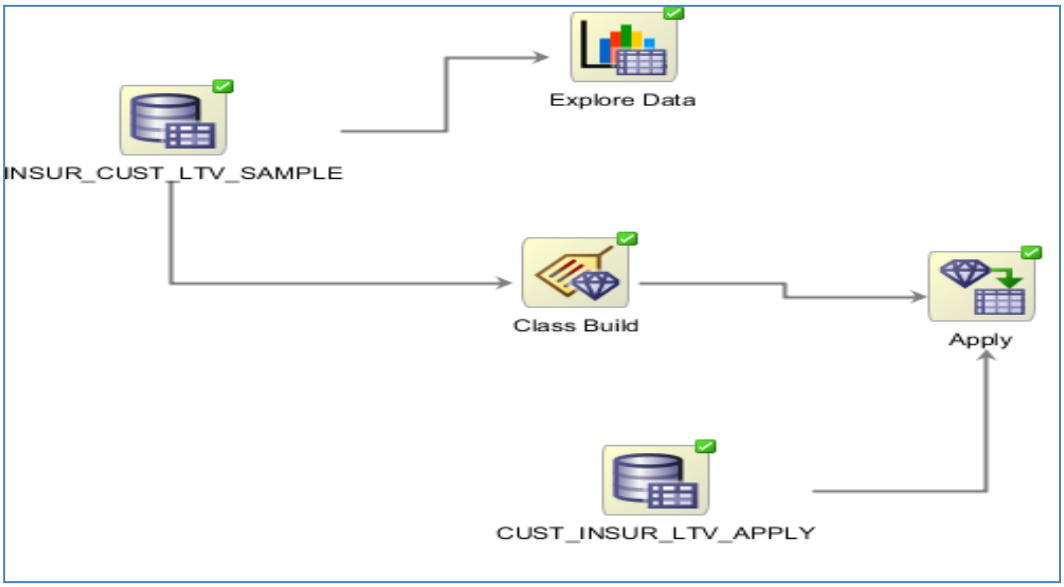


Рис. 3.96. Застосування інструменту Create Table or View для виведення результатів моделювання у таблицю

3.11. Результатом може бути не таблиця, а уявлення (можна поміняти в налаштуваннях вузла Create Table or View), що дозволить кожен раз динамічно обчислювати результати (рис. 3.97).

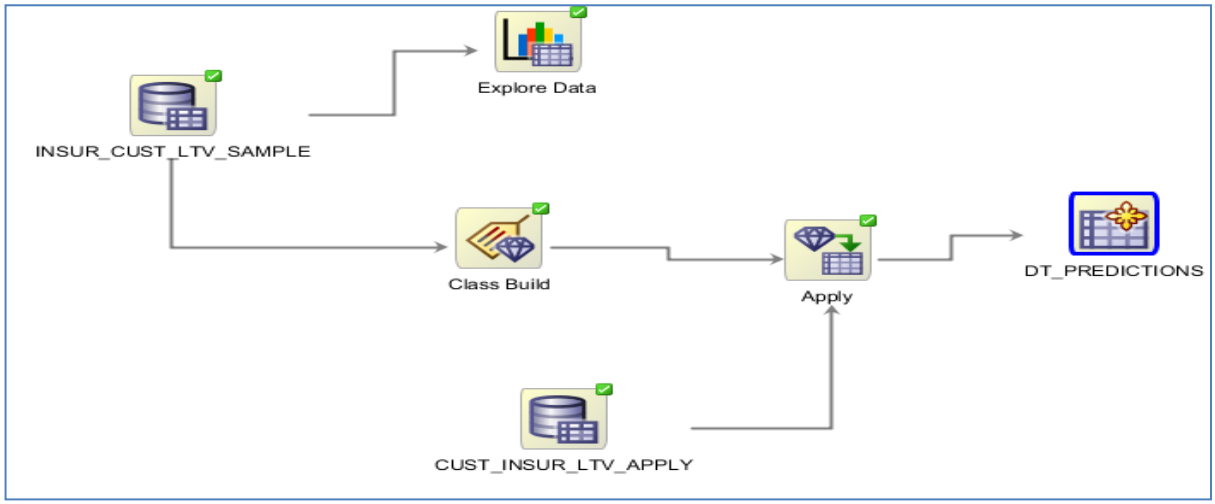


Рис. 3.97. Таблиця із виведеними результатами моделювання

3.12. Після додавання вузла Create Table or View повторно запустимо Workflow. У результаті в базі даних з'явиться таблиця DT_PREDICTIONS, що містить дані, які бачили раніше (рис. 3.98). Це дуже зручно. Вміст таблиці (або подання) може бути використано для виведення в Oracle Application Express, Oracle BI Answers, Oracle BI Dashboards тощо.

Columns | SQL

View: Actual Data | Sort... | Filter: Enter Where Clause

	CUSTOMER_ID	CLAS_DT_1_5_PRED	CLAS_DT_1_5_PROB	CLAS_DT_1_5_PCST
1	CU7854	No	0,9951	0,018
2	CU12993	No	0,9951	0,018
3	CU608	No	0,8031	0,7195
4	CU10025	No	0,9951	0,018
5	CU11680	No	0,9951	0,018
6	CU10706	No	0,9951	0,018
7	CU6752	No	0,8031	0,7195
8	CU6932	No	0,9951	0,018

Рис. 3.98. Робочий простір Data Miner із доданою таблицею DT_PREDICTION

3.13. Проведіть класифікаційний аналіз платників податків відповідно до даних вашого варіанта, розміщеного на СДН Moodle, для чого:

– визначте класифікаційну ознаку;

– визначте показники, що можна використати для класифікації платників податків за класифікаційною ознакою;

– сформуйте гіпотезу про можливі варіанти класифікації;

– проведіть класифікацію підприємств, проаналізуйте точність моделі, сформуйте висновки.

3.14. Підготуйте звіт про виконання роботи і розмістіть його для оцінювання у СДН Moodle.

Питання для самоконтролю

1. Перерахуйте основні процедури налагодження моделей, що інтегровані в Oracle Data Miner.

2. Запишіть на мові SQL команду надання привілеїв користувачу БД Orcle для доступу до репозитарію Oracle Data Miner?

3. Які основні компоненти проекту в Oracle Data Miner?

4. Поясніть математичну суть моделей класифікації, що інтегровані в Oracle Data Miner?

5. Що означає якість моделі, специфікація моделі, більша точність?

6. Які дані підприємств використовувались для їх класифікації щодо ймовірності купівлі послуги?

7. Запропонуйте застосування класифікаційних методів та Oracle Data Miner у адмініструванні податків, зокрема у контрольно-перевірочній роботі.

8. Які дані підприємств бажано мати для класифікації їх за ступенем податкових ризиків з точки зору податкового відомства?

ПРЕДМЕТНИЙ ПОКАЖЧИК

База даних (БД), сховище даних (СД) – структура даних, що дозволяє отримувати, нагромаджувати чи видавати інформацію на запити численних незалежних користувачів (ДСТУ 2505-94, п. 3.7.15).

Віртуальна приватна мережа (VPN – Virtual Private Network) – віртуальні канали зв'язку, які за допомогою з'єднання через пакетну мережу (Укртелеком) забезпечують об'єднання в єдину приватну мережу приміщень клієнта (ДПС), що територіально віддалені один від одного.

Доступність мережі провайдера – відношення часу знаходження компонентів мережі в робочому стані до загальної довжини інтервалу спостерігання (місяць).

Клієнт – організація, що орендує послуги, які надає провайдер.

Маршрутизатор – блок взаємодії, який забезпечує вибір маршруту пересилання даних між кількома ЛОМ, які мають різну архітектуру чи протоколи (ДСТУ 2229-93, п.85).

Мітка – короткий ідентифікатор фіксованої довжини, який має значення на локальному відрізку мережі для визначення FEC. На сьогодні стандартом визначений формат 32-бітної мітки, яка розміщується між заголовками другого рівня (Layer 2) і третього рівня (Layer 3).

Пакетна мережа MPLS провайдера – мережева інфраструктура провайдера послуг зв'язку (Укртелеком), яка забезпечує побудову віртуальних приватних мереж на основі технології багатопротокольної комутації по мітках (MPLS - Multi Protocol Label Switching), яка складається з опорних маршрутизаторів (Provider router (P), граничних маршрутизаторів (Provider Edge router (PE), а також магістральних каналів, що з'єднують ці маршрутизатори.

Сайт – сукупність підмереж, які об'єднані за допомогою маршрутизаторів.

Сховище даних (Data Warehouse) – база даних, що містить дані більш ніж з одного джерела, які зібрані для управління інформацією.

Провайдер – комерційна організація, що надає послуги зв'язку (Укртелеком).

Трафік – сукупність даних, переданих по мережі MPLS провайдера.

Трафік клієнта – сукупність даних, отриманих на порту клієнта (вхідний трафік) або надісланих з порту клієнта (вихідний трафік).

Фізичне середовище – абстрактне представлення середовища пересилання даних (ДСТУ 2229-93, п.133).

Fibre Channel – високошвидкісний волоконно-оптичний канал.

Frame Relay – високошвидкісна технологія передачі кадрів, яка передбачає ділення даних на кадри змінної довжини.

Автоматизована система митного оформлення (АСМО) – підсистема, що входить до складу Єдиної автоматизованої інформаційної системи Держмитслужби і забезпечує використання електронних документів та електронних копій документів на паперових носіях посадовими особами митного органу для здійснення митних процедур.

Автоматизована інформаційна система «Центр» (далі – АІС «Центр») – автоматизована інформаційна система центрального рівня, яка становить сукупність підсистем, комплексів задач, програмно-інформаційних комплексів та централізованого сховища зведеної інформації про митне оформлення – електронних копій вантажних митних декларацій (далі – ВМД) на паперовому носії та електронних вантажних митних декларацій (далі – ЕВМД), інших електронних документів та електронних копій документів на паперовому носії, які використовуються під час здійснення митного контролю та оформлення товарів і транспортних засобів, що переміщуються через митний кордон України, та забезпечує двосторонній обмін митною та іншою інформацією з автоматизованими системами митного оформлення та іншими системами (підсистемами) локального рівня.

Дворівнева архітектура (модель) – схема структурної організації взаємодії двох частин програмного забезпечення – клієнтського та серверного. Залежно від того, як між ними розподіляються функції, розрізняють:

модель «тонкого клієнта», в рамках якої вся логіка застосування та управління даними зосереджена на сервері. Клієнтська програма забезпечує тільки функції рівня представлення (відображення) інформації кінцевому користувачу;

модель «товстого клієнта», в якій сервер насамперед виконує функції зі збереження даних, а основна обробка інформації та інтерфейс користувача зосереджені з боку клієнта.

Єдина автоматизована інформаційна система Держмитслужби (далі – ЄАІС Держмитслужби) – багатофункціональна інтегрована автоматизована система, що забезпечує інформаційну підтримку та супроводження митної справи в Україні і становить сукупність кількох взаємопов'язаних інформаційних систем, зокрема, автоматизованої інформаційної системи «Центр», автоматизованої системи митного оформлення, інформаційно-телекомунікаційної системи «Електронна пошта» та інших систем, програмно-інформаційних комплексів.

Інформаційні ресурси баз даних – визначені групи взаємопов'язаних задокументованих одиниць інформації, які об'єднані в інформаційних та інформаційно-телекомунікаційних системах митних органів, спеціалізованих митних установ та організацій, описують процеси митного контролю та оформлення товарів і транспортних засобів, контролю за переміщенням вантажів, що перебувають під митним контролем, як через митний кордон України, так і в її межах, а також інші процеси діяльності митних органів.

Інцидент – будь-яка подія, що відбувається в процесі функціонування інформаційної системи, яка може призвести до погіршення якості роботи системи або до її повної зупинки;

Обробка інформації – виконання однієї або кількох операцій, зокрема: збирання, введення, записування, перетворення, зчитування, зберігання, знищення, реєстрація, приймання, отримання, передавання, які здійснюються у системі за допомогою технічних і програмних засобів.

ПІК ЄІАС Держмитслужби – програмно-інформаційні комплекси ЄАІС Держмитслужби, які призначені для обробки інформації в ЄАІС Держмитслужби та побудовані за дворівневою або трирівневою архітектурою.

Трирівнева архітектура (модель) – схема структурної організації певної програмної системи, яка складається з трьох частин:

– **«тонкого клієнта»**, який направляє інформаційний запит та отримує інформацію від сервера додатків (становить просте клієнтське програмне забезпечення, що забезпечує лише передачу даних, їх локальне кешування, представлення засобами інтерфейсу користувача, редагування і найпростішу обробку);

– **сервера додатків**, який приймає інформаційні запити клієнтів і інформацію від сервера бази даних, обробляє їх відповідно до запрограмованих правил, у разі необхідності перетворює у форму, зручну для сервера бази даних або для «тонкого клієнта»;

– **сервера баз даних**, який виконує отримані від сервера додатків інформаційні запити та відправляє результати до сервера додатків, що адресує дані тонкому клієнтові.

Служба активного каталогу – реалізація служби каталогів корпорації Microsoft для операційних систем сімейства Windows.

Служба каталогу – програмний засіб представлення інформації про ресурси інформаційної системи (користувачі, загальні папки, сервери баз даних, друку та інше) в ієрархічному вигляді, який зберігається в єдиному місці, що забезпечує централізоване управління як самими ресурсами, так і інформацією про них.

Сховище даних – сукупність структурованих та формалізованих інформаційних ресурсів баз даних, об’єднаних для проведення обробки даних з подальшим формуванням статистичних форм звітності та прийняття рішень у потаткових та митних органах для контролю та оформлення товарів і транспортних засобів, що переміщуються через митний кордон країни, контролю за переміщенням вантажів та інших процесів, що стосуються митної та подткової справи, діяльності митних та податкових органів, спеціалізованих митних установ, організацій.

ПОЗНАЧЕННЯ ТА СКОРОЧЕННЯ

ATM (Asynchronous Transfer Mode)	—	технологія асинхронного режиму передачі пакетів фіксованого розміру.
BGP (Border Gateway Protocol)	—	граничний міжмережевий протокол, який використовується для обміну маршрутною інформацією тільки з тими маршрутизаторами, які зазначені в його таблиці маршрутизації.
CE (Customer Edge router)	—	маршрутизатор клієнта, для підключення до магістралі провайдера.
DWDM	—	технологія передачі інформації, що використовує мультиплексування по довжині хвилі високої щільності.
FEC (Forwarding Equivalency Class)	—	сукупність пакетів, які обслуговуються в мережі однаковим способом. Пакети, що належать одному FEC у домені MPLS, просуваються загальним маршрутом – по одному LSR. FEC (клас еквівалентного пересування), до якого приписаний пакет, може бути однозначно ідентифікований міткою, розташованою в цьому пакеті.
IGP (Interior Gateway Protocol)	—	протокол внутрішнього шлюзу, який використовується для обміну інформацією між сумісно працюючими маршрутизаторами.

Intranet — технологія створення корпоративної локальної мережі підвищеної надійності з обмеженим доступом, яка використовує мережні стандарти та мережні програмно-апаратні засоби, аналогічні Internet.

IPSec (IP Security) — протокол наскрізної передачі даних (тільки відправник та кінцевий одержувач повідомлення можуть читати інформацію, що передається) з використанням алгоритмів кодування та аутентифікації.

IS-IS (Intermediate System to Intermediate System) — зв'язок між проміжними системами, за цим протоколом проміжні системи обмінюються інформацією про маршрутизацію.

ISDN (Integrated Services Digital Network) — цифрова мережа зв'язку з комплексними послугами.

LAN — локальна обчислювальна мережа.

Label Distribution Protocol (LDP) — протокол розповсюдження міток.

Label Switch Path (LSP) — точний маршрут прямування пакетів через мережу MPLS. Побудова маршрутів LSP забезпечується спеціальними протоколами розповсюдження міток (LDP), такими як RSVP-TE і/або CR-LDP.

Label Switch Router (LSR)	—	маршрутизатор з комутацією по мітках.
MAN	—	мережа масштабу міста.
NFS	—	мережна файлова система.
OSPF (Open Shortest Path First)	—	першочергове відкриття найкоротших маршрутів, алгоритм маршрутизації, при якому шлях вибирається на основі інформації про стан каналу (link state), є стандартом IGP для Internet.
P (Provider router)	—	Внутрішньомагістральні опорні маршрутизатори мережі MPLS.
PE (Provider Edge router)	—	пограничні маршрутизатори мережі MPLS.
RFC (Requests for Comments)	—	серія документів IETF, що розпочата в 1969 році та включає опис набору протоколів Internet і зв’язану з ними інформацію.
RIP (Routing Information Protocol)	—	протокол маршрутної інформації, стандарт IGP.
PPP (Point-to-Point Protocol)	—	протокол двоточкового з’єднання.
Proxy	—	Модуль доступу до засобів мережі Internet, за допомогою якого одна комунікаційна система представляє іншу у відповідь на запити Http-протоколу.

RSVP-TE і CR-LDP	—	механізми сигналюваня для керування трафіком у магістралі (MPLS Traffic Engineering). Призначені для організації процесу розповсюдження міток з використанням явної маршрутизації.
SAN (Storage Area Network)	—	мережа збереження даних.
SAP-(service access point)	—	інтерфейс, на якому провайдер надає послугу і гарантує узгоджений з клієнтом рівень якості послуги.
TCP\IP (Transmission Control Protocol/Internet Protocol)	—	протокол керування передачею даних.
VLAN	—	віртуальна локальна мережа.
VoIP (Voice over IP)	—	передача голосу по IP-протоколу.
VPN	—	віртуальна приватна мережа.
WAN	—	обчислювальна мережа, яка забезпечує передачу інформації на значні відстані.
X.25	—	протокол доступу до мереж з комутацією пакетів.

10/100 Mbps Ethernet.	—	специфікація широкополосної LAN з пропускною спроможністю 10 Mbps та 100 Mbps.
КСЗІ	—	комплексна система захисту інформації.
МКМ	—	мультисервісна корпоративна мережа.
ПЗ	—	програмне забезпечення.
ТМЗК	—	телефонна мережа загального користування.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Концепція розвитку цифрової економіки та суспільства України на 2018–2020 роки. Схвалена розпорядженням Кабінету Міністрів України від 17 січня 2018 р. № 67-р. – Режим доступу : https://zakon.rada.gov.ua/laws/show/67-2018-%D1%80 (Дата зверенення: 10.11.2018).

2. Бамбізов Є. Є. Інформаційна функція Державної фіскальної служби України: напрями реалізації. / Є. Є. Бамбізов. – Режим доступу : URL: http://goal-int.org/informatsijna-funktsiya-derzhavnoyi-fiskalnoyi-sluzhbi-ukrayini-napryami-realizatsiyi (Дата звернення 15.08. 2018).

3. Пантала О. П. Інформаційна система «Податковий блок» – ключове досягнення проекту «Модернізація державної податкової служби-1» / О. П. Пантала : матеріали науково-практичного семінару [«Модернізація державної податкової служби України у контексті інноваційного розвитку»]. – Ірпінь: Національний університет ДПС України, 2013. – 146, [22–31] с.

4. Вітлінський В. В. Методологічні основи управління ризиками у адмініструванні податків / В. В. Вітлінський, О. В. Редич //. Науковий вісник НУДПСУ. – 2009, № 3 (46). – С. 107–113. – Режим доступу : http://ir.nusta.edu.ua/ jspui/handle/doc/364 (Дата звернення 15.08. 2018).

5. Oleksandr Redych. (2019) Decision-Making in the Management of the Tax Authorities on theBasis of Integrated Performance Indicators. International Journal of Innovative Technologies in Economy. 3(23). doi: 10.31435/rsglobal_ijite/ 30042019/6454. С. 32–40.

6. Редич О. В. Погляд на податкову систему України як на складну саморегулюючу економічну систему / О. В. Редич // Бюджетна система: тези доповідей науково-практичної конференції, (м. Ірпінь, травень 2003 р.). – Національна академія ДПС України. – С. 124–126. – Режим доступу : http://ir.nusta.edu.ua/jspui/ handle/doc/386 (Дата зверенення: 09.10.2018).

7. Редич О. В. Управління податковим відомством на основі ключових показників ефективності / О. В. Редич // УДФСУ: економічний вісник. – 2018. – Випуск № 2. – С.169–177. – (Серія «Фінанси, облік, оподаткування»).

8. Редич О. В. Суб'єктивна оцінка податкового навантаження як фактора ризику економічної безпеки України / О. В. Редич, В. В. Романенко // Теоретичні та прикладні питання економіки: збірник наукових праць; за заг. ред. проф. Ю. І. Єханурова, проф. Г. М. Филюк. – К.: ТОВ «ЦП «КОМПРИНТ», 2017. – Випуск № 1 (34). – С. 175–184.

9. Шкуропат К. С. Розширення прогностичних функцій АС «Податковий блок» на основі використання технологій Data Mining / К. С. Шкуропат, О. В. Редич // Збірник статей студентів ОКР «магістр» Навчально-наукового інституту інформаційних технологій та менеджменту Університету ДФС України. – Ірпінь: Університет ДФС України, 2016. – 80, [46–55] с.

10. Сидор Марія-Василина В. Елементи CRM системи ДФС України на базі технологій Data Mining / Сидор Марія-Василина В., О. В. Редич // Збірник статей студентів ОКР «магістр» Навчально-наукового інституту інформаційних технологій та менеджменту Університету ДФС України. – Ірпінь: Університет ДФС України, 2016. – 80 [64–73] с.

11. Редич О. В. Концептуальні основи впровадження системи оцінювання результативних показників діяльності органів ДПС / О. В. Редич: матеріали науково-практичного семінару, [«Модернізація державної податкової служби України у контексті інноваційного розвитку»]: 8 вересня, м. Ірпінь. – Ірпінь: Національний університет ДПС України, 2010. – 224, [53–60] с. – Режим доступу : http://ir.nusta.edu.ua/jspui/handle/doc/372 (Дата звернення: 12.12.2018).

12. Про затвердження Порядку доступу до інформації в інформаційних, телекомунікаційних та інформаційно-телекомунікаційних системах ДФС : Наказ ДФС : від 24.11.2015 № 934 із змінами внесеними наказом від 24.10.2016 № 887.

13. Про затвердження Порядку призначення ролей в інформаційній системі «Податковий блок» : Наказ ДФС : від 18.08.2015 № 618.

14. Про затвердження Положення про Порядок створення (удосконалення) автоматизованих інформаційних систем : Наказ ДПС : від 12.10.2011 № 54.

15. Проміжний звіт про надання послуг Іноземним підприємством «ГФК ЮКРЕЙН» відповідно до Контракту № 63М від 28 лютого 2011 року щодо результатів розрахунку ключових показників досягнення прогресу впровадження проекту модернізації державної податкової служби України-1 за 2010 рік.

16. Економічний аналіз: базовий курс для податкового аналітика : навч. посібник для слухачів проф. прогр. підвищення кваліфікації / упоряд. І. С. Санжаревська, С. М. Шаповал ; Міндоходів України. – К. : ДП «СВЦ Міндоходів України», 2013. – 800 с.

17. Матвійчук А. В. Штучний інтелект в економіці: нейронні мережі, нечітка логіка : монографія / А. В. Матвійчук. – К. : КНЕУ, 2011. – 439 с.

18. Вітлінський В. В. Концептуальні засади впровадження ризикоорієнтованого підходу у адмініструванні податків / В. В. Вітлінський, О. В. Редич // Проблеми впровад-ження інформаційних технологій в економіці: тези доповідей 7-ї міжнародної науково-практичної конференції, (м. Ірпінь, квітень 2009 р.). – НУДПСУ. – С. 27–40.

19. Ситник В. Ф. Системи підтримки прийняття рішень: навч. посіб. / В. Ф. Ситник – К.: КНЕУ, 2004. – 614 с.

20. Редич О. В., Губатенко О. А. Результати застосування методології інтегральних критеріїв в оцінюванні діяльності органів ДПС України / О. В. Редич, О. А. Губатенко: матеріали науково-практичного семінару [«Модернізація державної податкової служби України у контексті інноваційного розвитку»]. – Ірпінь: Національний університет ДПС України, 2011. – 268, [86–97] с.

21. Редич О. В. Зведений звіт про застосування методології інтегральних критеріїв при оцінці процесів адміністрування податків та результативних показників діяльності органів ДПС України. ДРМ ДПС / О. В. Редич. – 2011. – С. 54.

22. Редич О. В. Оцінювання точок Лаффера на основі двопараметричного методу / О. В. Редич // Науковий вісник: збірник наукових праць Академії державної податкової служби України. – 2002. – № 3 (17). – С. 104–119.

23. Редич. О. В. Показники ефективності податків та їх економетричні моделі / О. В. Редич // Науковий вісник Волинського державного університету імені Лесі Українки. – № 2. – Луцьк. – 2002. – С. 444–448.

24. Мовшович С. М. Выпуск, налоги и кривая Лаффера / С. М.Мовшович, Л. Е. Соколовский // Экономика и математические методы.– 1989. – Т. 30. – Вып. 3.

25. Балацкий Е. В. Лафферовы эффекты и фиансовые критерии экономической деятельности / Е. В. Балацкий // Мировая экономика и международные отношения. – 1997. – № 11.

26. Балацкий Е. В. Точки Лаффера и их количественная оценка / Е. В. Балацкий // Мировая экономика и международные отношения. – 1997. – № 12.

27. Чугунов И. Взаимосвязь ставки налогов и налоговых поступлений / И. Чугунов // Бизнес – информ. – 1998. – № 9. – С. 27–31.

28. Офіційний сайт Oracle. Oracle Business Intelligence / Дистрибутиви / Література. URL:http://www.oracle.com/technetwork/middleware/bi-enterprise-edition/tutorials/obiee11g-453435.html (Дата звернення: 12.09.2018).

29. Офіційний сайт Oracle. Oracle Data Miner/ SQL Developer Extension Література / URL:http://www.oracle.com/technetwork/database/options/odm/dataminerworkflow-168677.html (Дата звернення: 12.09.2018).

30. Офіційний сайт Oracle. Introduction to Artificial Intelligence на Udacity. (Пітер Норвіг) URL: https://www.udacity.com/course/intro-to-artificial-intelligence-cs271 (Дата звернення: 10.09.2018).

31. Стандартні технології доступу до баз даних. – Режим доступу : https://sites.google.com/site/fieldbusbook/integracionnye-tehnologii/sql-odbc-oledb (Дата звернення: 11.01.2018).

32. Офіційний сайт Statsoft. Мультимедійний посібник по технологіям Statistica. – Режим доступу : http://statsoft.ru/home/textbook/default.htm (Дата звернення: 11.02.2018).

33. Определение Облачных Вычислений (Драфт) / Рекомендации Национального Института Стандартов и Технологий (США). Special Publication 800-145 (Draft). – 2012.

Навчальне видання

СЕРІЯ «ПОДАТКОВА ТА МИТНА СПРАВА В УКРАЇНІ»

Редич Олександр Володимирович

ІНФОРМАЦІЙНО-АНАЛІТИЧНІ ТЕХНОЛОГІЇ В ОПОДАТКУВАННІ

Навчальний посібник

Відповідальний за випуск — *Д. Ф. Салахова*

Редактор — *Н. І. Грицюк, О. В. Данильчик*

Форматування та комп'ютерна верстка — *Д. П. Завальніцька*

Здано до друку 11.07.2019. Формат 60×84/16
Папір офсетний № 1. Гарнітура «Times New Roman».
Друк. арк. 25,3.
Тираж 300 примірників. Замовлення № 790.

Підготовлено до друку Видавничо-поліграфічним центром
Університету ДФС України
08200, вул. Університетська, 31, м. Ірпінь, Київська область, Україна

Свідоцтво про внесення суб'єкта видавничої справи до державного реєстру видавців, виготовлювачів і розповсюджувачів видавничої продукції Серія ДК № 5104 від 20.05.2016