МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний економічний університет
Навчально-науковий інститут інноваційних освітніх технологій
Кафедра комп'ютерної інженерії

МАТВІСІВ Павло Миколайович

Алгоритми редукції інформативних ознак цитологічних зображень на основі методу головних компонент / Reduction algorithms of cytological images informative features based on the principal component analysis method

спеціальність: 123 - Комп’ютерна інженерія
магістерська програма - Комп’ютерна інженерія

Магістерська робота

Виконав студент групи КІзм-21
П. М. Матвісів
Науковий керівник:
к.т.н., доцент. К. М. Березька

Магістерську роботу допущено до захисту:

ТЕРНОПІЛЬ -2018

РЕЗЮМЕ

Магістерська робота на тему «Алгоритми редукції інформативних ознак цитологічних зображень на основі методу головних компонент» зі спеціальності 123 «Комп'ютерна інженерія» написана обсягом 77 сторінок і містить 26 ілюстрацій, 5 таблиць, 3 додатків та 46 джерел за переліком посилань.

Метою роботи є розроблення алгоритмів редукції інформативних ознак цитологічних зображень на основі методу головних компонент та проектування програмної системи факторного аналізу.

Методи досліджень. Для розв'язання поставлених задач у магістерській роботі використано методи статистичного моделювання: математична статистика, статистика, кореляційний, регресійний і факторний аналіз, метод головних компонент.

Результати дослідження: математична модель методу головних компонент алгоритми перевірки статистичних гіпотез в компонентному аналізі, програмна система комп'ютерного моделювання алгоритму головних компонент.

Результати роботи можуть бути використані при побудові автоматизованих систем аналізу зображень та в навчальному процесі.

Можливими напрямками подальших досліджень є продовження робіт по редукції інформативних ознак цитологічних зображень: розробка алгоритмів спеціальних методів факторного аналізу, розробка алгоритмів канонічного аналізу.

КЛЮЧОВІ СЛОВА: МЕТОД ГОЛОВНИХ КОМПОНЕНТ, ФАКТОРНИЙ АНАЛІЗ, СТАТИСТИЧНА ГІПОТЕЗА, МАТРИЦЯ КОВАРІАЦІЙ, ВЛАСНІ ЧИСЛА, ВЛАСНИЙ ВЕКТОР, ПРОГРАМНА СИСТЕМА.

RESUME

Diploma work: «Reduction algorithms of cytological images informative features based on the principal component analysis method» from the specialty 123 «Computer engineering» written 77 page volume and contains 26 illustrations, 5 tables, 3 applications and 46 sources for references.

The purpose of the work is to develop algorithms for reduction of informative features of cytological images based on the method of the main components and design of the software system of factor analysis.

Research methods. Methods for statistical modeling are used to solve the tasks set in the master's work: mathematical statistics, statistics, correlation, regression and factor analysis, principal component analysis.

Results of the research: mathematical model of the principal component analysis of the algorithms for checking statistical hypotheses in the component analysis, software system of computer simulation of the main components algorithm.

The results of the work can be used in the construction of automated image analysis systems and in the learning process.

Possible areas of further research are the continuation of work on the reduction of informative features of cytological images: the development of algorithms of special methods of factor analysis, the development of algorithms canonical analysis.

KEYWORDS: PRINCIPAL COMPONENT ANALYSIS, FACTOR ANALYSIS, STATISTICAL HYPOTESIS, MATRIX OF COVARIATIONS, OWN NUMERES, OWN VECTOR, SOFTWARE.

ЗМІСТ

ВСТУП

Актуальність теми. При вирішенні завдань інтелектуального аналізу даних [1], в різних прикладних областях нерідко доводиться оперувати вибірками даних великого обсягу. Це тягне за собою істотні витрати часу на обробку даних, а також вимагає наявності значних обсягів оперативної і дискової пам'яті ЕОМ. Тому актуальним завданням є скорочення розмірності вибірок даних [1-5]. Згідно означення, редукція (лат. reductio – повернення, відновлення) – процес або дія, що призводить до зменшення, послаблення або спрощення чого-небудь, іноді до повної втрати якихось об'єктів, ознак. Редукція зустрічається у різних галузях життєдіяльності: біології, економіці, інформатиці, математиці, хімії, лінгвістиці, музиці, філософії та ін. В роботі розглянуто алгоритми редукції інформативних ознак цитологічних зображень.

Мета і завдання дослідження. Метою роботи є розробка алгоритмів редукції інформативних ознак цитологічних зображень на основі методу головних компонент.

Для досягнення поставленої мети необхідно розв’язати такі завдання:

1. Зробити аналіз методів відбору інформативних ознак.
2. Проаналізувати інформативні ознаки цитологічних зображень.
3. Проаналізувати програмні засоби аналізу даних.
4. Розробити алгоритм методу головних компонент.
5. Побудувати математичну модель методу головних компонент.
6. Розробити алгоритми перевірки статистичних гіпотез в компонентному аналізі.
7. Здійснити комп’ютерне моделювання програмної системи.
8. Розробити модуль факторного аналізу.
9. Розробити модуль інтерпретації результатів факторного аналізу.

Об'єкт дослідження – цитологічні зображення раку молочної залози.

Предмет дослідження – алгоритми методу головних компонент.

Методи досліджень. В роботі застосовувалися лінійна алгебра, математична статистика, статистика, інтелектуальний аналіз даних, кореляційний, регресійний та факторний аналіз, метод головних компонент.

Наукова новизна

1. Алгоритм методу головних компонент.

2. Алгоритм перевірки статистичних гіпотез в компонентному аналізі.

Практичне значення отриманих результатів полягає в розробленні:

1. Модуля факторного аналізу.

2. Модуля інтерпретації результатів факторного аналізу

Публікації та апробація ДР. Основні результати досліджень опубліковано в тезах доповідей VII міжнародної науково-технічної конференції «Актуальні задачі сучасних технологій», (м. Тернопіль) [6]:

Струтинський П.І. Згорткові нейронні мережі як засіб обробки біомедичних зображень / П.І. Струтинський, Н.А. Поворозник, П.М. Матвісів // Матеріали VII міжнародної науково-технічної конференції «Актуальні задачі сучасних технологій», м. Тернопіль, 28–29 листопада 2018 р. – Тернопіль: ТНТУ, 2018. – Т.2. – С. 170-171.

Дипломна робота складається із трьох розділів, висновків, списку використаної літератури та додатків [7].

У першому розділі здійснено аналіз методів, алгоритмів і програмних засобів редукції інформативних ознак цитологічних зображень. Проаналізовано методи відбору інформативних ознак, інформативні ознаки цитологічних зображень і програмні засоби, що використовуються для редукції ознак.

У другому розділі розроблено алгоритми методів факторного аналізу, методу головних компонент, математичну модель методу головних компонент, алгоритми перевірки статистичних гіпотез в компонентному аналізі.

У третьому розділі розроблено програмну систему комп’ютерного

моделювання алгоритмів методу головних компонент, розроблено модуль факторного аналізу, модуль інтерпретації результатів факторного аналізу.

У додатках приведено фрагмент тексту програми, довідку про використання результатів дипломної роботи, світлокопію виданої публікації.

1 АНАЛІЗ МЕТОДІВ, АЛГОРИТМІВ І ПРОГРАМНИХ ЗАСОБІВ РЕДУКЦІЇ ІНФОРМАТИВНИХ ОЗНАК ЦИТОЛОГІЧНИХ ЗОБРАЖЕНЬ

1.1 Аналіз методів відбору інформативних ознак

У задачах розпізнавання образів для визначення значення вихідного параметра застосовуються математичні моделі. Для синтезу таких моделей використовується вхідна навчальна вибірка, що складається з великого набору ознак, які характеризують досліджуваний об’єкт або процес. Масиви даних великого розміру, як правило, містять надлишкові й неінформативні ознаки, які ускладнюють не тільки процес синтезу моделі, але й приводять до її надлишковості, що збільшує час класифікації за такою моделлю. Таким чином, при вирішенні задач розпізнавання образів важливим етапом є процес редукції вхідного набору ознак. Складність вирішення задач вибору максимально значимої комбінації ознак полягає в її комбінаторному характері. Використання повного перебору всіх можливих комбінацій при великій кількості ознак приводить до комбінаторного вибуху.

Традиційним і найбільш широко застосовуваним підходом при скороченні розмірності вибірок даних є використання методів відбору інформативних ознак [1-5] (видаляють з вихідного набору найменш інформативні ознаки) і методів конструювання ознак [5, 8, 9] (замінюють початковий набір ознак розрахованим на його основі набором штучних ознак меншого розміру). Однак, якщо спочатку заданий набір ознак не є надмірною, або обсяг вибірки (число екземплярів в ній) надзвичайно великий для подання та обробки в пам'яті ЕОМ, застосування цих методів виявляється надзвичайно складним, а результати їх роботи або призводять до втрати суттєвої для подальшого аналізу інформації, або не дозволяють зберегти вихідну інтерпретабельність даних.

Іншим, значно рідше використовуваним на практиці, підходом при вирішенні даного завдання є скорочення обсягу вибірки. Як правило, це реалізується за допомогою вилучення випадкових підвибірок з вихідної вибірки

[10-12], що може призводити до формування нерепрезентативних в топологічному сенсі вибірок внаслідок невключення до них екземплярів, які рідко зустрічаються, на границях класів, представлених у вихідній вибірці.

В [13-15] автором запропоновані методи перебору і еволюційні методи формування вибірок, а також модель (комплекс критеріїв) якості вибірки, які дозволяють забезпечити формування з вихідної вибірки підвибірок меншого обсягу, що володіють в системі використовуваних критеріїв кращими якостями. Однак для вибірок дуже великого обсягу застосування даних методів і моделі являється вельми витратним як з обчислювальної точки зору, так і з точки зору ресурсів оперативної і дискової пам'яті.

Виділення максимально значимої комбінації інформативних ознак є досить важким і ресурсномістким завданням, оскільки вона пов’язана з необхідністю комбінаторного перебору. У цей час запропоновані різні методи виділення набору ознак, серед яких найбільше поширення отримали [1]:

– метод повного перебору (exhaustive search);

– пошук у глибину (depth-first search);

– пошук у ширину (breath-first search);

– метод гілок і границь (branch and bound method) або скорочений пошук у глибину;

– метод групового врахування аргументів або скорочений пошук завширшки;

– метод послідовного додавання ознак (forward selection);

– метод послідовного видалення ознак (backward selection);

– метод почергового додавання й видалення ознак (combined selection);

– ранжування ознак;

– кластеризація ознак (unsupervised learning for feature selection);

– випадковий пошук з адаптацією (adaptive stochastic search);

– еволюційний пошук (evolutionary search).

Класифікація методів відбору ознак наведена на рисунку 1.1.

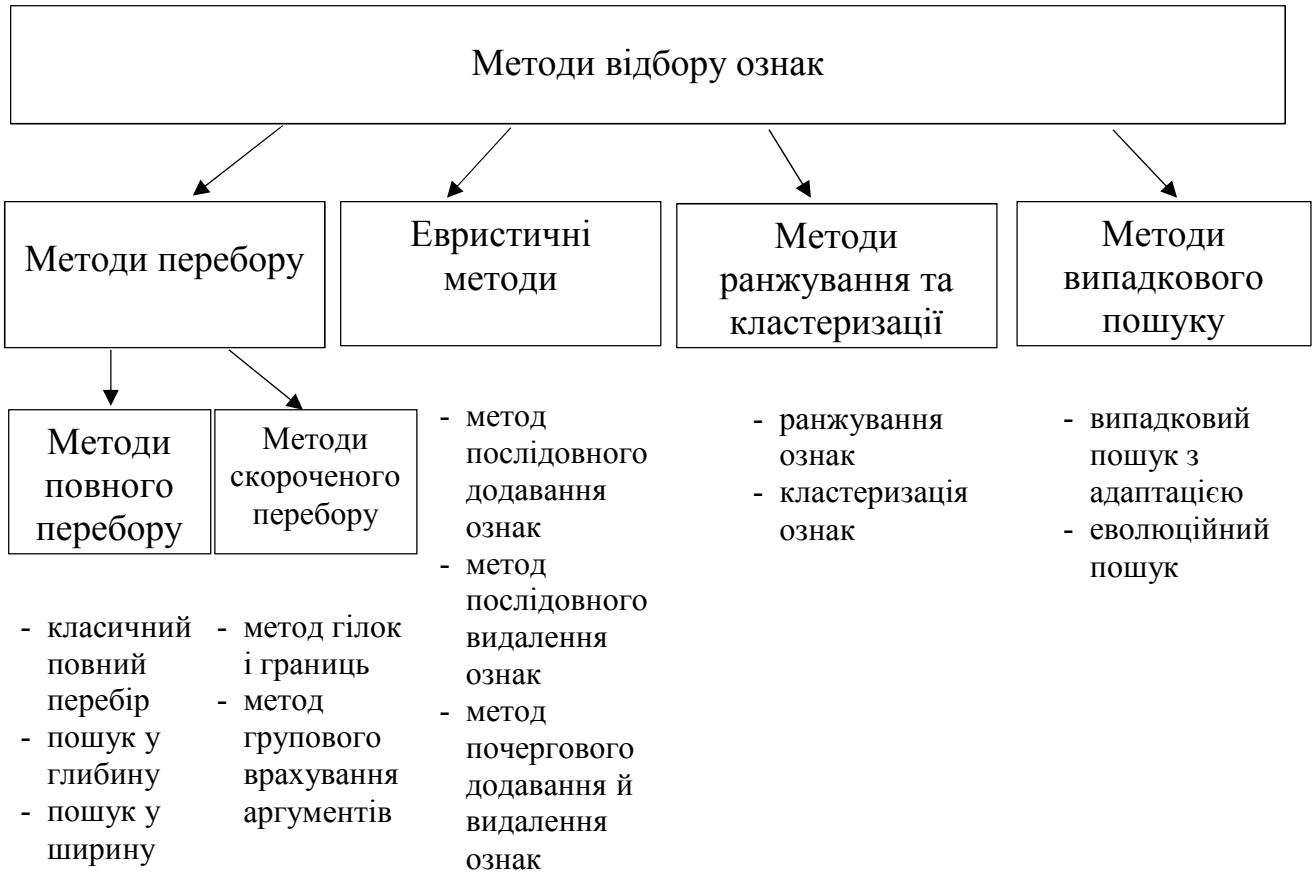


Рисунок 1.1 – Класифікація методів відбору ознак

У методі класичного повного перебору аналізуються всі можливі комбінації ознак, серед яких вибирається найкраща.

Виконати повний перебір всіх можливих комбінацій ознак можливо за допомогою обходу дерева можливих наборів ознак, у якому вузли відповідають наборам ознак. Кореневий вузол відповідає порожньому набору. Кожний наступний набір утвориться шляхом приєднання деякої ознаки до попереднього батьківського вузла. З метою уникання появи в дереві вузлів, що відповідають однаковим наборам, що відрізняються тільки порядком ознак, до дочірніх вузлів додають тільки ті ознаки, номери яких перевищують максимальний номер ознаки в батьківському наборі. При використанні методу пошуку в глибину обхід дерева можливих наборів ознак відбувається по напрямку від кореневого (батьківського) вузла до вузла-нащадка, що характеризується найбільшою кількістю ознак. Для обходу дерева використовується рекурсивний виклик процедури нарощування поточного вузла Наростити(*Xe*), у такий спосіб дерево в явному виді не будується.

Процедура нарощування по черзі приєднує до поточного набору по одній ознаці, і для кожного з отриманих наборів спочатку обчислює значення критерію оцінювання, а потім викликає себе рекурсивно.

При обході дерева можливих наборів ознак за допомогою пошуку в ширину відбувається послідовний перегляд вузлів дерева по рівнях, тобто комбінації ознак аналізуються по збільшенню їхнього розміру. Таким чином, на початку проглядаються всі можливі одноознакові комбінації, потім комбінації, що складаються із двох ознак і т. д.

Аналогічно методу повного перебору недоліком методів пошуку в глибину й пошуку в ширину є їхня обмеженість практичного застосування. Пошук оптимальних комбінацій за допомогою дерева наборів ознак може бути легко перетворений від повного перебору рішень до скороченого, заснованому на використанні деякої додаткової інформації, отриманої з вихідного масиву даних. До методів скороченого перебору відносяться метод гілок і границь (скорочений пошук у глибину) і метод групового обліку аргументів (скорочений пошук в ширину). Скорочення кількості комбінацій, що перебираються, у методі гілок і границь досягається за рахунок відмови від нарощування гілки дерева у випадку, якщо вже є краща гілка. Тобто вузол, що відповідає набору ознак *Xe*, не нарощується, якщо значення критерію оцінювання набору ознак $J(Xe)$ виявиться гірше, ніж на найкращому із уже оцінених наборів меншої розмірності. З метою пошуку більш оптимального набору спочатку ознаки ранжуються в порядку спадання їхньої індивідуальної значимості.

Застосування методу гілок і границь дозволяє скоротити час, необхідний для пошуку. Недоліком такого методу є послідовне додавання ознак до оцінюваного набору ознак, що часто приводить до виключення з розгляду комбінацій ознак, що володіють максимальною інформативністю. У методі групового врахування аргументів (МГВА) на кожній t-ій ітерації оцінюється не один набір ознак, а множина P_t ($|P_t| = N$) наборів, що називається t-им рядом. Для переходу від поточного P_t до наступного P_{t+1} ряду від кожного набору $Xe \in P_t$ породжується $L - t$ нових наборів шляхом приєднання одного з ознак, що не

належать набору *Xe*. Зі згенерованих $N(L - t)$ наборів ознак у наступний ряд відбирається не більше N наборів, кращих за значенням критерію оцінювання набору ознак. Таким чином, на кожній ітерації розмір наборів ознак збільшується на одиницю. Кількість наборів ознак на кожній ітерації N називається шириною пошуку. Зокрема при $N = 1$ метод групового обліку аргументів являє собою евристичний метод послідовного додавання ознак.

Такий метод дозволяє позбутися необхідності оцінювання кожної з $2^L - 1$ можливих комбінацій ознак, але є більше складним у порівнянні з методом гілок і границь. Евристичні методи відбору ознак використовують жадібні стратегії (greedy strategy) для додавання або видалення ознак на кожній ітерації. До евристичних методів відносяться метод послідовного додавання ознак, метод послідовного видалення ознак, а також метод почергового додавання й видалення ознак. У методі послідовного додавання ознак на основі оптимального набору ознак, знайденого на попередній ітерації X^*_{t-1}, на поточній ітерації формуються всі можливі комбінації ознак $Xe_{t,k}$ шляхом додавання однієї ознаки, ще не включеного в набір X^*_{t-1}. У наступній ітерації формування нових рішень відбувається на основі оптимального набору $X^*_t = argmin\ J(Xe_t,k)$, знайденого на поточній ітерації.

Метод послідовного додавання ознак є простим у реалізації, а також не вимагає значних тимчасових витрат при його використанні: обчислювальна складність методу $O(L^2)$, є значно нижча, ніж у методів повного перебору.

Недолік такого методу викликаний неоптимальністю жадібної стратегії пошуку: при використанні методу послідовного додавання ознак часто в оптимальний набір включаються надлишкові ознаки. У методі послідовного видалення ознак на основі оптимального набору ознак, знайденого на попередній ітерації X^*_{t-1}, на поточній ітерації формуються всі можливі комбінації ознак *Xet*,*k* шляхом видалення однієї ознаки, ще з набору X^*_{t-1}. У наступній ітерації формування нових рішень відбувається на основі оптимального набору $X^*_t = argminJ(Xe_{t,k})$, знайденого на поточній ітерації. Таким чином, у методі послідовного видалення ознак на кожній ітерації виключаються ознака, що

мінімально погіршує критерій оцінювання набору ознак.

Метод послідовного видалення ознак також, як і попередній метод, простий у реалізації. До основного недоліку такого методу варто віднести неоптимальність жадібної стратегії. Важливо відзначити, що метод послідовного видалення ознак працює повільніше в порівнянні з методом послідовного додавання ознак, оскільки на початкових ітераціях необхідно оцінювати набори ознак, що складаються із всіх або майже всіх ознак. Такий метод застосовують у випадках, коли відомо, що інформативних ознак значно більше, ніж малоінформативних або надлишкових.

Метод послідовного додавання та видалення ознак сполучає в собі ідеї двох розглянутих раніше методів, що діють протилежно, у результаті чого виходить нежадібна стратегія пошуку.

Ідея такого методу полягає в тому, щоб використовувати стратегію додавання ознак доти, поки не виникне ситуація, при якій збільшення кількості ознак в оптимальному наборі не приводить до поліпшення значення критерію оцінювання набору ознак $J(X^*)$. Після цього запускається процедура видалення ознак, що припиняє своє функціонування за тих самих умов, що й процедура додавання ознак. Процедури додавання й видалення ознак чергуються доти, поки значення критерію $J(X^*)$ в оптимальних точках X^* не перестане поліпшуватися.

Метод почергового додавання й видалення ознак є більше складним у реалізації й працює довше в порівнянні з методами послідовного додавання й послідовного видалення ознак окремо й також не гарантує оптимальності знайденого рішення. Однак рішення, отримані за допомогою такого методу, як правило, виявляються більш оптимальними в порівнянні з рішеннями, отриманими шляхом застосування методів послідовного додавання й послідовного видалення ознак.

Методи ранжування й кластеризації не використовують критерії оцінювання спільного впливу набору ознак на вихідний параметр. Так при ранжуванні використовуються критерії оцінювання індивідуальної значимості ознак. У методах кластеризації застосовуються метрики відстані на ознаках.

Для відбору ознак також використовується ранжування ознак. При такому підході виконується сортування ознак по оптимальності обраного критерію індивідуальної значимості. Після індивідуального оцінювання кожної ознаки і їхнього сортування відбувається вибір певної кількості ознак, індивідуальна значимість яких задовольняє заданим умовам.

Для оцінювання індивідуальної значимості ознак можуть використовуватися критерії кореляції (парний, Фехнера, знаків), інформаційний критерій, ентропія ознак.

Перевагою такого підходу є простота реалізації, а недоліком – можливість застосування тільки у випадку, якщо ознаки у вихідному наборі є статистично незалежними. Як правило, при рішенні практичних завдань ознаки статистично залежать друг від друга, у результаті чого при використанні такого підходу для відбору ознак виходять комбінації, що містять надлишкові ознаки, отже, такі комбінації виявляються далеко не оптимальними.

Виділити максимально значиму комбінацію можна за допомогою кластеризації ознак (unsupervised learning for feature selection).

Методи кластеризації застосовують для розбивки вибірки на кластери, що складаються зі схожих екземплярів, і виділення в кожній групі одного найбільш типового екземпляра. Аналогічні дії можна виконати не над екземплярами, а над ознаками, якщо ввести функцію відстані на ознаках.

Недоліком кластеризації ознак є можливість існування кластерів, що цілком складаються з неінформативних ознак, у результаті чого типові представники таких кластерів можуть увійти в комбінацію ознак, що вважається найбільш інформативною.

Кластеризацію ознак доцільно застосовувати на початкових етапах інших методів відбору ознак для формування груп (кластерів) схожих ознак. Після цього в процесі пошуку оптимальної комбінації занадто схожим ознакам, що належать одному класу, забороняється входити в той самий набір.

У методі випадкового пошуку з адаптацією на кожній t-ій ітерації використовується популяція P_t, що складається з N рішень. Формування нових

рішень відбувається шляхом випадкової генерації N наборів ознак залежно від розподілу ймовірностей включення ознак у набори, що генеруються.

Ідея адаптації полягає у тому, щоб при генерації нових наборів ознак імовірність включення в них ознак, які частіше входять у кращі набори, була більшою в порівнянні з імовірністю мало використовуваних у кращих наборах ознак.

Кількість наборів ознак N у кожній ітерації повинна бути по можливості мінімальною, але одночасно достатньою для того, щоб кращий набір X_{min} був близький до оптимального, а гірший X_{max} – до самого неінформативного набору ознак.

Параметр h, що задає ступінь адаптації, вибирається таким чином, щоб імовірність включення ознаки не могла стати нульовою. При $h = 0$ метод випадкового пошуку з адаптацією перетвориться до неадаптивного випадкового пошуку.

Перевага методу випадкового пошуку з адаптацією полягає в тому, що в більшості випадків він знаходить більш оптимальні набори ознак у порівнянні з методом почергового додавання й видалення ознак. Недоліком такого методу є досить повільна збіжність.

Для відбору інформативних ознак з вихідного масиву, що містить L ознак, за допомогою методів еволюційного пошуку рішення (хромосома) представляється бітовим рядком розміру L. Якщо біт хромосоми приймає одиничне значення, то відповідний йому ознака вважається інформативним і враховується при оцінюванні набору ознак, що відповідає хромосомі. У противному випадку, коли біт приймає нульове значення, ознака вважається неінформативним і не використовується при оцінюванні комбінації ознак.

Перевага такого подання полягає в тому, що класичні еволюційні оператори схрещування й мутації можуть бути застосовані для відбору ознак без внесення в них яких-небудь змін.

Перевагою еволюційного пошуку є те, що він має можливості для виходу з локальних оптимумів і пристосований для знаходження нових рішень за рахунок

об’єднання кращих рішень, отриманих на різних ітераціях. Крім того, еволюційний пошук адаптується до особливостей цільової функції. Створені в процесі схрещування, нові рішення тестують усе більш широкі області простору ознак і переважно розташовуються в області оптимуму. Відносно рідкі мутації перешкоджають виродженню популяції, що рівносильно рідкому, але безупинному пошуку оптимуму у всіх інших областях простору ознак.

Недоліками еволюційного пошуку є відносно повільна збіжність і залежність від початкових умов пошуку.

1.2 Аналіз інформативних ознак цитологічних зображень

До одного з найпоширеніших видів раку відноситься рак молочної залози, що є злоякісною пухлиною, утвореною з клітин молочної залози. Виявляють рак молочної залози у клінічній практиці за допомогою рентгенівського дослідження – мамографії. Після проведення мамографії проходить діагностування злоякісних новоутворень (ЗН) біопсією. Біопсія – забір клітин і тканин ураженого органу на цитологічне або гістологічне дослідження. Після проведення цитологічного дослідження формулюється діагноз, завдяки якому розробляють тактику лікування пацієнта. При цьому дослідженні використовують цитологічні зображення (ЦЗ). Мікрооб’єктами на ЦЗ є клітини та їх складові. Цитологічна діагностика ЗН ґрунтується на морфології клітини і насамперед та особливо ядра.

Для того, щоб підтвердити істинність попереднього діагнозу, спланувати подальше лікування і терапію використовуються гістологічні зображення (ГЗ), що отримуються за допомогою мікроскопа. Об’єктами на ГЗ є зрізи тканин певних органів, утворені впорядковано розміщеними клітинами. Обидва види зображень – цитологічні і гістологічні входять до класу біомедичних зображень (БМЗ). БМЗ – це зображення, які отримані за допомогою будь-якої біомедичної техніки. Якщо

відбувається зміна структури тканин та клітин то протікає злоякісний процес. Прикладом ознак таких змін є: інфільтрація – проникнення в тканину клітин, що не є їх нормальною складовою частиною, інвазія – здатність клітин злоякісної пухлини відділятися та проникати в сусідні тканини, атипія – це порушення розташування клітин, характерного для нормального стану тканини, з якої розвинулася пухлина, порушення стратифікації – розділення тканини на шари. Ракові клітини розташовуються безладно, у вигляді хаотичних скупчень.

Злоякісні процеси на основі ЦЗ і ГЗ аналізуються візуально. В результаті цього аналізу отримується якісна експертна оцінка, яка є необхідною, але не є достатньою, оскільки для повної характеристики патології процесів потрібні кількісні ознаки, які отримуються шляхом числового морфометричного аналізу. В умовах клінічної практики для автоматизації аналізу ГЗ і ЦЗ застосовуються системи автоматизованої мікроскопії (САМ). САМ складається з мікроскопа, цифрової камери і спеціалізованого програмного забезпечення.

Цитологічна діагностика ЗН ґрунтується як на геометричних ознаках зображень клітини та їх ядер так і на оцінці сукупності клітин та фону препарату [16].

На основі візуального спостереження зображень рак діагностується за наступними інформативними ознаками:

- описом форми,
- кольору,
- текстури окремих клітин,
- розміщенням одна відносно одної,
- щільністю розміщення,
- формою груп клітин.

Остання описується спеціальними термінами, що походять від назв предметів подібних до них візуально. В спеціальній літературі для опису препарату характерним є використання якісних термінів, що описують частоту появи певної ознаки, наприклад: "часто буває", "де-не-де", "характерно", "інколи" та ін. Згідно [17] типове правило розміщення виглядає наступним чином: клітини

розташовуються переважно в щільних, пухких скупченнях. Правило опису клітин може мати такий вигляд [17]: відзначається рясний клітинний склад, неправильна форма ядер, нерівні і нечіткі контури ядерної мембрани.

Цитологічний метод дає можливість виявити передракову проліферацію епітелію – дисплазію. Допомагає діагностувати рак в тому числі його ранні форми, створює передумови для вивчення на клітинному рівні передпухлинних станів, передуючих виникненню раку і патогенетично з ним зв'язаних [18, 19].

Основним і неодмінним правилом онкологічної діагностики до початку лікування є морфологічна верифікація діагнозу раку і іншого злоякісного новоутворення. Ігнорування цього правила приводить до клінічних помилок більш ніж біля 1/4 хворих і затримує початок лікування. Тому цитологічне дослідження є одним з найважливіших діагностичних заходів в комплексному обстеженні хворих.

Цитолог також проводить морфологічне розпізнавання не тільки онкологічних, а й непухлинних уражень. Він дає гістологічну характеристику патологічного процесу. Іншими словами, він виконує ті ж завдання, що й патоморфолог. Тому цитолог не має права на помилки.

Помилки цитологічної діагностики можуть бути при розбіжностях цитологічних і гістологічних даних про природу ураження, злоякісність або доброякісність процесу, гістологічну форму пухлинного або не пухлинного процесу, ступінь злоякісності.

Вірогідність помилки досить велика у випадку, коли клітинна атипія (нерівність контурів і гіперхромія ядер при відносно невеликих розмірах пухлинних клітин) при раку виражена слабо. У зв'язку з цим цитологічний діагноз злоякісної пухлини (рак, саркома і ін.) встановлюють по сукупності багатьох ознак, що виявлених у ряду клітин і відображають різний ступінь змін (атипії) в морфології клітини і ядра.

При гіпердіагностиці раку молочної залози більшість помилок пов'язана з неправильним трактуванням цитологічної картини при фіброаденомі і проліферативному аденозі: проліферація епітелію дольок і проток з укрупненням і

поліморфізмом приводить до гіпердіагностики раку. Тільки аналіз ядерних характеристик клітин пухлини: наявність рівних контурів ядра і рівномірний розподіл хроматину допомагає правильному трактуванню змін. Також на дорозі подолання цих труднощів стоїть використання сучасних методів уточнюючої діагностики, перш за все імуноцитохімічне дослідження [20-23].

На даний час існують обмежені можливості використання сучасних методів уточнюючої діагностики: морфометрії, імуноцитохімії, проточної цитофлуорометрії, сучасних методів молекулярної діагностики. Щоб не допустити помилки цитолог повинен мати матеріал, узятий у хворого саме з основного патологічного осередку, а не з тканини, що оточує його, або супутнього патологічного процесу. А такої впевненості у кожному конкретному випадку немає.

Проведемо аналіз зображень пухлини молочної залози в медичних термінах цитологів.

Почнемо з нормального стану. В зрізі нормальної молочної залози жінок 20-35 років серед рихлої сполучної тканини є в значній кількості зрілі і незрілі часточки молочної залози. Альвеоли залоз і молочні протоки вистелені однорідним кубічним епітелієм, до якого із зовнішньої сторони прилягають зірчасті міоепітеліоцити (міоепітелій). Епітелій молочних проток має призматичну форму. Після 35-40 років кількість часточок значно зменшується, видно молочні протоки і огрубілу фіброзну тканину, що розрослася В постклімактеричному періоді часточки атрофуються і зустрічаються рідко. В цитологічних препаратах виявляються еритроцити, зрідка гранулоцити і лімфоцити. В багатьох препаратах місцями визначаються одиничні або невеликими пучками фібробласти, фіброцити. Цитоплазма містить невелику кількість дрібно розсіяних гранул, добре помітних при фазоконтрастній мікроскопії. Цитоплазма люмінесціює червоним світлом. Ядра в основному однорідні, округлої, рідше овоїдної, форми, майже однакових розмірів (5,5 ± 0,1 мкм), розміщуються в центрі клітини і займають приблизно половину її об'єму, люмінесціюють зеленим світлом. Ядерця невеликі. В суправітально забарвленій

суспензії клітин цитоплазма слабо забарвлена, ядра – блідо, а ядерця – інтенсивно. В ядрах 70-80% клітин помітна ніжна сітка хроматину.

Фіброзна форма. Цитологічні препарати зіскобу і пунктата бідні клітинами, основу їх складає однорідний залозистий епітелій, розташований комплексами малих розмірів, рідше видно фіброцити і обривки фіброзної тканини, зрідка – молозивоподібні клітини і тканинні базофіли (огрядні клітини). Є трохи жирових крапель і дрібнозернистого детриту. Однорідні клітини залізистого епітелію цитоморфологічно і цитохімічно майже ідентичні епітелію нормальної молочної залози. В частині клітин менше елементів залозистого епітелію нормальної молочної залози, цитоплазма їх більш багата дрібними гранулами, ядра мають густу сітку хроматину, інтенсивніше сприймають фарбник. В люмінесцентних препаратах ядра зелені з жовтим відтінком. Зміст нуклеопротеїдів помірний.

Прострумова форма. В цитологічних препаратах зіскобів і пунктатів виявляються клітини однорідного, рідше гіперплазованого залізистого епітелію. Останній подібний до нормального залозистого епітелію, розташований більш компактно, з менш чіткими межами цитоплазми, що містить дрібні гранули. Трохи були збільшені розміри клітин (11,09 ± 0,7 мкм), ядер (6,75 ± 0,32 мкм), ядерець (1,15 ± 0,088 мкм). Частина клітин і ядер приймають овоїдну або дещо подовжену форму, що робить їх особливо добре помітними серед розрізнено розміщених елементів. Зрідка в комплексі виявляються одиничні або по 3-4 рядом розташовані клітини з жовтуватим відтінком люмінесценції ядра і оранжевою цитоплазмою. Суправітально забарвлені клітини мають таке ж забарвлення, як і клітини однорідного залізистого епітелію. Легше визначається незначний поліморфізм ядер і клітин. В мазках і відбитках, забарвлених азуреозином, цитоплазма голубого, ядра блідо-бузкового, ядерця голубого кольору. В цитоплазмі невелика кількість дрібних гранул синюватого кольору, зрідка одиничні дрібні вакуолі. Хроматин ядер сітчастий, частково дрібнозернистий. Вміст нуклеопротеїдів в клітинах помірний, глікогену і кислих глікозаміногліканів - у вигляді слідів.

Аденома молочної залози. В цитологічних препаратах виявлені епітеліальні клітини відносно невеликого розміру, по будові подібні до епітелію альвеол і залозистих трубок незміненої молочної залози (рисунок 1.2). Вони можуть розташовуватись розрізнено, групами, великими полями іноді багатошарово мікроскопічно малими тканинними ділянками і багаточисельними округлими структурами без просвіту. Ознаками проліферації можуть служити зміни морфологічних властивостей клітин або поява незвичайних для незміненого органа багатоклітинних епітеліальних структур. До клітинних критеріїв проліферації слід віднести збільшення розмірів клітин та їх ядер. У клітинах переважають округлі, або злегка овальні, ніби набряклі ядра з чітко верифікованим ніжнопетлистим інтенсивно профарбованим хроматином. В окремих ядрах можуть визначатись поодинокі, відносно невеликі, інтенсивно профарбовані ядерця. Завжди виявляються епітеліальні клітини, ядра яких світяться жовто-зеленим світлом і більш інтенсивно, ніж сусідні клітини. Розрізнені клітини в основному позбавлені цитоплазми. В порівнянні з нормальним залозистим епітелієм вміст РНК в цитоплазмі епітеліальних клітин фіброаденом помірний і підвищений, в ядерцях – помірний. ДНК визначається в основному в помірній концентрації, глікоген – в помірній і великій, кислі глікозаміноглікани – у вигляді слідів. Нейтральні глікопротеїди, ліпіди у більшості хворих не виявляються. Розподіл речовин в клітині рівномірний. Цитоплазма переважної більшості клітин пофарбована в насичені голубі тони, іноді бузкові або сині. Клітинні комплекси, які характеризують проліферацію, мають вигляд округлих утворів, подібних на первинні залозисті ацинуси, але не мають просвітів. Зустрічаються також сосочкоподібні комплекси з щільним розміщенням клітин і багатошарові пласти.

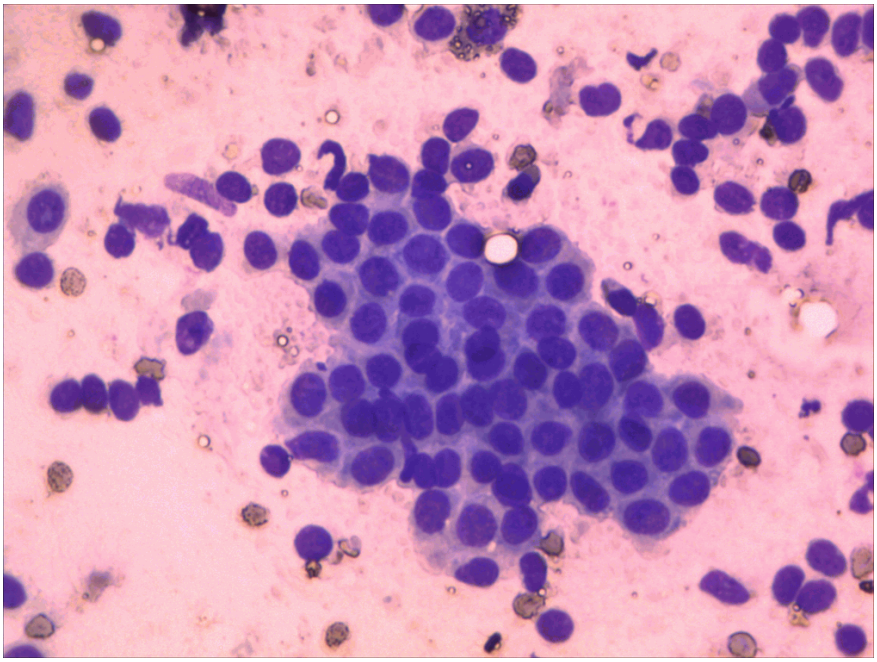


Рисунок 1.2 – Цитограма протокового епітелію при аденомі молочної залози , помірна проліферація епітелію. Забарвлення по Романовському –Гімза. ×200

Цитологічна картина (рисунок 1.3) характеризується присутністю округлих або кубічних, або призматичних епітеліоцитів середнього та дрібного калібру, рідше великого розміру. Клітини розміщуються розрізнено групами або залозистоподібними сосочковими структурами із накопиченням розеток.

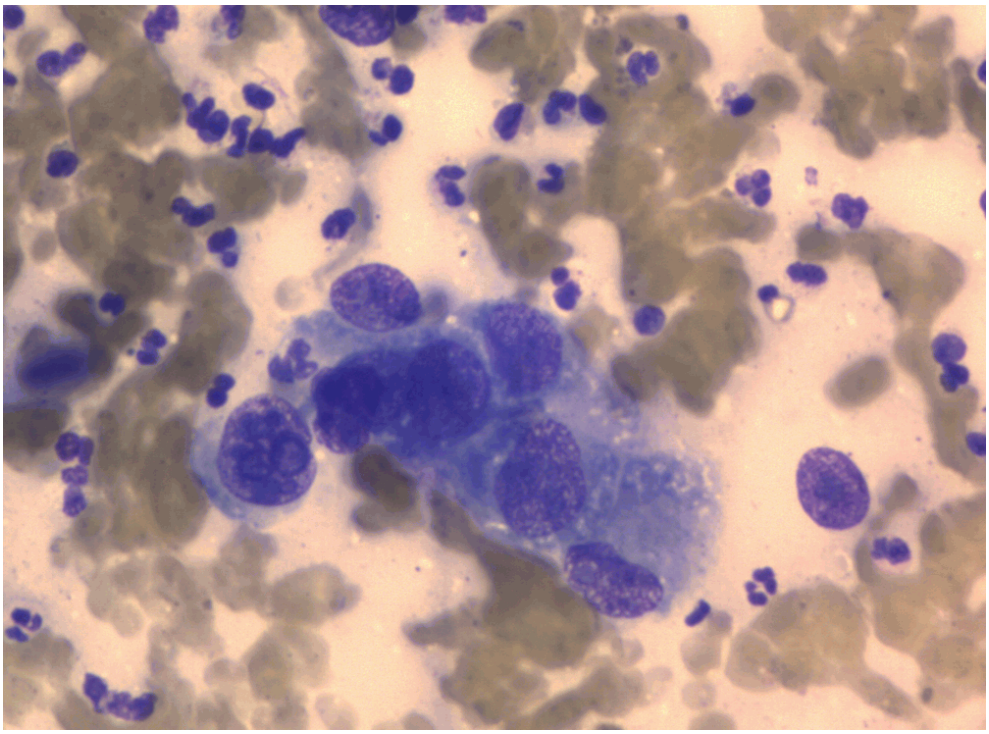


Рисунок 1.3 – Цитограма протокового епітелію при внутрішньопротоковому раку молочної залози. Забарвлення по Романовському – Гімза. ×200

Цитоплазма клітин добре виражена, насичена, виявляє ознаки секреторної активності і представлена дрібнозернистою пінистою нерівномірно пофарбованою, з оксіфільними (просвітленими) ділянками на фоні вираженої базофілії. Ядра, переважно округлі і овальні, різні за величиною часто із грубим нагромадженням хроматину, займають значну частину клітини, інтенсивно пофарбовані. Ядерця у них визначаються не завжди. Поліморфізм та гіперхромія ядер виражені у більшості клітин. При аденокарциномі чітко виражені сосочкові та округлі структури.

При цитологічному дослідженні клітинного компоненту інфільтративного раку молочної залози спостерігається наявність атипових епітеліальних клітин великого, середнього та дрібного розміру; округлої, кубічної та неправильної форми. Всі вони містять поліморфні ядра (рисунок 1.4), часто збільшені ядерця, і зазвичай мають насичену цитоплазму. В деяких клітинах можуть зустрічатись ознаки секреторної активності, а саме: дрібнозерниста, піниста, нерівномірно пофарбована іноді ортохромна цитоплазма, іноді із розривом апікальної частини клітини. Такі клітини розміщуються або розрізнено, або у вигляді комплексів і щільних клітинних груп (рисунок 1.5). Для даного виду клітин характерний різкий поліморфізм гіперхромія ядер та їх дифузне пофарбування.

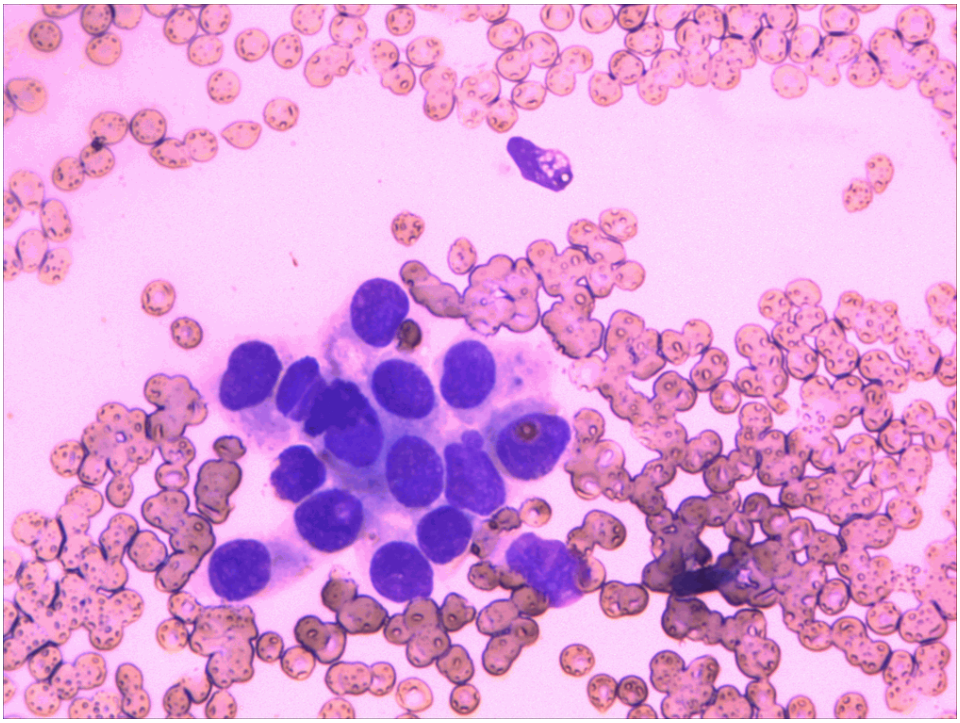


Рисунок 1.4 – Цитограма інфільтративного раку молочної залози. Забарвлення по Романовському – Гімза. × 200

Цитограми аденокарциноми характеризуються присутністю своєрідних комплексів, що мають будову атипових залозистих бульбашок або залозистих ходів. Ці комплекси мають округлу або сосочкоподібну форму, ядра клітин орієнтовані до периферії структури. Нерідко просвіт структури через дискомплексацію клітин не визначається поряд із комплексами зустрічається значна кількість розрізнених пухлинних клітин.

Цитологічний діагноз аденокарциноми базується на особливостях морфології пухлинних клітин (насичена негомогенна цитоплазма та ексцентрично розміщені ядра) і виявлення характерних багатоклітинних комплексів, нерідко чіткі клітинні межі в комплексах не виявляються. В таких випадках на залозистоподібні структури вказує розміщення ядер у вигляді розеток.

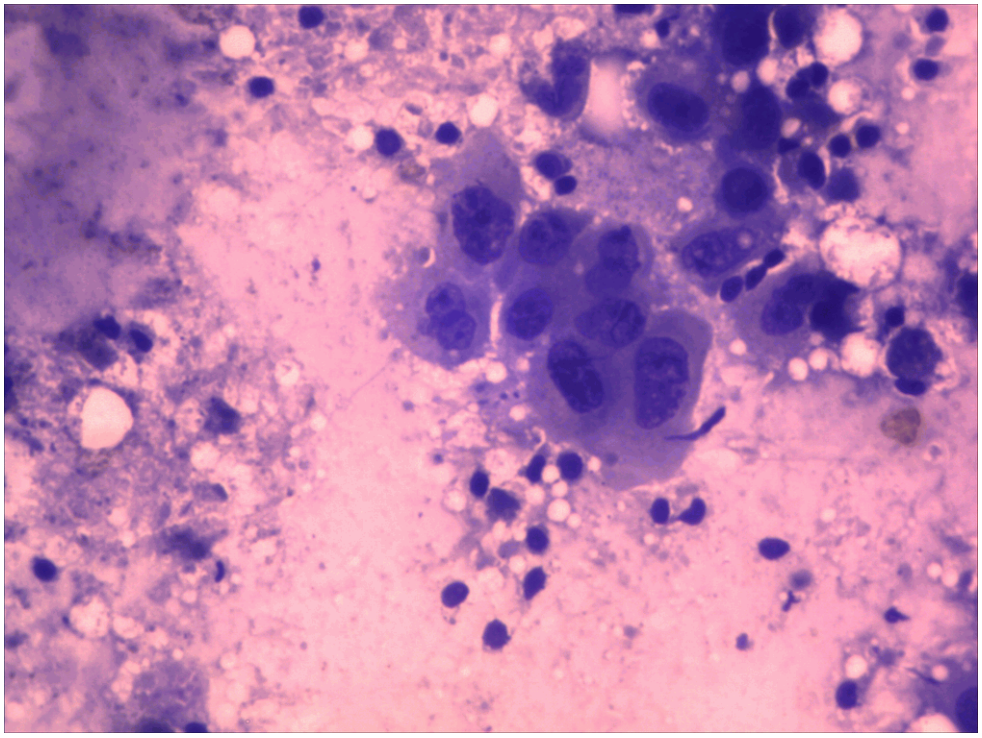


Рисунок 1.5 – Цитограма інфільтративного раку молочної залози. Забарвлення по Романовському – Гімза. × 200

Проведено експертний аналіз цитологічних зображень декількох видів раку та нормальних тканин молочної залози, в результаті чого отримано набір характерних ознак мікрооб'єктів, які використано для побудови бази знань інтелектуальної системи діагностування.

Аналіз цитологічних зображень різних форм раку молочної залози показав, що для внутрішньопротокового раку молочної залози характерні епітеліоцити різного розміру, які розташовуються групами, у вигляді залозистоподібних сосочкових структур. Цитоплазма клітин зерниста, насичена, з ознаками секреторної активності.

Цитологічний аналіз інфільтративного раку базується на таких ознаках:

- формування атипових епітеліальних клітин різного розміру,
- неправильної форми, із вираженим поліморфізмом ядер,
- наявністю гіперхромних ядерець, розміщених ексцентрично.

Рак молочної залози – одне з щільних злоякісних новоутворень, яке найбільш піддається лікуванню. Невеликі пухлини, локалізовані в тканинах

залози, видаляють, і, найчастіше, випадків рецидивування не відзначається.

Лікування раку молочної залози – хірургічне. Вибір операції залежить від розміру пухлини, ступеня ураженості навколишніх тканин і лімфовузлів. Довгий час практично всім жінкам з виявленою злоякісною пухлиною молочної залози проводилася радикальна мастектомія (повне видалення залози, розташованих поруч лімфатичних вузлів і м'язів грудної клітки, розташованих під нею). Зараз все частіше роблять модифікований аналог операції, коли грудні м'язи зберігають (якщо вони не зачеплені злоякісним процесом).

В даний підрозділі охарактеризовано інформативні ознаки цитологічних зображень.

1.3 Програмні засоби аналізу даних

Більшу частину статистичних пакетів можна розбити на три групи – це статистичні пакети загального призначення, спеціалізовані програмні продукти, професійні пакети.

Пакети загального призначення або універсальні (Statgraphics, SPSS, Statistica, Excel, Minitab, EViews тощо) є найбільш зручними для користувача-початківця завдяки відсутності орієнтації на специфічну предметну галузь, широкому діапазону статистичних методів і дружньому інтерфейсу користувача. Вони більш доступні для практики й можуть використовуватися широким колом фахівців різного профілю.

Спеціалізовані пакети – як правило, реалізують кілька статистичних методів або методи, які застосовуються в конкретній предметній області. Найчастіше це системи, орієнтовані на аналіз часових рядів, кореляційно-регресійний, факторний або кластерний аналіз. Застосовувати такі пакети доцільно в тих випадках, коли потрібно систематично вирішувати завдання з цієї області, для якої призначено спеціалізований пакет, а можливостей пакетів

загального призначення недостатньо. З російських пакетів більш відомі STADIA, Minitab, EViews, Статистик-Консультант; ODA, WinSTAT, Statit і т.д.

Професійні пакети призначені для користувачів, які мають справу із надзвичайно великими обсягами даних або вузькоспеціалізованими методами аналізу [24].

Статистичний пакет повинен відповідати певним вимогам:

- модульність;

- асистування при виборі способу обробки даних;

- використання простої проблемно-орієнтованої мови для формулювання завдання користувача;

- автоматична організація процесу обробки даних та зв'язків з модулями пакету;

- ведення банку даних користувача і складання звіту про результати зробленого аналізу;

- діалоговий режим роботи користувача з пакетом;

- сумісність з іншим програмним забезпеченням.

Пакет Statgraphics розроблявся ще для роботи в середовищі DOS, а потім був адаптований до операційної системи Windows і отримав нову назву Statgraphics Plus. Сучасний пакет STATGRAPHICS PLUS – це досить потужна статистична програма, яка містить більше 250 статистичних функцій. За своїми характеристиками пакет займає проміжне місце між SPSS і Statistica.

Пакет SPSS (Statistical Package For Social Science) – це один із найбільш часто використовуваних пакетів статистичної обробки даних. Цей пакет створювався ще для «великих» електронно-обчислювальних машин і послідовно переводився для роботи в середовищі DOS, а потім Windows. Пакет SPSS досить потужний і добре відпрацьований, наближається за своїми можливостями до професійних пакетів, і реалізація статистичних процедур добре пристосована до практичної роботи.

Поряд з пакетом SPSS, великою популярністю в освітніх та наукових закладах США користується пакет STATA. Це професійний статистичний

програмний пакет з data-management system, який досить часто застосовують для біомедичних цілей. Програма добре документована, видається спеціальний журнал для користувачів системи. Офіційний сайт http://www.stata.com/.

Не менш популярним інструментом розробки користувацьких додатків в медицині, бізнесі, економіці, фінансах, промисловості тощо є інтегрована система аналізу та управління даними STATISTICA [24].

Пакет Statistica спеціально створювався для роботи в середовищі Windows і відповідає всім стандартам Windows, що дозволяє зробити аналіз високоінтерактивним. STATISTICA складається з набору модулів, в кожному з яких зібрані тематично пов'язані групи процедур, що мають високу швидкість і точність обчислень. Система STATISTICA містить повний набір класичних методів аналізу даних [25, 26]: від основних методів статистики до просунутих методів, що дозволяє гнучко організувати аналіз. Дані системи STATISTICA легко конвертувати в різні бази даних і електронні таблиці. Ця система відрізняється найбільш розвиненим інтерфейсом із користувачем і багатими графічними можливостями, підтримує високоякісну графіку, що дозволяє ефектно візуалізувати дані і проводити графічний аналіз. Гнучка і потужна технологія доступу до даних дозволяє ефективно працювати як з таблицями даних на локальному диску, так і з віддаленими сховищами даних.

Система STATISTICA є відкритою системою: містить мови програмування, які дозволяють розширювати систему, запускати її з інших Windows-додатків, наприклад, з Excel.

Електронна таблиця Excel найпоширеніша і найбільш доступна, вона встановлюється автоматично при інсталяції пакета MS Office. Електронна таблиця Excel тісно інтегрується з іншими програмами пакета MS Office, наприклад, MS Word і PowerPoint, а тому найчастіше використовується при оформленні результатів роботи. MS Excel – це електронна таблиця з достатньо потужними математичними можливостями, проте деякі статистичні функції є просто додатковими [27,28]. MS Excel, як правило, використовується при найпростішому статистичному аналізі даних. Окрім того, в MS Excel неможливо побудувати

якісні наукові графіки. MS Excel добре підходить для накопичення даних, попередніх статистичних прикидок, для побудови деяких видів діаграм, проте остаточний статистичний аналіз бажано робити в програмах, які спеціально створені для цих цілей. Існує макрос-додаток XLSTAT-Pro (http://www.xlstat.com) для MS Excel, який включає в себе більше 50 статистичних функцій.

Програма STADIA включає в себе всі необхідні статистичні функції, призначені для статистичного аналізу даних. Проте ця програма фактично не змінюється з 1996 року, а тому графіки і діаграми, побудовані за допомогою STADIA, виглядають в сучасних презентаціях архаїчно. До позитивних якостей програми можна віднести наявність книг, що описують роботу [27].

Слід зазначити, що всі ці пакети постійно оновлюються і з кожним роком з'являються їх нові версії (таблиця 1.1).

Таблиця 1.1 – Хронологія розвитку пакетів прикладних програм з обробки й аналізу даних

Роки	Основні пакети аналізу даних	Операційні системи
1970-1985	SSP, BMDP, SAS	DOS
1985-1995	Statgraphics, STATA, SAS, Systat, STADIA, МЕЗОЗАВР, САНІ, Евріста, Клас-майстер тощо	
1995-2017	Statgraphics Plus, SAS, SPSS, Statistica, Excel тощо	Windows

При виборі пакету для аналізу даних можна виділити два аспекти: а) початковий вибір пакету аналізу; б) поточний вибір при переході на більш сучасний, більш потужний пакет. Підходи в обох випадках дещо відрізняються.

У першому випадку на вибір накладаються такі обмеження:

1. Можливості комп'ютера.

2. Можливості одержання установчої версії пакету.

3. Характеристики пакету.

Що стосується першого пункту, то варто вибирати найбільш сучасні версії пакетів із тих, що можуть бути встановлені на наявний комп'ютер. Другий пункт очевидний – вибирати можна з тих пакетів, що доступні. Що стосується характеристик пакету, то тут варто розглянути такі аспекти: а) обчислювальні можливості, б) зручність роботи, в) складність освоєння.

а) Обчислювальні можливості. У випадку, коли необхідно обробляти дані помірних обсягів (до декількох тисяч спостережень) стандартними статистичними методами, то найкраще використовувати універсальні пакети. Проте завжди варто переконатися, що обраний пакет містить необхідні методи обробки.

б) Зручність роботи. Всі сучасні пакети досить зручні в роботі (коли вони вже освоєні).

в) Складність освоєння. За складністю освоєння пакети дещо відрізняються і тут варто віддати перевагу пакетам, з яких є доступні джерела або є ймовірність пройти курс навчання.

1.4 Аналіз завдання на магістерську роботу та постановка задач

В результаті аналізу цитологічних зображень інфільтративного раку молочної залози встановлено, що зображення характеризуються великою кількістю ознак. Тому при обробці зображень необхідно створювати методи формування і редукції вибірок, що дозволить обробляти вихідні вибірки великого обсягу.

Показано, що методами аналізу даних є методи відбору інформативних ознак, методи формування і редукції вибірок, які забезпечують збереження в сформованій підвибірці важливих для наступного аналізу властивостей вихідної вибірки. Це дозволить суттєво зменшити обсяг вибірки, зменшити вимоги до ресурсів ЕОМ.

Необхідно розробити програмне забезпечення, що реалізує запропонований метод формування та редукції вибірок, а також провести експерименти по їх дослідженню, результати яких дозволяють рекомендувати розроблений метод для використання на практиці при вирішенні завдань інтелектуального аналізу даних.

Метою роботи є розробка алгоритмів редукції інформативних ознак цитологічних зображень на основі методу головних компонент.

Для досягнення поставленої мети необхідно розв’язати такі задачі:

1. Проаналізувати інформативні ознаки цитологічних зображень.
2. Проаналізувати програмні засоби аналізу даних.
3. Розробити алгоритм методу головних компонент
4. Розробити математичну модель методу головних компонент.
5. Розробити алгоритми перевірки статистичних гіпотез в компонентному аналізі.
6. Спроектувати і програмно реалізувати систему факторного аналізу.
7. Розробити модуль інтерпретації результатів факторного аналізу.

2 АЛГОРИТМИ МЕТОДІВ ФАКТОРНОГО АНАЛІЗУ

2.1 Алгоритм методу головних компонент

В аналізі даних, як і в будь-якому іншому аналізі, часом буває не зайвим створити спрощену модель, яка максимально точно описує реальний стан справ. Часто буває так, що ознаки досить сильно залежать одна від одної і їх одночасна наявність надмірна.

Наприклад, витрата палива у нас вимірюються в літрах на 100 км, а в США – миль на галон. На перший погляд, різні величини, але насправді вони строго залежать одна від одної. Милі 1600м, а в галоні 3.8 л. Одна ознака строго залежить від іншої, знаючи одну, знаємо і іншу.

Але набагато частіше буває так, що ознаки залежать одна від одної не так суворо і (що важливо!) не так явно. Об'єм двигуна в цілому позитивно впливає на розгін до 100 км/год, але це вірно не завжди. А ще може виявитися, що з урахуванням непомітних на перший погляд факторів (типу поліпшення якості палива, використання більш легких матеріалів та інших сучасних досягнень), вік автомобіля не сильно, але теж впливає на його розгін.

Знаючи залежності та їх силу, ми можемо висловити кілька ознак через одну, злити воєдино, так сказати, і працювати вже з більш простою моделлю. Звичайно, уникнути втрат інформації, швидше за все, не вдасться, але мінімізувати її нам допоможе якраз метод головних компонент.

Висловлюючись більш строго, даний метод апроксимує *n*-розмірну хмару спостережень до еліпсоїда (теж *n*-мірного), півосі якого і будуть майбутніми головними компонентами. І при проекції на осі (зниження розмірності) зберігається найбільша кількість інформації.

Теоретичні аспекти методу головних компонент (МГК) розроблені були вже давно (1901 р.) англійським статистиком К. Пірсоном, але практичне використання не було поширене через труднощі обчислення і складність інтерпретації. З появою комп’ютерних технологій цей метод набуває

популярності. Метод головних компонент має ряд переваг порівняно з іншими, а також недоліки. Виділимо насамперед переваги. По-перше, головні компоненти дозволяють візуалізувати складний набір даних і, по-друге, побачити найбільш інформативні змінні. По-третє, головні компоненти є некорельованими змінними. По цій причині МГК часто використовується для боротьби з мультиколінеарністю. По-четверте, головні компоненти дозволяють побачити особливі спостереження.

Негативною стороною МГК є складність математичного апарату, зумовлена абсолютністю знань теорії ймовірностей, математичної статистики, лінійної алгебри. Нерозуміння математичної суті обчислення може призвести до необґрунтованих висновків.

Метод використовується при узагальненні значень елементарних ознак в моделюванні складних причинних комплексів. Множина ознак моделі замінюється меншою кількістю некорельованих величин (головних компонент), які зберігають всю інформацію щодо механізму формування процесу (явища) і не впливають на точність результатів аналізу.

Розв'язування задач методом головних компонент зводиться до поетапного перетворення матриці вихідних даних X – прямокутної таблиці чисел розмірністю m рядків і n стовпців [29-31]:

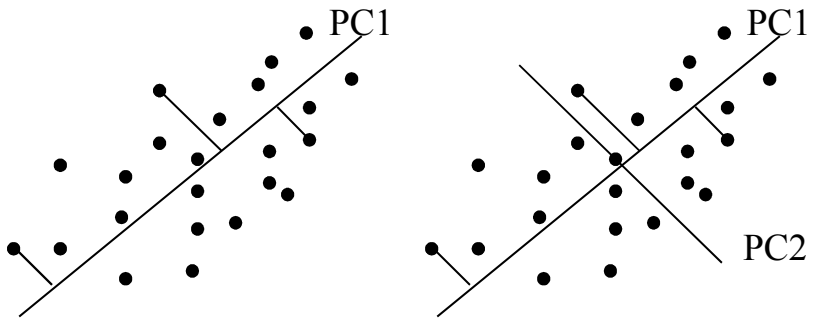
$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{in} \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mj} & \dots & x_{mn} \end{bmatrix}. \qquad (2.1)$$

В більшості випадків стовпці – це змінні (ознаки). Вони нумеруються індексом j ($j = \overline{1,n}$), а рядки – їх реалізації. Вони нумеруються індексом i ($i = \overline{1,m}$).

Метою МГК є витягнення з цих даних потрібної інформації для розв'язуваної задачі. Інформація може як міститися в даних так і не міститися. В першому випадку дані можуть бути надлишковими, містити шум. Що є шумом, а що інформацією залежить від поставленої мети і методів її досягнення. В будь-якому випадку зайвих даних не існує. Краще, коли даних багато, ніж мало.

Шум і надлишковість в даних проявляють себе через кореляційні зв'язки між змінними. Щоб забезпечити адекватність реального процесу, роблять заміну множини ознак на меншу кількість величин (компонент), які є некорельованими і зберігають всю інформацію про процес і не впливають на точність результату аналізу. Всі змінні при цьому враховуються, несуттєва частина даних відділяється, перетворюється в шум.

Розглянемо геометричну суть методу головних компонент для випадку, коли є тільки дві змінні x_1 і x_2 (рисунок 2.1). Такі дані легко зобразити на площині.



а) б)

Рисунок 2.1– Геометрична суть методу головних компонент

Нехай дані приведені на діаграмі розсіювання (рисунок 2.1(а)). Проведемо через точки діаграми пряму, так, щоб уздовж неї відбувалася максимальна зміна даних. Ця пряма називається першою головною компонентою – PC1. Знайдемо проекції всіх вихідних точок на пряму (це показано для трьох точок). Якщо припустити, що насправді всі наші експериментальні точки мали б лежати на цій новій осі, тоді всі відхилення від нової осі можна вважати шумом. Щоб перевірити чи це шум, чи все ще важлива частина даних, необхідно знайти в них вісь максимальних змін. Для цього обчислюється перпендикуляр ортогональної

лінії до PC1 (рисунок 2.1(б)). Ця лінія є другою головною компонентою (PC2). І так треба діяти, до тих пір, поки шум вже не стане дійсно шумом, тобто випадковим хаотичним набором величин.

В багатовимірному випадку, процес виділення головних компонент відбувається аналогічно.

В результаті застосування методу головних компонент до матриці X істотно знижується її розмірність. Вихідна матриця X замінюється двома новими матрицями T і P (рисунок 2.2), розмірність яких A є меншою, ніж число змінних (стовпців) n у вихідній матриці X. Якщо декомпозиція виконана правильно – розмірність A обрана вірно, то матриця T несе в собі стільки ж інформації, скільки її було на початку, в матриці X. При цьому матриця T менша, і, отже, простіша, ніж X.

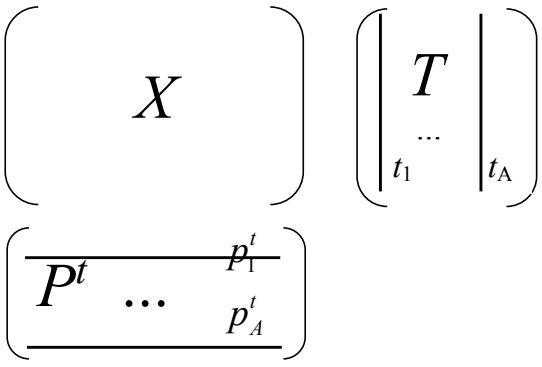


Рисунок 2.2 – Декомпозиція матриці X

В МГК використовуються нові змінні $t_a = p_{a1}x_1 + ... + p_{an}x_n$ $(a = \overline{1, A})$, що є лінійною комбінацією вихідних змінних x_j $(j = \overline{1, n})$.

Матриця X розкладається на добуток двох матриць T і P:

$$X = TP^t + E = \sum_{a=1}^{A} t_a p_a^t + E,$$

де T – матриця ваг, її розмірність – $(m \times A)$, P – матриця навантажень, її

розмірність – ($n \times A$), E – матриця залишків, її розмірність – ($m \times n$).

Змінні t_a називаються головними компонентами, число A – числом головних компонент.

На рисунку 2.3 приведено алгоритм методу головних компонент. Опишемо кроки реалізації методу.

Крок 1. Центрування і нормування вихідних даних за формулою:

$$\frac{x_{ij} - \bar{x}_j}{\sigma_j},$$

де j – номер вихідної змінної, i – номер реалізації j-ої змінної.

Цей крок є підготовчим і обов'язковим в тому випадку, коли різні одиниці вимірювання, або великий розкид значень у вихідних даних.

Крок 2. Обчислення матриці коваріацій (S) або кореляцій (R) (в нашому випадку обчислюється матриця коваріацій).

Крок 3. Знаходження власних чисел λ_1, λ_2, …, λ_p матриці S (або R) з характеристичного рівняння

$$\det(S - \lambda E) = 0 \text{ або } \det(R - \lambda E) = 0,$$

де E – одинична матриця (квадратна матриця на діагоналі якої стоять одиниці, решта елементів – нулі).

Крок 4. Знаходження для кожного власного числа λ_j власного вектора. Власний вектор – це розв'язок системи рівнянь:

$$(S - \lambda E) \cdot \vec{w} = 0, \text{ або } (R - \lambda E) \cdot \vec{w} = 0,$$

де $\vec{w}$ – власний вектор.

Рисунок 2.3 – Схема алгоритму методу головних компонент

Крок 5. Знаходження лінійних комбінацій для головних компонент y_j:

$$y_j = w_{1j}x_1 + w_{2j}x_2 + ... + w_{pj}x_p .$$

Крок 6. Аналіз вкладу кожної головної компоненти і їх ранжування за зростанням.

Загальний вигляд таблиці аналізу вкладу кожної компоненти і їх ранжування за зростанням приведено в таблиці 2.1. Вклад кожної компоненти обчислюється за формулою:

$$I_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p},$$

де j – номер компоненти.

Таблиця 2.1 – Аналіз вкладу кожної компоненти і їх ранжування за зростанням

Головна компонента	y_1	y_2	…	y_p
Власне число	λ_1	λ_2	…	λ_p
Вклад головної компоненти	I_1	I_2	…	I_p
Сумарний вклад	I_1	I_1+I_2	…	$I_1+I_2+…+I_p$

Розрахунок факторних навантажень. Факторні навантаження – коефіцієнти кореляції кожної з аналізованих змінних із кожним виділеним фактором (компонентою). Чим тісніший зв'язок змінної з фактором, тим більшим є її

факторне навантаження. Додатний знак факторного навантаження вказує на прямий зв'язок змінної з компонентою, від'ємний – на обернений.

2.2 Математична модель методу головних компонент

Щоб записати математичну модель МГК введемо позначення:

k – номер змінної (ознаки) ($k = \overline{1,m}$);

i – номер досліджуваного об'єкту (спостереження) ($i = \overline{1,n}$);

L – номер головної компоненти ($L = \overline{1,m}$);

x_{ki} – значення i-ої реалізації k-ої змінної;

F_{Li} – значення L-ої головної компоненти в i-ого об'єкту;

a_{kL} – коефіцієнт парної кореляції між k-ою ознакою і L-ою головною компонентою (факторна нагрузка);

z_{ki} – значення k-ої стандартизованої ознаки в i-ого об'єкту.

Тоді основне рівняння методу має вигляд:

$$Z = AF \tag{2.2}$$

де Z – матриця стандартизованих значень спостережених ознак розміру $m \times n$;

A – матриця факторних навантажень розміру $m \times m$;

F – матриця головних компонент розміру $m \times n$.

Матриця Z в моделі (2.2) є транспонованою. Тут рядки відповідають окремим ознакам, а стовпці – окремим об'єктам.

(2.2) є лінійним перетворенням змінних $F_1, F_2, \ldots, F_m$ в змінні $z_1, z_2, \ldots, z_m$. Якщо матриця факторних навантажень A невироджена, то матриця головних компонент F визначається з рівняння (2.2) наступним чином:

$$F = A^{-1}Z \tag{2.3}$$

В рівняннях (2.2) і (2.3) A і F є невідомі. Відомою є лише матриця Z. Тому для розв'язання цих рівнянь необхідні ще додаткові обмеження. Введення обмежень є основною посилкою всіх методів факторного аналізу і в тому числі методу головних компонент.

Розглянемо матрицю коефіцієнтів R. Для неї виконується вираз:

$$R = Z^T Z / n \tag{2.4}$$

де Z^T – транспонована матриця до матриці Z;

Z – матриця стандартизованих ознак.

Для моделі (2.2) рівняння (2.4) набуде наступного виду:

$$R = ZZ^T / n \tag{2.5}$$

Підставимо (2.2) в (2.5):

$$R = ZZ^T / n = AF(AF)^T / n = AFF^T A^T / n = A(FF^T / n)A^T \tag{2.6}$$

За аналогією до (2.5) можна стверджувати, що

$$FF^T / n = R_F \tag{2.7}$$

є матрицею кореляції, що відображає зв'язки між головними компонентами. Тоді (2.6) набуде виду:

$$R = AR_F A^T . \tag{2.8}$$

І, на кінець, врахувавши, що головною вимогою методу головних компонент є некорельованість головних компонент ($R_F = I$, де I – одинична матриця розміру $m \times m$), маємо:

$$R = AA^T . \qquad (2.9)$$

Умова (2.9) є додатковим обмеженням, що дозволяє знайти єдиний розв'язок рівнянь (2.2) і (2.3). З (2.9) випливає, що ранг A має бути рівним рангу R.

Так як елементи матриці A є коефіцієнтами парної кореляції між головними компонентами і вихідними ознаками, тобто $a_{kL} = r_{kL}$. Тоді $-1 \le a_{kL} \le 1$.

Якщо a_{kL} по модулю близький до одиниці ($|a_{kL}| \ge 0,7$), то зв'язок між k-ою ознакою і L-ою головною компонентою є тісним. В такому випадку, говорять, що k-а ознака навантажує L-у головну компоненту. Якщо $|a_{kL}| \approx 0$, k-а ознака не зв'язана з L-ою головною компонентою.

Компонента називається генеральною, якщо вона тісно зв'язана з усіма без винятку вихідними ознаками. Компонента називається груповою, якщо вона тісно зв'язана більше, ніж з одною вихідною ознакою.

2.3 Алгоритми перевірки статистичних гіпотез в компонентному аналізі

Статистичними гіпотезами називають припущення про можливі значення параметрів розподілу випадкової величини в генеральній сукупності. Перевірка та аналіз статистичних гіпотез здійснюються в результаті збору і побудови статистики. Інструментом в такій роботі виступають статистичні тести, або критерії, кожен з яких представляє собою деякий набір стандартизованих правил.

На основі цих правил приймається рішення про істинність або хибність статистичної гіпотези.

Якщо припустити, що матрицю X утворюють дані, що представляють собою вибірку з деякої реальної або гіпотетичної сукупності з нормальним законом розподілу змінних x_1, x_2, …, x_m в ній, то всі параметри і матриці методу головних компонент є вибіркові оцінки їх істинних значень. Тому необхідно перевіряти важливі статистичні гіпотези відносно значущості окремих матриць і параметрів компонентного аналізу.

Розглянемо наступну нульову гіпотезу $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_m$ проти альтернативної $H_1 : \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ [31]. Статистика

$$\chi_p{}^2 = -[n - \frac{1}{6}(2m + 5)] \ln \det[R]$$

підлягає χ^2-розподілу з рівнем значущості α і $K = \frac{1}{2} m(m-1)$ ступенями вільності.

Гіпотезі $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_m$ проти альтернативи $H_1 : \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ відповідає гіпотеза $H_0 : r = I$ проти альтернативи $H_1 : r \neq I$. Тому даний критерій дозволяє перевірити статистичну значущість всієї матриці коефіцієнтів парної кореляції, відмінність її від одиничної матриці. Якщо виясниться, що кореляційні зв'язки між ознаками дуже слабкі, то проводити компонентний аналіз взагалі немає змісту [32-35].

Порівнюється розрахункове χ_p^2 і табличне значення $\chi_{кр}^2$, знайдене за таблицми χ^2-розподілу з заданим рівнем значущості α і числом ступенів вільності $K = \frac{1}{2} m(m-1)$. Якщо $\chi_p^2 \leq \chi_{кр}^2$ то нульова гіпотеза не відхиляється, можна вважати, кореляційна матриця [R] не відрізняються від одиничної. Значить кореляційні зв'язки між ознаками дуже слабкі і проводити компонентний аналіз взагалі немає змісту.

Якщо $\chi_p^2 > \chi_{кр}^2$, то нульова гіпотеза відхиляється, можна вважати з достовірністю $1-\alpha$, що кореляційна матриця [R] відрізняються суттєво від одиничної. Значить кореляційні зв'язки між ознаками достатньо тісні і проводити компонентний аналіз можна.

Різновидом цієї перевірки є також простіший критерій для перевірки значущості всієї матриці коефіцієнтів парної кореляції:

$$\chi_p{}^2 = n\sum_{i<j} r_{ij}^2 ,\ i, j = 1, 2, ..., m$$

який застосовується аналогічно вище описаному критерію.

3 КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ АЛГОРИТМІВ РЕДУКЦІЇ ІНФОРМАТИВНИХ ОЗНАК ЦИТОЛОГІЧНИХ ЗОБРАЖЕНЬ

3.1 Програмно-апаратні засоби

Для побудови алгоритмів редукції інформативних ознак цитологічних зображень необхідно правильно підібрати програмно-апаратний комплекс для швидкої та безперебійної роботи системи. Характеристики апаратного забезпечення наведено в таблиці 3.1.

Таблиця 3.1 – Характеристики апаратного забезпечення

Система	Характеристики
Виробник	Dell Inc.
Модель	Latitude
Процесор	Intel(R) Core(TM) i5-2520M CPU 2,50GHz
Оперативна пам'ять	4,00 ГБ
Тип системи	64-розрядна ОС на базі процесора x64
Операційна система	Windows 10 Pro

Програмним засобом для проведення експерименту є Matlab. Система Matlab – математична лабораторія, представляє собою інтегроване програмне середовище для виконання числових розрахунків, комп'ютерного моделювання і обчислювальних експериментів, які в тій чи іншій мірі охоплюють області класичної та сучасної математики, спектр інженерних додатків [46]. В таблиці 3.2 наведено основну інформацію про використовуваний програмний засіб Matlab.

Пакет STATISTICA – це програмний комплекс, призначений для проведення статистичного аналізу і володіє широким набором функцій. Серед них – тисячі типів графіків, аналіз даних, факторний аналіз, використання різних критеріїв, непараметричних статистик, коваріаційний і дисперсійний аналіз, таблиці частот і ін. [36-38].

Таблиця 3.2 – Характеристики програмного засобу Matlab

Програмне забезпечення	Характеристики
Розробник	The MathWorks
Версія	R2015a (8.5.0.197613)
Тип ПЗ	64-bit (win64)
Зчитувані формати файлів	Файл MATLAB-M
Створювані формати файлів	Файл MATLAB-M

Пакет STATISTICA – це програмний комплекс, призначений для проведення статистичного аналізу і володіє широким набором функцій. Серед них – тисячі типів графіків, аналіз даних, кореляційний аналіз, використання Т-критеріїв (і не тільки їх), непараметричні статистики, коваріаційний і дисперсійний аналіз, застосування множинної регресії, таблиць частот, і ін. [39-42].

Переваги пакету:

- широкий вибір різноманітних інструментів для статистичного аналізу даних;
- дозволяє обробити масивні обсяги даних;
- можливість проведення кластерного, дискримінантного, канонічного, дисперсійного і факторного аналізу;
- наявність функції угрупування даних;
- можливість побудови рядів, лінійних і нелінійних моделей;
- великий набір інструментів для дослідження кореляції між різними змінними;
- підтримка нейронних мереж;
- можливість імпортування даних з Excel-документів;
- наявність блоку інструментів data mining;
- можливість побудови діаграм і 2D / 3D графіків (більше 10000 типів);
- можливість категоризації по змінним (ієрархічна угруповання);
- якісна технічна підтримка;

- інтеграція з Visual Basic (можна використовувати DLL-бібліотеки) і ін.

3.2 Структура програмної системи

В даному підрозділі на базі алгоритмів методу головних компонент розроблена програмна система редукції інформативних ознак цитологічних зображень (рисунок 3.1) за алгоритмом рисунку 3.2.

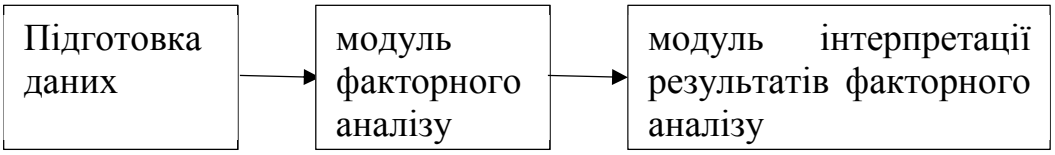


Рисунок 3.1 – Програмна система редукції інформативних ознак

Рисунок 3.2 – Етапи компонентного аналізу

3.3 Модуль факторного аналізу

Модуль Factor Analysis (Факторний аналіз) містить широкий набір методів,

які дають користувачу вичерпні можливості для виділення факторів і представлення результатів (рисунок 3.3).

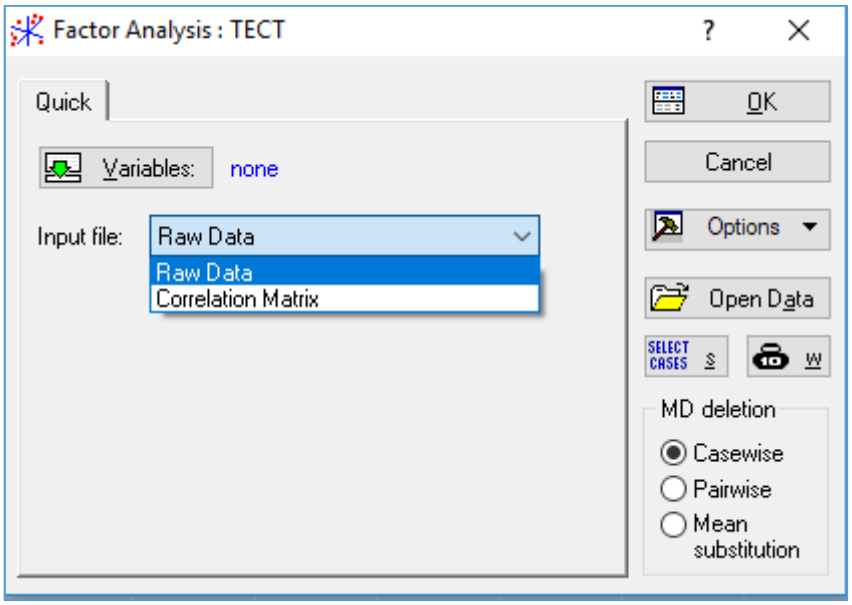


Рисунок 3.3 – Діалогове вікно Factor Analysis

Вхідні змінні:

- площа;
- периметр;
- довжина;
- ширина;
- окружність;
- координата Xc;
- координата Yc;
- довжина головної осі;
- довжина побічної осі;
- кут нахилу головної осі до осі OX;
- периметр обмежуючого прямокутника;
- координата Bx;
- координата By;
- ширина обмежуючого прямокутника;
- довжина обмежуючого прямокутника;

- площа прямокутника;
- відношення сторін

Числові значення показано на рисунку 3.4.

	1 площа	2 периметр	3 довжина	4 ширина	5 окружність	6 координата Xc	7 координата Yc	8 довжина головної осі	9 довжина побічної осі	кут нахилу головної осі до	периметр обмежуючого прямоку	12 координата Bx	13 координата By	ширина обмежуючого прямоку	довжина обмежуючого прямоку	16 площа прямокутника	17 відношення сторін
1	3164	224,6518	76	59	39760	1402,494	2230,526	81,55885	50,1784	24,14214	63,4707	1362	2183	82	97	7954	1,625378
2	3231,5	231,4386	50	85	40608,23	3139,049	2138,653	88,535	46,09697	101,5564	64,14416	3091	2107	98	65	6370	1,920625
3	3862,5	174,0244	54	52	26716,1	1499,627	2027,167	54,27808	50,67153	131,6197	52,02795	1462	1990	76	76	5776	1,071175
4	4408,5	259,865	73	85	55398,84	204,9863	1738,779	95,47722	60,39172	120,0496	74,92047	148	1688	115	102	11730	1,580965
5	1934	167,8822	48	52	24303,36	2697,128	1072,14	51,65514	49,6713	156,1781	49,62303	2663	1038	69	69	4761	1,039939
6	4658,5	258,6934	70	88	58540,44	497,8377	1142,093	87,08671	70,67023	102,0924	77,01549	447	1098	102	89	9078	1,232297
7	3303	224,1665	66	68	41506,72	3015,015	2275,642	75,24791	57,23547	125,1439	64,84991	2967	2230	97	92	8924	1,314708
8	1855	175,397	39	63	23310,62	3032,455	559,8697	62,55547	40,2323	83,05925	48,59896	2998	536	69	49	3381	1,554857
9	4228	249,3381	64	82	53130,61	898,5469	1037,602	84,79321	66,5555	94,30136	73,37068	853	1001	92	74	6808	1,274023
10	1654	161,5391	55	42	20784,78	1664,283	2313,565	54,27321	40,06574	168,8672	45,8905	1639	2283	52	63	3276	1,354604
11	2062	193,3381	74	44	25911,86	1487,539	1329,319	76,46916	35,7642	20,48398	51,23885	1457	1287	62	86	5332	2,138148
12	4632	262,3087	78	78	58207,43	743,8984	1021,752	82,63897	71,77	42,6727	76,79613	689	967	111	111	12321	1,151442
13	2528,5	200,468	46	70	31774,07	1934,758	1603,423	70,13614	48,31947	104,0584	56,73964	1894	1571	82	66	5412	1,451509
14	1296	139,397	38	46	16286,02	2090,577	1150,805	44,74844	36,94402	111,8568	40,62165	2062	1125	58	53	3074	1,21125
15	2782	212,3675	66	58	34959,64	496,3932	581,1967	76,55018	48,68476	46,87732	59,51598	451	537	91	90	8190	1,572364
16	4204,5	261,0366	64	94	52835,31	2764,293	2179,856	91,01789	61,95533	100,0674	73,16649	2714	2141	102	79	8058	1,469089
17	3862,5	254,267	52	95	48537,61	710,1561	2273,507	98,27068	52,20684	91,21426	70,12765	660	2246	101	56	5656	1,882333

Рисунок 3.4 – Фрагмент файлу даних

I етап факторного аналізу – обчислення кореляційної матриці.

Верхня частина вікна Define Method of Factor Extraction (Визначити метод виділення чинників), зображеного на рисунку 3.5, є інформаційною. Вона містить дані про вибраний метод обробки пропущених значень – Casewise, кількість оброблених спостережень – 17, кількість спостережень, придатних для подальших обчислень – 17. Кореляційна матриця обчислена для 17 змінних.

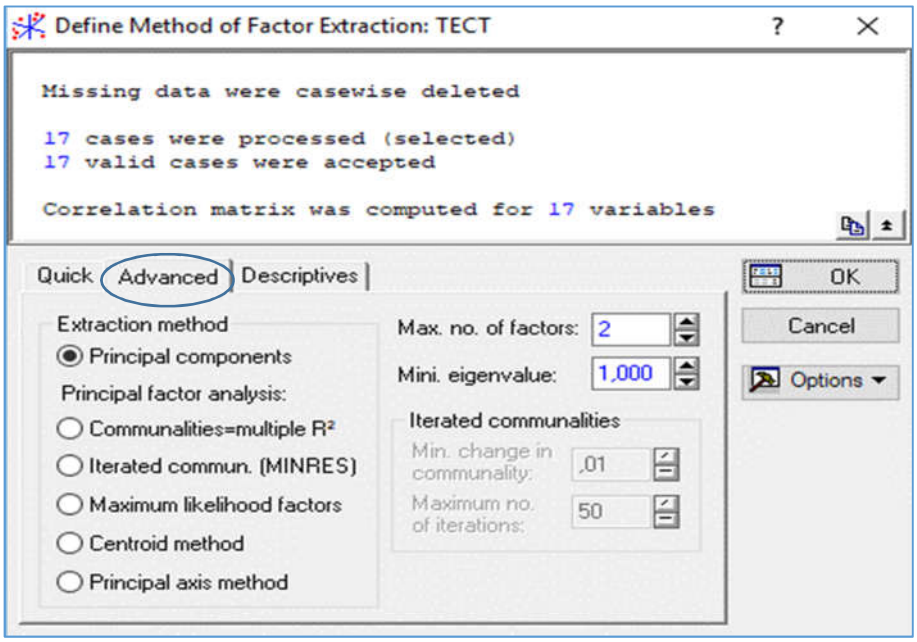


Рисунок 3.5 – Вкладка Advanced вікна вибору методу виділення факторів

У вкладці Descriptives вікна методу виділення факторів (рисунок 3.6) можна обчислити середні, стандартні відхилення, кореляції, коваріації, побудувати різноманітні графіки, провести додатковий аналіз поточних даних, перевірити відповідність вибіркових змінних нормальному закону розподілу і т.д. (рисунок 3.7).

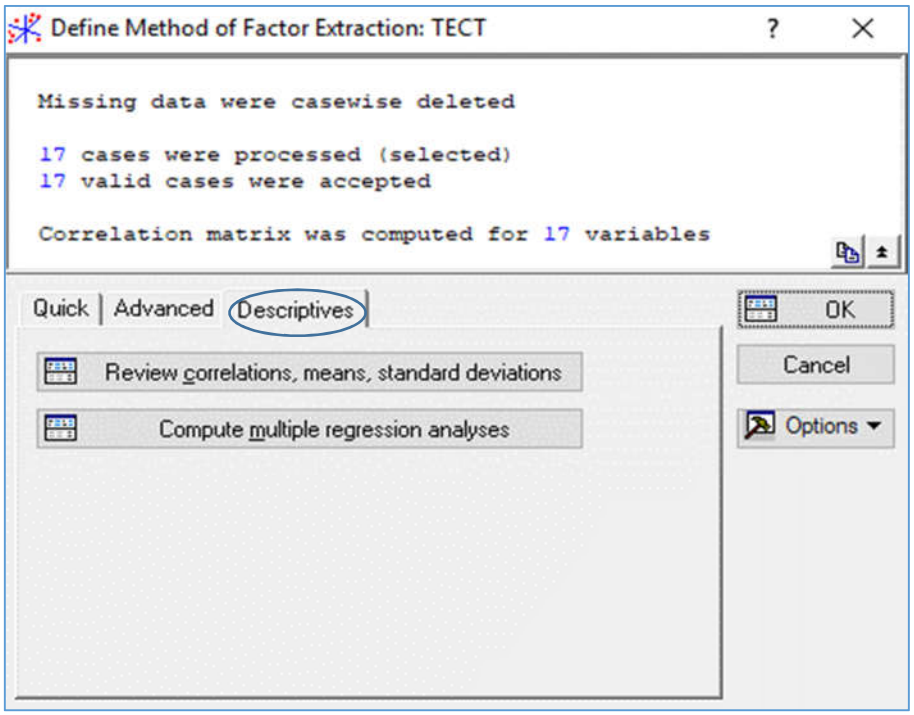


Рисунок 3.6 – Вкладка Descriptives вікна вибору методу виділення факторів

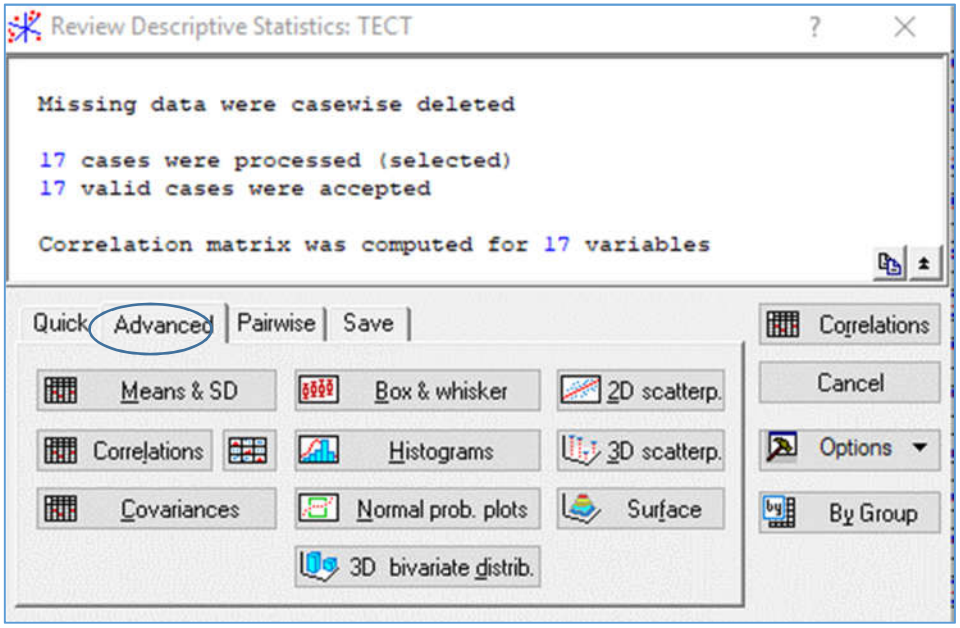


Рисунок 3.7 – Вкладка Advanced вікна Descriptives Statistics

Матриця кореляцій для даних рисунку 3.4 показана на рисунку 3.8, знайдена за екранною кнопкою Correlations (рисунок 3.6).

Вибравши екранну кнопку Principal components (рисунок 3.5) отримуємо результати факторного аналізу (рисунок 3.9).

Correlations (TECT)
Casewise deletion of MD
N=17

Variable	площа	периметр	довжина	ширина	окружність	координата Xc	координата Yc	довжина головної осі	довжина побічної осі	кут нахилу головної осі до осі OX	пер обмеж прямо
площа	1,00	0,89	0,60	0,79	0,93	-0,48	0,18	0,75	0,90	-0,15	
периметр	0,89	1,00	0,65	0,88	0,98	-0,41	0,17	0,94	0,81	-0,32	
довжина	0,60	0,65	1,00	0,22	0,64	-0,55	0,09	0,58	0,56	-0,51	
ширина	0,79	0,88	0,22	1,00	0,87	-0,18	0,18	0,84	0,71	-0,06	
окружність	0,93	0,98	0,64	0,87	1,00	-0,46	0,11	0,87	0,90	-0,23	
координата Xc	-0,48	-0,41	-0,55	-0,18	-0,46	1,00	0,19	-0,38	-0,41	0,30	
координата Yc	0,18	0,17	0,09	0,18	0,11	0,19	1,00	0,23	-0,03	0,29	
довжина головної осі	0,75	0,94	0,58	0,84	0,87	-0,38	0,23	1,00	0,58	-0,41	
довжина побічної осі	0,90	0,81	0,56	0,71	0,90	-0,41	-0,03	0,58	1,00	-0,03	
кут нахилу головної осі до осі OX	-0,15	-0,32	-0,51	-0,06	-0,23	0,30	0,29	-0,41	-0,03	1,00	
периметр обмежуючого прямокутника	0,93	0,99	0,65	0,87	1,00	-0,44	0,13	0,88	0,89	-0,25	
координата Bx	-0,49	-0,42	-0,55	-0,19	-0,46	1,00	0,19	-0,39	-0,42	0,30	
координата By	0,17	0,16	0,08	0,17	0,10	0,19	1,00	0,22	-0,03	0,30	
ширина обмежуючого прямокутника	0,88	0,93	0,51	0,87	0,93	-0,35	0,11	0,84	0,81	-0,19	
довжина обмежуючого прямокутника	0,59	0,57	0,92	0,17	0,59	-0,48	0,01	0,44	0,60	-0,42	
площа прямокутника	0,80	0,80	0,80	0,55	0,83	-0,48	0,03	0,66	0,80	-0,31	
відношення сторін	-0,10	0,21	0,13	0,15	0,02	-0,01	0,21	0,50	-0,39	-0,51	

Рисунок 3.8 – Матриця кореляцій для даних рисунку 3.4

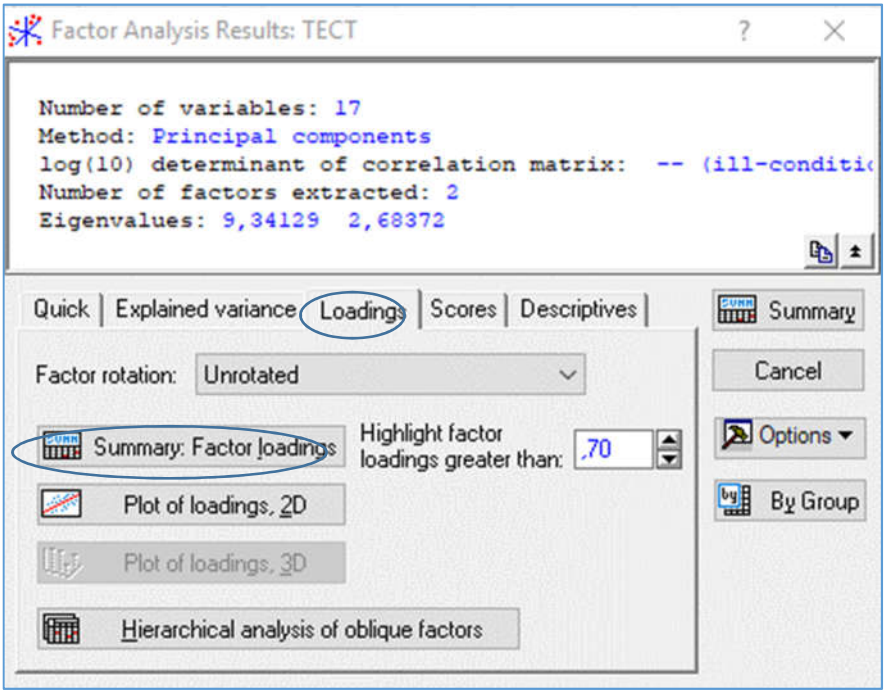


Рисунок 3.9 – Вкладка Loadings вікна результатів факторного аналізу

3.4 Модуль інтерпретації результатів факторного аналізу

Інформаційна частина вікна результатів факторного аналізу містить дані про:

- кількість аналізованих змінних – 17;
- метод – головних компонент;
- десятковий логарифм визначника матриці кореляції;
- кількість виділених факторів – 2;
- власні значення – 9,34129, 2,68372.

Нижня частина вікна – різноманітні числові та графічні можливості перегляду результатів аналізу:

- факторні навантаження;
- факторні поля;
- факторні ваги;
- власні значення факторів.

Будуємо таблицю факторних навантажень (рисунок 3.10), застосувавши екранну кнопку Summary. Factor loadings вікна результатів факторного аналізу (рисунок 3.9). Її рядки – це змінні, а стовпці – виділені компоненти.

Для кожної компоненти вказано навантаження кожної початкової змінної, яке відображає величину проекції змінної на факторну координатну вісь (рисунок 3.10). Факторні навантаження можна інтерпретувати як кореляції між відповідними змінними та компонентами – чим вище навантаження за модулем, тим ближче компонента до вихідної змінної. Факторні навантаження дають найбільш важливу інформацію для інтерпретації отриманих компонент.

Variable	Factor Loadings (Unrotated) (ТЕСТ) Extraction: Principal components (Marked loadings are >,700000) Factor 1	Factor 2
площа	0,930062	0,115608
периметр	0,965469	0,166686
довжина	0,753929	-0,259835
ширина	0,778245	0,393526
окружність	0,974332	0,102113
координата Xc	-0,562509	0,543321
координата Yc	0,105962	0,824972
довжина головної осі	0,870384	0,233751
довжина побічної осі	0,862471	-0,028791
кут нахилу головної осі до осі OX	-0,337040	0,440998
периметр обмежуючого прямокутника	0,977713	0,123953
координата Bx	-0,569342	0,540003
координата By	0,095465	0,829533
ширина обмежуючого прямокутника	0,923290	0,165663
довжина обмежуючого прямокутника	0,718818	-0,319073
площа прямокутника	0,902130	-0,135494
відношення сторін	0,076785	0,203315
Expl.Var	9,341294	2,683717
Prp.Totl	0,549488	0,157866

Рисунок 3.10 – Факторні навантаження

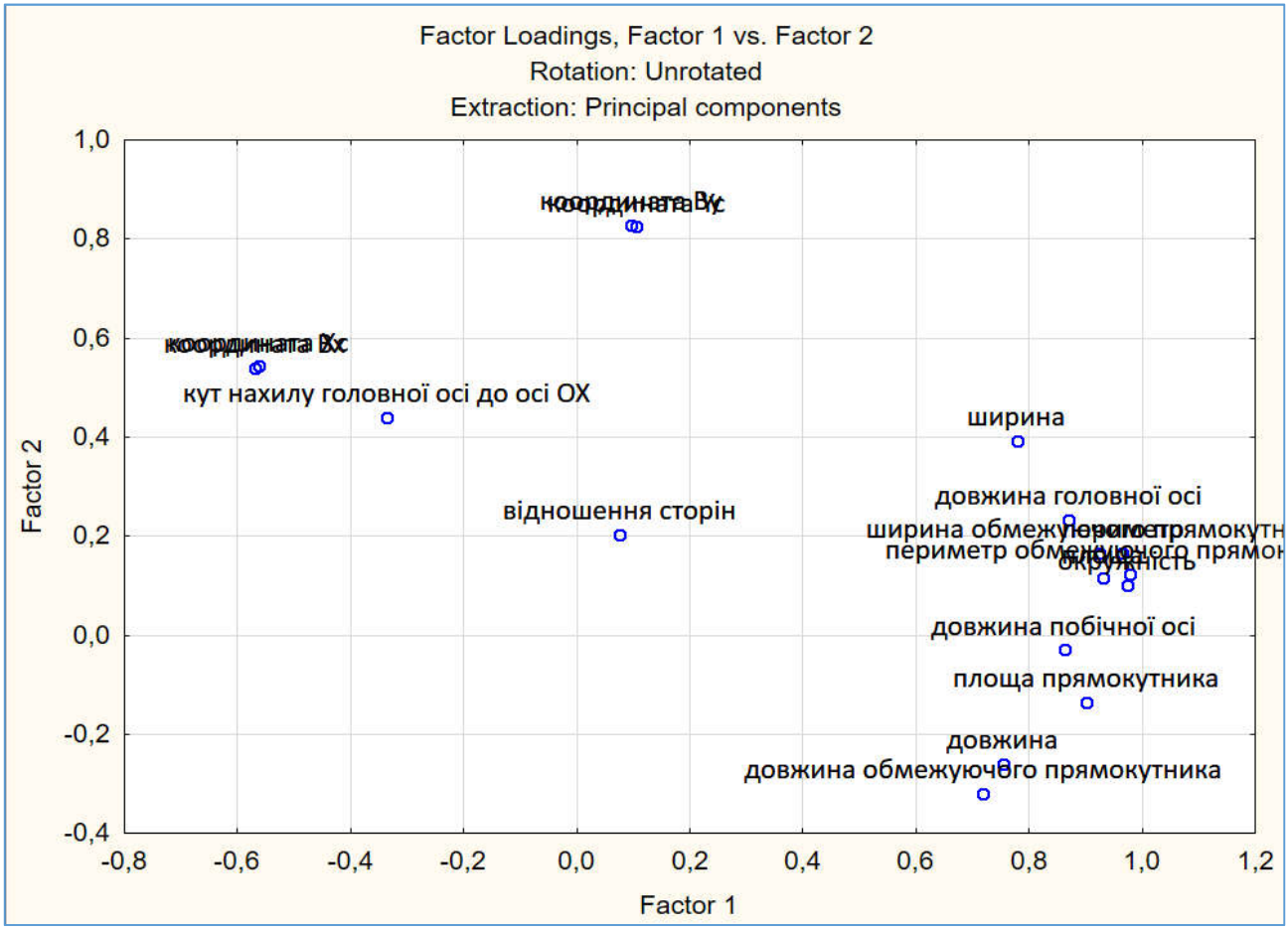


Рисунок 3.11 – Графічне представлення факторних навантажень (факторне поле)

З рисунку 3.10 видно, що виділено дві компоненти. З першою компонентою найтісніше пов’язаними виявилися

- площа;
- периметр;
- довжина;
- ширина;
- окружність;
- довжина головної осі;
- довжина побічної осі;
- периметр обмежуючого прямокутника;
- ширина обмежуючого прямокутника;
- довжина обмежуючого прямокутника;
- площа прямокутника.

Друга виділена компонента пов’язана лише зі змінними:

- координата Yc;
- координата By.

Факторні навантаження видно і у факторному полі (рисунок 3.11).

Факторні ваги – кількісні значення зв’язку виділених компонент з спостереженнями. Щоб їх обчислити, потрібно натиснути кнопку Factor scores у вікні результатів факторного аналізу (рисунок 3.12).

Спостереженню з більшою вагою притаманний більший рівень прояву властивостей виділеної компоненти (більший рівень зв’язку з компонентою). Додатні факторні ваги відповідають спостереженням, які мають рівень прояву властивостей компоненти більше середнього, а від’ємні – нижче середнього.

Найбільші факторні ваги по 1-ій компоненті (рисунок 3.13) в спостережень №14, №4, по 2-ій – в спостережень №15, №2.

Для встановлення, яка з компонент є найбільш значущою проведемо аналіз власних значень компонент (рисунки 3.14, 3.15).

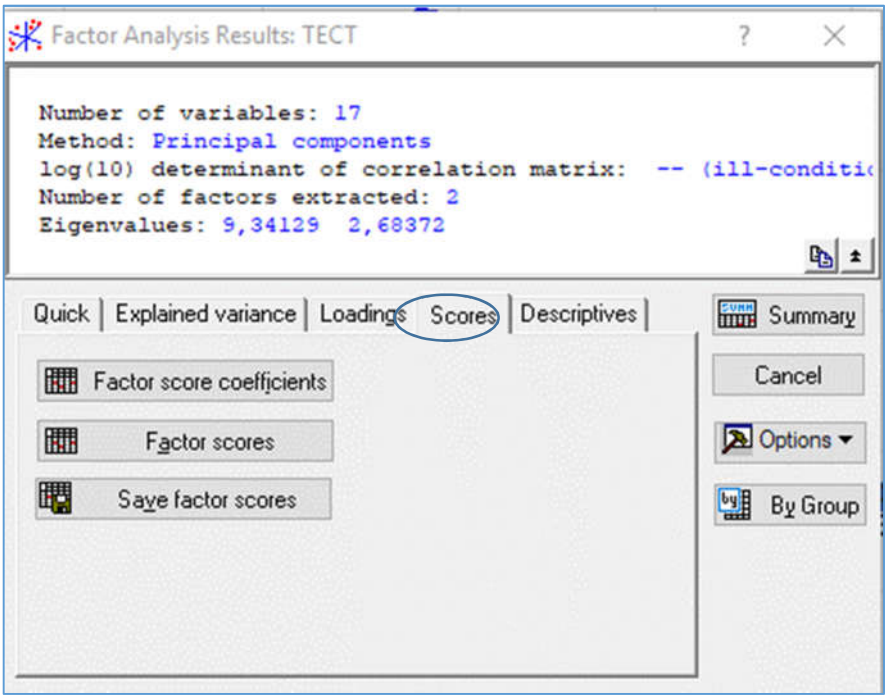


Рисунок 3.12 – Вкладка Scores вікна результатів факторного аналізу

Case	Factor Scores (TECT) Rotation: Varimax normalized Extraction: Principal components Factor 1	Factor 2
1	0,35970	-0,01864
2	0,11631	1,76405
3	-0,54004	0,17194
4	1,40471	-0,22584
5	-1,16149	-0,17644
6	1,18915	-0,72515
7	0,29501	1,19512
8	-1,31476	-0,30556
9	0,70768	-0,60561
10	-1,26571	0,63434
11	-0,60518	-0,92919
12	1,40051	-1,36269
13	-0,45446	0,31769
14	-1,75586	-0,44589
15	0,02471	-1,87686
16	0,88112	1,47611
17	0,71859	1,11260

Рисунок 3.13 – Таблиця факторних ваг

Власні значення – це дисперсії, які пояснюються компонентами.

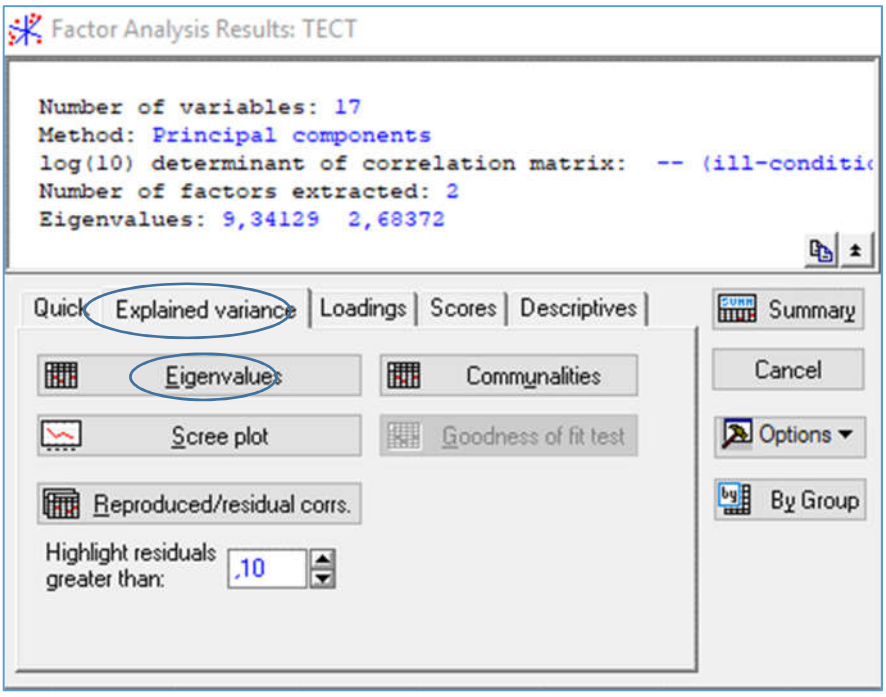


Рисунок 3.14 – Вкладка Explained variance вікна результатів факторного аналізу

Value	Eigenvalues (TECT) Extraction: Principal components			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	9,341294	54,94879	9,34129	54,94879
2	2,683717	15,78657	12,02501	70,73536

Рисунок 3.15 – Власні значення виділених компонент

- Eigenvalue – дисперсії кожної компоненти;
- % Total variance – відсоток від загальної дисперсії для кожної компоненти;
- Cumulative Eigenvalue – накопичена дисперсія виділених компонент;
- Cumulative % - накопичений відсоток від загальної дисперсії.

Накопичений відсоток від загальної дисперсії показує, що дві компоненти описують 70,7% початкової сукупності даних. Перша компонента пояснює 54,95%, а друга – 15,79%. Це означає, що 29,3% дисперсії даних припадає на інші не враховані фактори.

Для проведення якісного факторного аналізу необхідно встановити, скільки факторів необхідно виділити, щоб вони максимально повно описали дані і були значущими.

Рисунок 3.16 зображає графік власних значень кожної виділеної компоненти в порядку спадання – це є критерій кам’янистого насипу.

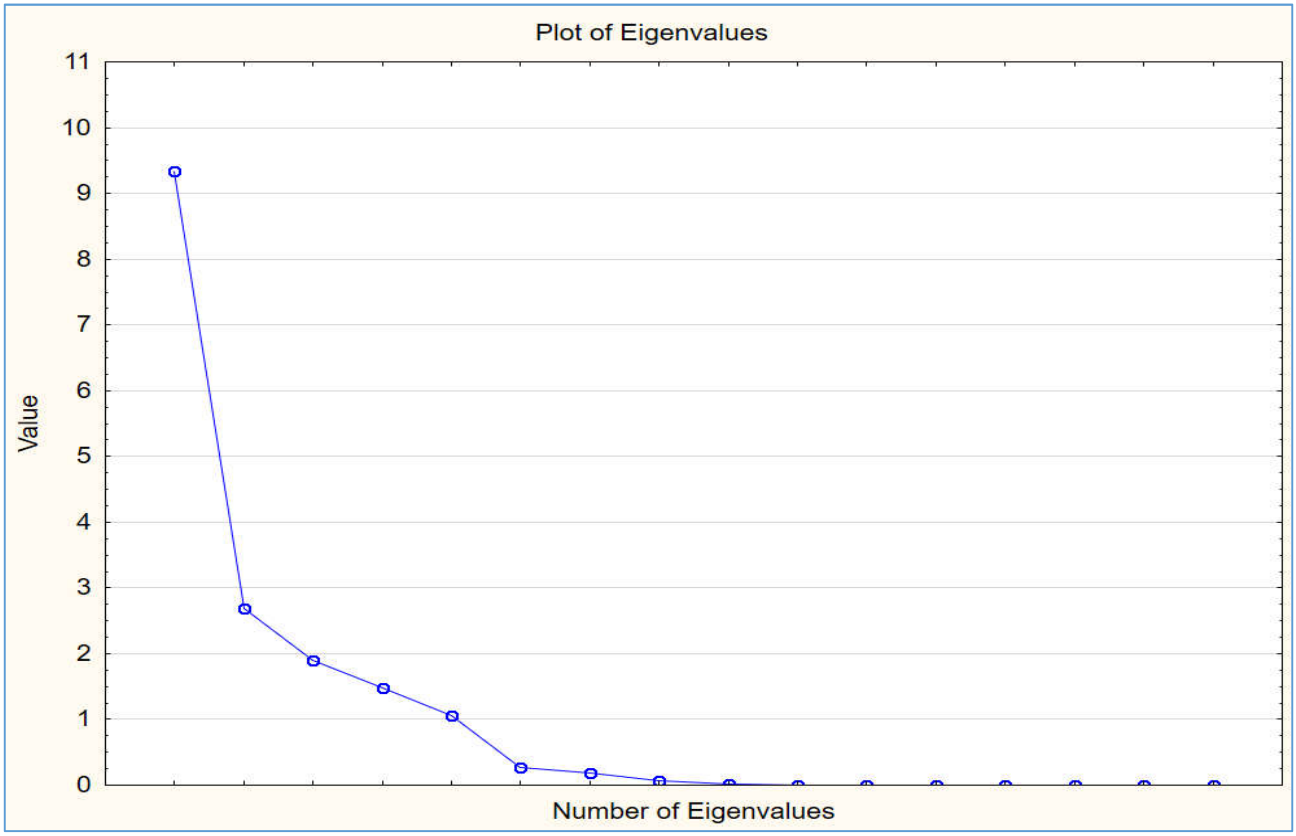


Рисунок 3.16 – Графік власних значень

З графіку видно, що більшими 1 є п’ять власних чисел, а після третьої компоненти графік спадає сповільнено. Тому збільшимо кількість компонент до числа 3.

Таблиця факторних навантажень показана на рисунку 3.17. З третьою компонентою найтісніше пов’язаною виявилася ознака – відношення сторін.

Проведемо аналіз власних значень компонент (рисунок 3.18).

Третя компонента описує 11,18% початкової сукупності даних. Накопичений відсоток від загальної дисперсії зріс при появі третьої компоненти з 70,74% до 81,92%.Це означає, що 18,08% дисперсії даних припадає на інші не враховані фактори.

Аналіз вкладу кожної компоненти і їх ранжування за зростанням приведено в таблиці 3.3.

Variable	Factor Loadings (Unrotated) (ТЕСТ) Extraction: Principal components (Marked loadings are >,700000) Factor 1	Factor 2	Factor 3
площа	0,930062	0,115608	0,202682
периметр	0,965469	0,166686	-0,055015
довжина	0,753929	-0,259835	-0,288082
ширина	0,778245	0,393526	0,142044
окружність	0,974332	0,102113	0,116891
координата Xc	-0,562509	0,543321	0,114089
координата Yc	0,105962	0,824972	-0,210445
довжина головної осі	0,870384	0,233751	-0,327477
довжина побічної осі	0,862471	-0,028791	0,472105
кут нахилу головної осі до осі OX	-0,337040	0,440998	0,663066
периметр обмежуючого прямокутника	0,977713	0,123953	0,084530
координата Bx	-0,569342	0,540003	0,112482
координата By	0,095465	0,829533	-0,209305
ширина обмежуючого прямокутника	0,923290	0,165663	0,124006
довжина обмежуючого прямокутника	0,718818	-0,319073	-0,088216
площа прямокутника	0,902130	-0,135494	0,072055
відношення сторін	0,076785	0,203315	-0,905569
Expl.Var	9,341294	2,683717	1,900034
Prp.Totl	0,549488	0,157866	0,111767

Рисунок 3.17 – Факторні навантаження (3 компоненти)

Value	Eigenvalues (ТЕСТ) Extraction: Principal components Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	9,341294	54,94879	9,34129	54,94879
2	2,683717	15,78657	12,02501	70,73535
3	1,900034	11,17667	13,92504	81,91202

Рисунок 3.18 – Власні значення виділених трьох компонент

Таблиця 3.3 – Аналіз вкладу кожної компоненти і їх ранжування за зростанням

Головна компонента	y_1	y_2	y_3
Власне число	9,341	2,684	1,900
Вклад головної компоненти	54,95	15,79	11,18
Сумарний вклад	54,95	70,74%	81,92%

Отримано 3 компоненти виду:

$$y_1 = 0,930 \times площа + 0,965 \times периметр + 0,754 \times довжина + 0,778 \times ширина +$$

$$+0,974 \times окружність + 0,870 \times довжина\ головної\ осі + 0,862 \times довж.\ побічн.\ осі +$$
$$+0,977 \times периметр\ обмеж.\ прямокутн. + 0,923 \times ширина\ обмеж.\ прямокутн. +$$

$$+0,719 \times довжина\ обмеж.\ прямокутн. + 0,902 \times площа\ прямокутника;$$

$$y_2 = 0,825 \times координата\ \mathrm{Yc} + 0,830 \times координата\ \mathrm{By};$$

$$y_2 = -0,906 \times відношення\ сторін.$$

Таким чином, в даному розділі зроблено комп'ютерне моделювання алгоритмів редукції інформативних ознак цитологічних зображень.

ВИСНОВКИ

1. Проведено аналіз методів, алгоритмів і програмних засобів редукції інформативних ознак цитологічних зображень.
2. Проаналізовано методи відбору інформативних ознак.
3. Проведено аналіз інформативних ознак цитологічних зображень.
4. Розроблено алгоритм методу головних компонент.
5. Розроблено математичну модель методу головних компонент.
6. Розроблено алгоритми перевірки статистичних гіпотез в компонентному аналізі.
7. Здійснено комп’ютерне моделювання алгоритмів факторного аналізу.
8. Розроблено модуль факторного аналізу.
9. Розроблено модуль інтерпретації результатів факторного аналізу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Олійник А. О. Інтелектуальний аналіз даних: навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя: ЗНТУ, 2012. – 271 с.

2. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский; пер. с польск. И. Д. Рудинского. – М.: Горячая линия – Телеком, 2004. – 452 с.

3. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография / [С. А. Субботин, Ан. А. Олейник, Е. А. Гофман, С. А. Зайцев, Ал. А. Олейник; под ред. С. А. Субботина]. – Харьков: ООО «Компания Смит», 2012. – 317 с.

4. Субботин С. А. Формирование выборок и анализ качества моделей на основе нейронных и нейро-нечетких сетей в задачах диагностики и распознавания образов / С. А. Субботин: монография. – Saarbr cken: LAP Lambert academic publishing, 2012. – 232 с.

5. Субботин С.А. Формирование и редукция выборок для интеллектуального анализа данных / С.А. Субботин // Радіоелектроніка, інформатика, управління. – 2013. – № 1. – С. 113-118.

6. Струтинський П.І. Згорткові нейронні мережі як засіб обробки біомедичних зображень / П.І. Струтинський, Н.А. Поворозник, П.М. Матвісів // Матеріали VII міжнародної науково-технічної конференції «Актуальні задачі сучасних технологій», м. Тернопіль, 28–29 листопада 2018 р. – Тернопіль: ТНТУ, 2018. – Т.2. – С. 170-171.

7. Методичні рекомендації до виконання дипломної роботи з освітньо-кваліфікаційного рівня «Магістр». Спеціальність «Комп’ютерні системи та мережі» / О.М. Березький, Л.О. Дубчак, Г.М. Мельник /Під ред. О.М. Березького – Тернопіль: ТНЕУ, 2016.– 47 с.

8. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография / [С. А. Субботин, Ан. А. Олейник, Е. А. Гофман, С. А. Зайцев, Ал. А. Олейник; под ред. С. А. Субботина]. – Харьков : ООО «Компания Смит», 2012. – 317 с.

9. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 339 p.

10. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York: Chapman & Hall, 2005. – 416 p.

11. Encyclopedia of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 p.

12. Кокрен У. Методы выборочного исследования / У. Кокрен; пер. с англ. И. М. Сонина; под ред. А. Г. Волкова, Н. К. Дружинина. – М. : Статистика, 1976. – 440 с.

13. Subbotin S. A. The training set quality measures for neural network learning / S. A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2010. – Vol. 19. – № 2. – P. 126–139.

14. Субботин С. А. Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов / С. А. Субботин // Математичні машини і системи. – 2010. – № 1. – С. 25–39.

15. Субботин С. А. Синтез нейро-нечетких моделей для выделения и распознавания объектов на сложном фоне по двумерному изображению / С. А. Субботин // Комп’ютерне моделювання та інтелектуальні системи: збірник наукових праць за ред. Д. М. Пізи, С. О. Субботіна. – Запоріжжя : ЗНТУ, 2007. – С. 68–91.

16. Автандилов Г. Г. Основы количественной патологической анатомии / Г. Г. Автандилов – М.: Медицина, 2002. – 240 с.

17. Шабалова И.П. Цитологический атлас. Диагностика заболеваний молочной железы / И.П. Шабалова, Т.В. Джангирова, Н.Н. Волченко, К.К. Пугачев. – М.-Тверь: ООО «Издательство «Триада», 2005. – 119 с.

18. Волченко Н. Н. Истинная и ложная гипердиагностика опухолевого процесса при цитологическом исследовании // Российский онкологический журнал. / Волченко Н.Н., Славнова Е.Н., Гладунова З.Д., Ермолаева А.Г. – 2006. – № 6. – С. 29-32.

19. Хмельницкий О.К. Цитологическая и гистологическая диагностика заболеваний шейки и тела матки. – СПб., 2004. – 333 с.

20. Koss L.G. Diagnostic Cytology and its hystopathologic bases/ Koss L.G., Melamed M.R. – Vol. 1-2. NY, 2006. – 1752 p.

21. Dabbs D.J. Diagnostic immunohistochemistry. – USA Philadelphia. – 2006. – 828 p.

22. Gray W., McKee G. T. Diagnostic Cytopathology. USA. Churchill livingstone, – 2004. – 1042 p.

23. Мельник А. Н. Цитоморфологическая диагностика опухолей / А.Н. Мельник. – К.: Здоров'я, 1983. – 240 с.

24. Гойко О. В. Аналіз сучасного програмного забезпечення для статистичного оброблення й аналізу біомедичних досліджень / О. В. Гойко, С. І. Мохначов // Медична інформатика та інженерія. – 2012. – №4. – С. 49-52.

25. Боровиков В. Statistica. Искусство анализа данных на компьютере / В. Боровиков. - СПб.: Питер, 2001. - 656 с.

26. Гойко О. В. Практичне використання пакета STATISTICA для аналізу медико-біологічних даних : навчальний посібник для студентів вищих навчальних закладів / О. В. Гойко. - Київ, 2004. - 76 с.

27. Кулаічев А. П. Методи і засоби аналізу даних у середовищі Windows /А. П. Кулаічев. - М. : Інко, 2002. - 341 с.

28. Лапач С. Н. Статистические методы в медико-биологических исследованиях с использованием Excel / С. Н. Лапач, А. В. Чубенко, П. Н. Бабич. - Издательство "Морион Лтд", 2000. - 320 с.

29. «Компонентний аналіз» [Електронний ресурс] – Режим доступу: https://stud.com.ua/93349/statistika/komponentniy_analiz

30. «Метод Главных Компонент (PCA)» [Електронний ресурс] – Режим доступу: http://bourabai.kz/cm/pca.htm

31. Янковой А. Г. Многомерный анализ в системе STATISTICA / А. Г. Янковой. – Одесса: Оптимум, 2002. Вып. 2. – 325 с.

32. Магнус Я. Р. Эконометрика. Начальный курс: Учебник. – 6-е изд., перераб. и доп. / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М.: Дело, 2004. – 576 с.

33. Грубер Й. Эконометрия. – К.: 1996. – Т.1. Введение в эконометрию. – 400 с.

34. Маленво Э. Статистические методы эконометрии. – М.: Статистика, 1975. – 423 с.

35. Джонстон Дж. Эконометрические методы. – М.: Статистика, 1980. – 444 с.

36. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL: учеб.пособ.– 2-е изд., испр. и доп. – М.: Форум, 2008.– 464 с.

37. Майборода Р.Є., Сугакова О.В. Статистичний аналіз даних за допомогою пакету STATISTICA [Електронний ресурс]. – Режим доступу: http://matphys.rpd.univ.kiev.ua/downloads/courses/mmatstat/StatAn.doc.

38. Електронний учебник по статистике [Електронний ресурс]. – Режим доступу: http://statsoft.ru/home/textbook/default.htm.

39. Роїк, М. В. Огляд програмних засобів статистичного аналізу даних / М. В. Роїк, О. І. Присяжнюк, В. О. Денисюк // Ефективна економіка. – 2017. – №7, [Електронний ресурс]. – Режим доступу: http://www.economy.nayka.com.ua/?op=1&z=5676

40. System requirements [Electronic resourse]. – Mode of access: https://en.wikipedia.org/wiki/System_requirements.

41. Статистические ППП [Електронний ресурс]. – Режим доступу: http://helpiks.org/6-54552.html.

42. Ковальчук О. Я. Математичне моделювання та прогнозування в міжнародних відносинах (Ч. 1): навчальний посібник / О. Я. Ковальчук. – Тернопіль: ТНЕУ, 2016. – 423 с.

43. Бахрушин В. Є. Методи аналізу даних: навчальний посібник / В. Є. Бахрушин. – Запоріжжя: КПУ, 2011. – 268 с.

44. Василенко О. А. Математико-статистичні методи аналізу у прикладних дослідженнях: навчальний посібник / О. А. Василенко, І. А. Сенча. – Одеса: ОНАЗ ім. О. С. Попова, 2011. 166 с.

45. Мамчич Т. І. Статистичний аналіз даних з пакетом Statistica: навчально-методичний посібник / Т. І. Мамчич, А. Я. Оленко, М. М. Осипчук, В. Г. Шпортюк; Нац. ун-т «Києво-Могилянська академія». – Дрогобич: Вид. фірма «Відродження», 2006. – 207 с.

46. Чен К. MATLAB в математических исследованиях: Пер. с англ. / Чен К., Джиблин П., Ирвинг А. – М.: Мир, 2001. – 346 с.