МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

І. Г. Красовська, В. О. Ковальова

УПРАВЛІННЯ ЯКІСТЮ ДАНИХ ГІС

Навчальний посібник

Харків «ХАІ» 2019

УДК 911 : 004.9 – 021.4 (075.8)
К78

Рецензенти: д-р техн. наук, проф. Р. Е. Пащенко,
канд. техн. наук, доц. Б. М. Іващук

Красовська, І. Г.

К78 Управління якістю даних ГІС [Текст] : навч. посіб. / І. Г. Красовська, В. О. Ковальова. – Харків : Нац. аерокосм. ун-т ім. М. Є. Жуковського «Харків. авіац. ін-т», 2019. – 64 с.

ISBN 978-966-662-670-0

Розглянуто типи та джерела помилок, які притаманні ГІС-даним. Подано стратегію управління якістю даних ГІС. Описано методи кількісного оцінювання помилок.

Для студентів, що вивчають курс «Організація і управління геодезичними та земельно-кадастровими роботами» за спеціальністю «Геодезія і землеустрій».

Іл. 23. Табл. 11. Бібліогр.: 7 назв

УДК 911 : 004.9 – 021.4 (075.8)

ISBN 978-966-662-670-0

ВСТУП

Дані геоінформаційних систем (ГІС-дані) відрізняються від інших промислових товарів тим, що про їхню якість і придатність не завжди можна судити за їхнім виглядом. Зазвичай присутнє внутрішнє переконання, що щось, створене комп'ютером, має бути правильним. До того ж ГІС-інструментарій здатен демонструвати дані у дуже привабливому вигляді, що може приховувати їхні проблеми.

Термін якість, який використовується як властивість для опису об'єкта, має багато тлумачень, що узагальнено є адекватними відсутності дефектів чи проблем, або ж відповідно певним вимогам, тобто коли певна річ уважається гарною.

Щодо даних термін якість спирається на те, наскільки набір даних є ефективним для його передбачуваного призначення. Стало загальновизнаним, що кращі та інформаційно більш обґрунтовані рішення приймаються завдяки використанню ГІС. Доступ до просторових даних та аналітичних інструментів означає, що все більше рішень підтримуються ГІС-інструментарієм. Однак рішення, що приймаються на основі помилкових даних, імовірно, є менш точними, ніж рішення, прийняті навіть на основі неповних даних і низки припущень. Тобто рішення, що приймаються за відсутності інформації, все одно повинні спиратися принаймні на точно визначені припущення та обмеження. В англійській мові є вислів "сміття на вході – сміття на виході". Зміст цього в контексті ГІС дуже простий: якщо використовуються неточні або помилкові дані для аналізу, то результати аналізу будуть також неточними або помилковими. Навіть простіше: результати аналізу є настільки гарними, наскільки якісними є вхідні дані, використані для його проведення. Якщо необхідно мати певний рівень довіри до аналітичних результатів, то треба розуміти рівні помилок вхідних даних та усвідомлювати те, що відбувається з цими рівнями під час певних необхідних кроків з оброблення даних.

Дехто може розглядати якість даних з точки зору порівняльного аналізу витрат і вигід (результатів). Застосовуючи ГІС, зазвичай порівнюється вартість апаратного і програмного забезпечення, даних і навчання персоналу із заощадженнями, отриманими внаслідок підвищення ефективності та покращення рішень. Якщо ж бізнес є неефективним або рішення приймаються на основі недостатніх даних, то вигоди від застосування ГІС для відшкодування витрат не простежуються.

Підвищення якості даних ГІС зазвичай підвищує вартість збору даних. Однією з найбільш важливих задач для менеджера ГІС є визначення належного компромісу між вартістю збору даних та їхньою якістю. Структурно недостатні дані та помилкові їхні значення можуть зробити навіть прості завдання, такі, як розміщення даних, дуже неефективними, а відповідні неточності можуть призвести до підвищення витрат на бізнес. Наприклад, розгляньмо збір даних щодо місцезнаходження підземної інфраструктури, такої, як водопровідна,

каналізаційна чи інша комунальна інфраструктура. Вартість збору високоточних даних щодо такого місцезнаходження може бути високою і завжди буде залучати збір польових даних. Однак використання таких даних може забезпечити значні заощадження під час експлуатації зазначеної інфраструктури. Зокрема, відправка ремонтної бригади зробити невелику яму у правильному місці є значно дешевшою, ніж розкопування великої ями для пошуку труби або розкопування ями у хибному місці. Окрім того, це забезпечить менше втручання в інше майно, транспортний рух і прилеглі комунікації. Витрати на помилкові рішення коливаються від незначних (наприклад, надсилання повідомлення щодо тимчасового призупинення постачання води не тим домовласникам) до катастрофічних (відправка швидкої медичної допомоги за неправильною адресою, що може призвести до смертельних наслідків).

З витратами внаслідок низької якості даних поєднане й зростання ймовірності стикання зі складнощами, зумовленими помилками в цих даних. Зростання генерування даних у приватному секторі та збільшення спільного використання даних між різними організаціями та установами призводить до того, що у багатьох випадках дані, що використовуються певною організацією, не були нею достатньо підготовлені для спільного використання. Так, умови та методики збору таких даних, поданих для інших користувачів, можуть бути невідомими, що спричинює невпевненість щодо будь-якого аналізу на основі таких даних. Вимоги до точності даних залежать від їхнього застосування, а збільшення спільного використання даних означатиме підвищену ймовірність того, що дані будуть застосовуватися зовсім по-іншому, ніж це передбачалося первинним їхнім призначенням. Без чіткого розуміння помилок або невизначеності даного набору даних неможливо прийняти ефективне управлінське рішення щодо придатності цих даних для ваших цілей.

Оцінка якості даних є по суті комплексом вправ з управління ризиками. Ризик полягає в тому, що дані є непридатними для призначених цілей і наслідки можуть коливатися від додаткових витрат до неправильних відповідей на вельми визначальні запитання.

1 ТИПИ ПОМИЛОК ДАНИХ

Є кілька аспектів якості даних. Так, ми можемо розглядати наведені терміни або як різні критерії оцінювання помилок, або як різні типи помилок у наших даних:

- позиційна точність;
- тематична (атрибутивна) точність;
- часова точність;
- логічна узгодженість;
- повнота.

Багато із щойно зазначеного спирається на термін точність. Точність може вважатися відсутністю помилок або “правильністю” даних. Точні дані ГІС відображають об’єкти в базах даних у такий спосіб, щоб ці об'єкти були дуже подібними до того, чим вони є в реальності. Інформація у базах даних не може по-справжньому подавати реальний світ; завжди буде певний рівень абстракції або генералізації, коли ми намагаємося відобразити цей світ, використовуючи комп’ютерну модель. Через це ГІС-дані ніколи не можуть бути цілковито точними, утім вони можуть бути прийнятними для передбаченої мети або достатньо точними для певного аналізу або візуалізації.

Це означає, що ми можемо порівнювати величини, що зберігаються в наших базах даних, з реальністю, яку спостерігаємо. Утім це не завжди так. Наприклад, деякі явища, що передаються через карти, є суб’єктивними за їхнім тлумаченням, як, наприклад, поняття "оточення (сусідство)". Дехто може провести опитування людей, котрі мешкають на певній території, для з’ясування усвідомлення ними міри зазначеного оточення (сусідства), однак неможливо визначити цю міру лише на основі досліджень особистого досвіду. Історичні карти також описують попередні стани реальності, які неможливо перевірити кількісно.

Позиційна точність

При оперуванні просторовими даними найбільш очевидним різновидом помилок є помилки у подаванні простору. Ми можемо розглядати таку просторову або позиційну точність як близькість просторової інформації у базі даних, наприклад пари x та y координат, до їхнього дійсного місцезнаходження. Просторову помилку точки може бути виражено у помилці за однією з осей (наприклад, x, y, z) або як горизонтальну, вертикальну чи загальну помилку. На рисунку 1 показано приклад визначення позиційної (просторової) помилки точки за трьома осями.

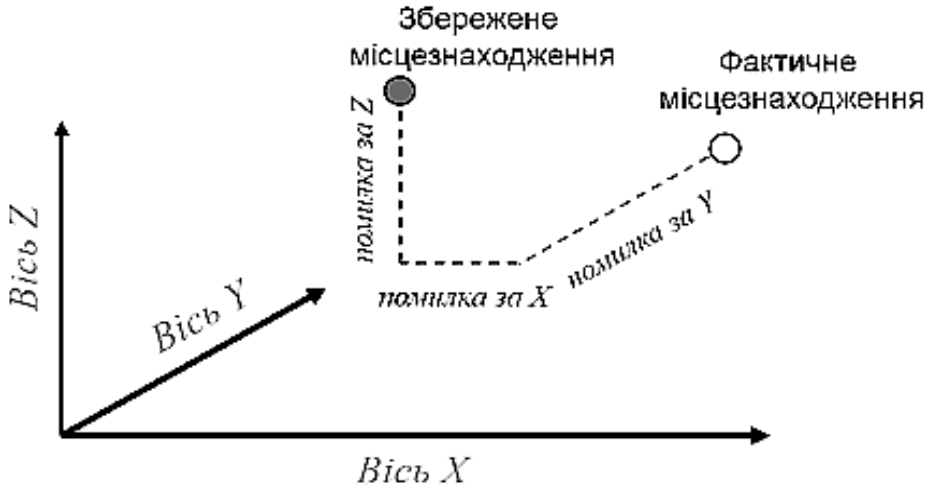


Рисунок 1 – Визначення позиційної (просторової) помилки точки

Помилки ліній та меж полігонів визначити кількісно складніше через таку просту причину, як те, що лінії в ГІС ніколи не можуть збігатися точно з їхньою формою, що існує в реальності. Наприклад, коли водотік створюється як лінія в базах даних ГІС, то він узагальнюється або генералізується, щоб відповідати певному масштабу введення даних у цю базу (зокрема шляхом цифрування). При цьому деякі точки на лінії, такі, як злиття двох водотоків, мають звірятися з реальним їхнім місцезнаходженням, однак більшість точок, які визначають форму водотоку, не відповідають їхньому конкретному місцезнаходженню, виміряному на земній поверхні. На рисунку 2 показано приклад помилок при цифруванні ліній.

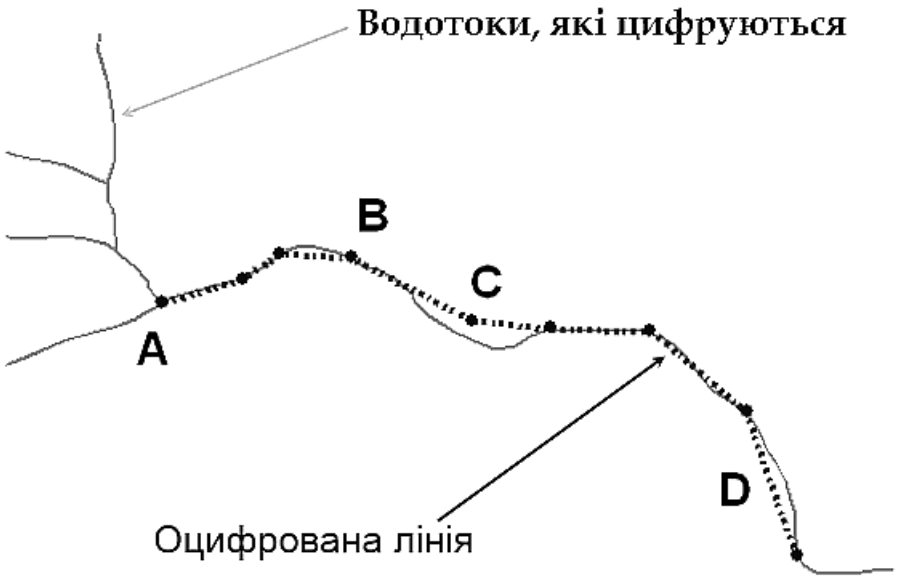


Рисунок 2 – Помилки ліній при їхньому цифруванні

Точка в положенні А відповідає точному її місцезнаходженню (злиттю двох невеликих водотоків), яке може бути виявлено в реальному просторі. Ми можемо оцінити точність цієї точки так само, як і будь-яких інших точкових даних (див. рисунок 1).

Точка В здається точною просто тому, що вона лежить безпосередньо на лінії, яку ми намагаємося відобразити. Однак якщо посунути цю точку значно вище або нижче за течією (залишаючи її лежати на лінії водотоку), чи буде вона все ще точною? Відповідь полягає в тому, як впливає пересування точки В на всю лінію. Цей різновид точки (наприклад, В, С) часто називають точкою форми (або формотвірною точкою), оскільки вона не відмічає якесь конкретне визначене положення, а лише надає форми лінійному просторовому об’єкту. Ця точка як формотвірна також може називатися вершиною (вертексом) (на відміну від вузла векторної структури даних, який править за кінцеву чи початкову точку лінійних дуг цієї структури). Точки форми є певною мірою довільними і технік із введення даних просто додає точки, щоб лінія слідувала за

напрямком призначеного для цифрування просторового об’єкта з детальністю, достатньою для вимогового масштабу. Це робить надзвичайно складним кількісне визначення величини помилки для таких точок. Так само з певним рівнем імовірності ми можемо стверджувати, що точка С має помилку, оскільки ця точка не лежить на просторовому об’єкті, який ми намагаємося відобразити. Утім, знову ж таки, величина цієї помилки не є очевидною. Ми можемо стверджувати, що зазначена помилка є відстанню від точки С до найближчої точки на водотоці. Хоча, як і у випадку з точкою В, можуть бути точки вище або нижче за течією, що зроблять загальну форму результувальної лінії більш наближеною до справжнього просторового об’єкта, ніж точка на водотоці, найближча до точки С. Позначка D на рисунку 2 маркує відрізок нової лінії, що істотно відхиляється від справжнього водотоку як просторового об’єкта. Це спричинено відсутністю точки форми, а не її неправильним місцезнаходженням, що зумовлює відхилення від дійсного місцезнаходження водотоку. Розташування точок форми, як і їхня частота, визначає, наскільки добре лінійний об’єкт у ГІС відображає лінійний об’єкт в реальності. Створення або вилучення формотвірних точок є частиною генералізації, яка є притаманною процесу подавання об’єктів реального світу в комп’ютерних базах даних.

Ще одним зі способів характеристики помилки лінійних просторових об’єктів є задавання буфера навколо оцифрованої лінії (або межі полігону), ширина якого відповідає величині помилки в даних. Цей буфер відображає територію, усередині якої знаходиться дійсна лінія з певним ступенем вірогідності. Однак буфер не може бути однаковим за всіх умов, оскільки, наприклад, по-перше, прямі лінії точно цифрувати простіше, ніж складні криві, а по-друге, лінії з багатьма чітко визначеними місцезнаходженнями (зокрема перехрестя вулиць) можуть бути більш точно оцифровані, ніж окремі ізольовані у просторі лінії. На рисунку 3 показано приклад такого буфера помилки.

Рисунок 3 – Помилки ліній відображені за допомогою буфера

Ширина буфера помилки є функцією вірогідності даних. Чим ширший буфер, тим більша вірогідність, що дійсне місцезнаходження просторового об’єкта розміщене десь у його межах. На рисунку 4 відображено ідею зростання зазначеної вірогідності із зростанням ширини буфера. На цьому рисунку, наприклад, буфер із 70 % вірогідності міститиме територію, оточену і буфером 50 %, і буфером 70 % вірогідності. Наземна одиниця ширини буфера заданої вірогідності, такої як 90 %, буде функцією масштабу даних (наприклад, чим крупніший масштаб, тим більш вузьким буде буфер і тим більше ми можемо бути упевненими щодо позиційної точності об'єктів).

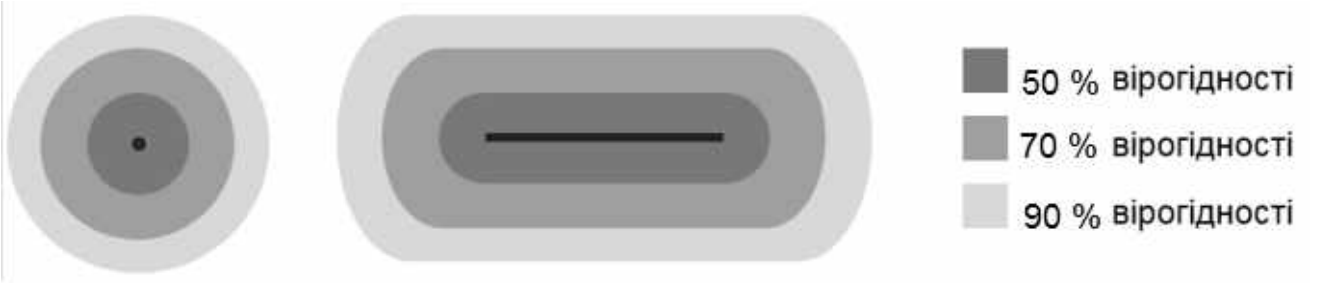


Рисунок 4 – Позиційна помилка та межі вірогідності точності точкових і лінійних просторових об'єктів

Масштаб є відношенням відстані або розмірів на карті (наприклад, між двома просторовими об’єктами, або як розмір одиничного просторового об’єкту) до відповідної відстані на земній поверхні. Він може бути поданий як чисельний дріб (масштаб), такий, як 1:50 000. Це означає, що одна одиниця на карті відображає 50 000 таких одиниць на земній поверхні. 1:5 000 є прикладом крупного масштабу карти, який відображає невелику ділянку з істотною детальністю. 1:2 000 000 є дрібним масштабом карти та подає велику територію з незначною детальністю. ГІС значною мірою не залежить від масштабу, оскільки за допомогою ГІС-інструментарію ми можемо генерувати на екрані монітора або на папері карти будь-якого масштабу. Однак є обмеження щодо масштабу, застосовного для практики. Проблеми виникають у випадках використання ГІС-даних для аналізу або відображення у масштабах, значно відмінних від масштабу вихідного джерела даних для ГІС (геодезичної чи аерофотозйомки, паперової карти тощо), який ще називають масштабом уведення даних. Масштаб уведення даних буде визначати:

- Належні рівні точності, чіткості та генералізації даних; крупніші масштаби зазвичай потребують вищих рівнів точності і чіткості та нижчого рівня генералізації.
- Які просторові об’єкти може бути відображено; невеликі об’єкти, такі як обриси будинку, не може бути подано на карті дрібного масштабу, що зображує державу в цілому.

- Як просторові об’єкти будуть подаватися; у масштабі 1:1 000 000 місто може бути точкою, тоді як в масштабі 1:20 000 воно буде великим полігоном.

Наприклад, карту масштабу 1:250 000 не може бути ефективно використано для аналізу або відображення в масштабі 1:10 000 із зазначених вище причин. Міста, які подано на карті в масштабі 1:250 000 як точки, не стають полігоном просто завдяки тому, що Ви маєте спроможність збільшити масштаб до 1:10 000. Багато просторових об’єктів (таких, як невелике озеро) взагалі не будуть відображені на карті масштабу 1:250 000 і питання про їхню позиційну точність взагалі не виникне.

Чіткість даних може тлумачитися і як точність їхнього вимірювання або розрізнювання (розрізнювальна здатність) даних. З функціональної точки зору це розмір найменшого просторового об’єкта, який може бути відображено у базах даних. Розрізнювання щільно поєднано з масштабом даних, оскільки розмір найменшого об’єкта, що може бути відображено, буде коливатися залежно від масштабу відображення або масштабу введення даних. Наприклад, якщо ми припустимо, що найменший розмір об’єкта, який обґрунтовано може бути розміщений на паперову карту, становить 0,5 мм, то об’єкт такого розміру буде 25 м у масштабі 1:50 000 і 500 м у масштабі 1:1 000 000. Карта крупнішого масштабу дозволяє відстежувати просторові об’єкти краще, а отже загалом потребує вищого розрізнювання. До речі, у растровій структурі (базі) даних розрізнювання – це просто розмір піксела (прямокутної комірки растра). Хоча чіткість як розрізнювання даних зазвичай використовується для опису їхнього просторового розрізнювання (чіткості), як було описано вище, вона також може бути застосована і до атрибутивного розрізнювання (атрибутивної чіткості). Атрибутивне розрізнювання (чіткість атрибутів) стосується того, наскільки чітко описано просторовий об’єкт, або наскільки ретельно виконано його категорування. Вимірювання довжини межі земельної ділянки до трьох знаків після коми (наприклад, міліметрів) або зберігання детальної типізації землекористування, такої, як “односімейна житлова забудова” (замість просто “міська забудова”) є прикладами атрибутів високої роздізнювальної здатності. Просте збільшення розрізнювання даних не ґарантує відповідне покращення їхньої точності. Можна мати дані крупного масштабу із високим розрізнюванням але з поганою точністю. Супутниковий знімок з дуже малим розміром піксела, який однак був дуже погано просторово прив’язаний, може правити за приклад такої ситуації. Ми регулярно зберігаємо векторні дані, використовуючи координати високої точності (наприклад, до метра), у той час, коли їхнє місцезнаходження може бути точним лише в межах 25 метрів. У такому випадку точність джерела даних є нижчою, ніж точність даної бази даних. Однак при зменшенні розрізнювання даних має відбуватися і відповідне зменшення їхньої точності. Наприклад, растр з пікселами у 10 метрів може

бути точним лише в межах 10 метрів.

Тематична (атрибутивна) точність

Тематична точність (яка також називається атрибутивною точністю) є мірою близькості значень, які зберігаються у атрибутивних базах даних, до дійсних значень. Атрибутивні значення можуть бути просто хибними, такими, як неправильна назва вулиці як певного просторового об’єкта. Такі різновиди помилок відносно легко виявити та виправити. Інші атрибутивні помилки може бути значно складніше знайти. Наприклад, значення абсолютної висоти, яке є просто неправильним, виявити дуже складно. Подібна ситуація стосується і необ'єктивних класифікацій, в яких полігон закодовано як “молодий ліс” замість “старий ліс”. Тематична точність не залежить від масштабу на відміну від просторової (позиційної) точності.

Часова точність

Там, де бази даних мають часовий складник, треба враховувати часову точність даних. Наприклад, якщо бази даних охоплюють період часу, коли існувала певна будівля, то дати її зведення та знесення мають бути розглянуті на предмет їхньої точності. Часова точність не є тим самим, що й сучасність даних. Часові дані можуть бути точними у часі, але не сучасними. Наприклад, полігон, який відображає положення лісу, може мати атрибут, який інформує, що деревину на цьому полігоні було цілковито вирубано у 2003 році. Якщо атрибути не сучасні, ми можемо не знати, що на зазначеному полігоні було знову висаджено ліс у 2005 році і наразі це є молодим деревостаном. Отже, з огляду на те, що ліс на полігоні дійсно було вирубано у 2003 році, дані щодо полігону мають часову точність, однак не є сучасними, оскільки відновлення лісу ще не зафіксовано. Часова точність загалом є важливою для історичних прикладних досліджень.

Логічна узгодженість

Логічна узгодженість стосується можливої наявності протиріч у даних. Помилки в логічній узгодженості можуть виявлятися у двох формах.

<u>Просторова узгодженість</u>

Стосується способу взаємодії просторових даних і тому може бути охарактеризована як помилки топологічної узгодженості. Прості приклади можуть містити наявність топологічних особливостей, таких, як помилкові висячі вузли, причинами виникнення яких є незамкненість межі полігона або ситуація з дугою, яку називають "недоліт". Більшість прикладних

програм геоінформаційних систем мають інструменти для знаходження та виправлення таких помилок, однак є низка неузгодженостей, які можуть виникнути в просторових базах даних і не так очевидно порушують елементарні правила просторової узгодженості.

Наприклад, розгляньмо таку ситуацію. Сейсморозвідувальні профілі є дуже довгими, і при перетині ними лісу в ньому розчищено вузькі прогалини з метою надати можливість нафтовому та газову розвідувальному обладнанню взяти зразки на регулярних інтервалах уздовж сітки розвідки. Ці профілі можуть бути довжиною у кілька кілометрів і перетинати уздовж і впоперек території, потенційні на нафтові поклади. Одночасно існують суцільні вирубки, які є досить великими ділянками лісу, розчищеними з метою збору деревини.

Запитання з точки зору баз даних геоінформаційних систем полягає у тому, чи сейсморозвідувальний профіль, що проходить через суцільну вирубку, дійсно знаходиться там? Хтось може заперечити це, мотивуючи тим, що розчищені вузькі лісові прогалини не можуть бути в межах великої суцільної вирубки і що сейсморозвідувальні профілі повинні перериватися, коли вони "підходять" до полігону суцільної вирубки і знову відновлюватися на іншому боці полігону, коли знову починається заліснена ділянка. Власне помилки в логічній узгодженості можуть виникати, якщо, наприклад, різні техніки в межах однієї організації виконували цифрування згаданих вище просторових об’єктів по-різному. Так, якщо один технік виходив з того, що сейсморозвідувальні профілі мають бути перервані та видалені там, де вони перетинають суцільні вирубки, а інший оператор просто безперервно цифрував сейсморозвідувальні профілі, що проходять через ділянки суцільних вирубок, то у базах даних будуть присутні відповідні неузгодженості. Наприклад, якщо будь-яка аналітична модель буде обчислювати площу ділянок вирубок, то вона може порахувати двічі площі проходження сейсморозвідувальних профілів через ділянки суцільних вирубок.

Інший приклад просторової узгодженості стосується топології вуличної мережі. Неузгодженість може існувати, якщо у деяких частинах мережі з'єднання вулиць розбите звичайним з’єднувальним вузлом (точкою), розміщеним на перехресті, тоді як інші частини мережі доріг як просторових об'єктів просто перетинаються в просторі без наявності нових вузлів перетину. Зверніть увагу на те, що ця остання ситуація створює проблему для мережного аналізу, оскільки програмне забезпечення не вважає перетином ліній випадок, якщо вони не з’єднані спільним вузлом перетину. Лінії ж, що перетинаються у просторі без з’єднувального вузла, розглядаються для відображення таких складних об'єктів, як, наприклад, естакада, коли автотранспорт не може повернути з однієї вулиці на іншу у їхньому перетині.

Непросторова узгодженість

ЇЇ помилки відображають неузгодженості атрибутів у ГІС. Такі помилки можуть бути помилками у методиці, коли, наприклад, ми порівнюємо густоту населення двох держав, але дані щодо населення для кожної держави було отримано в різні роки. Помилки також можуть бути наслідком протиріч між уже накопиченими атрибутивними значеннями, наприклад, якщо в таблиці збережено дані щодо площі певної території і кількості та густоти населення на ній, однак ці дані не є узгодженими (тобто густота населення не дорівнює частці від ділення його кількості на площу тощо).

Повнота

Помилки повноти даних стосуються речей, які пропущено в базах даних. Ці помилки можуть бути спричиненими відсутністю цілих класів просторових об’єктів, необхідних для певних потреб, однак більш імовірно стосуватимуться окремих просторових об’єктів, якими знехтували у базах даних.

2 ДЖЕРЕЛА ПОМИЛОК ДАНИХ

Обговорення, наведені вище, подають узагальнення щодо типів помилок, які притаманні ГІС-даним. Ці помилки можуть виникати внаслідок низки різних причин, і розуміння джерел таких помилок є основою для скорочення їхньої кількості та управління помилками у Ваших даних. Нижче подано узагальнення того, яким чином помилки виникають у наборах даних.

2.1 Очевидні помилки

Строк дії:

Існує низка очевидних підстав уважати, що дані є неточними. За першу з них править те, що дані є просто застарілими. Потреба в оновленні даних безсумнівно буде залежати від суті об'єктів і явищ, які картографуються. Наприклад, дані щодо ґрунтів і рельєфу змінюються досить повільно і ми можемо використовувати цю інформацію без оновлення значно довше, ніж такі більш динамічні набори даних, як дані щодо землекористування або транспортної мережі. Застарілі дані означатимуть, що просторові об’єкти фізично змінилися, у тому числі виникли або зникли з моменту останнього оновлення наших даних. Однак, окрім того, Барроу і МакДоннелл (1998) запропонували, що ми маємо враховувати ймовірність того, що сутність усього набору даних могла

змінитися. Так, дані можуть бути щойно зібрані з огляду на технічні умови, що змінилися (наприклад, щодо мінімального розміру полігону або принципів їхньої класифікації) або на стандарти, що теж змінилися.

Часто постачальники даних (наприклад, орган центральної влади, відповідальний за створення та ведення спеціальних наборів даних, таких, що стосуються землекористування або планіметрії) у процесі оновлення набору даних можуть змінювати спосіб збору або постачання даних замовнику. Це може бути зроблено завдяки залученню кращих технологій оброблення даних або зміни вимог до даних від їхніх споживачів. У таких випадках корисність і придатність нових, оновлених даних можуть бути геть відмінними від старої версії тих самих даних.

Просторове охоплення

Іншим джерелом очевидних помилок є просто просторове охоплення даних. Якщо необхідні дані не поширюються на весь географічний простір, де вони нам потрібні, то будь-який аналіз обов'язково буде неповним. Зазвичай дослідження, що стосуються великих територій, можуть бути частково підкріплені дуже високоякісним, крупномасштабним картографічним матеріалом. Однак на деяких територіях доступними можуть бути лише дані у дрібнішому масштабі і значно меншою точністю. У такій ситуації аналітики повинні вирішувати, чи генералізувати дані крупного масштабу (чи просто відкинути їх і використати дані дрібнішого масштабу), щоб можна було оперувати цілісним набором даних, чи зібрати додаткові дані у необхідному масштабі, щоб створити набір даних крупнішого масштабу на всю територію дослідження. Не рекомендується проводити дослідження з даними змішаних масштабів, оскільки будь-який аналітичний результат буде неузгодженим і ефективно зробити виконання порівнянь за територією дослідження буде неможливо.

Масштаб та генералізація

Цей останній приклад приводить нас до останньої очевидної проблеми з даними, а саме проблеми масштабу. Ми вже обговорювали, що дані крупного масштабу зазвичай мають вище просторове розрізнювання та більш детальні атрибути. Через це важливо узгоджувати масштаб даних з вимогами до їхнього аналізу чи візуалізації. Використання даних занадто дрібного масштабу для певних потреб призводить до неточності позиційних даних і, можливо, їхніх атрибутів, що не забезпечує необхідну функціональність даних. Використання даних занадто крупного масштабу призводить до перевантаження роботи обчислювальної системи, спричинене наявністю надто детальних позиційних і атрибутивних даних, які можуть ніколи реально не використовуватися. Утім останні міркування викликають все менше

занепокоєння, позаяк потужність сучасних комп'ютерних робочих станцій та можливості зберігання і пошуку даних у сучасних їхніх базах призводять до того, що, наприклад, наявність додаткових детальних просторових даних не знижують відчутно продуктивність обчислювальної системи. Наприклад, рисунок 5 ілюструє подану берегову лінію прибережних просторових об’єктів, початково отриману в масштабі 1:250 000 і відтворену в масштабі 1:10 000. Ця берегова лінія не є придатною для аналізу або візуалізації в масштабі 1:10 000, оскільки у цьому масштабі вона має недостатнє просторове розрізнювання, спричинюючи блоковий характер її обрисів та очевидні позиційні розбіжності до 40 метрів.

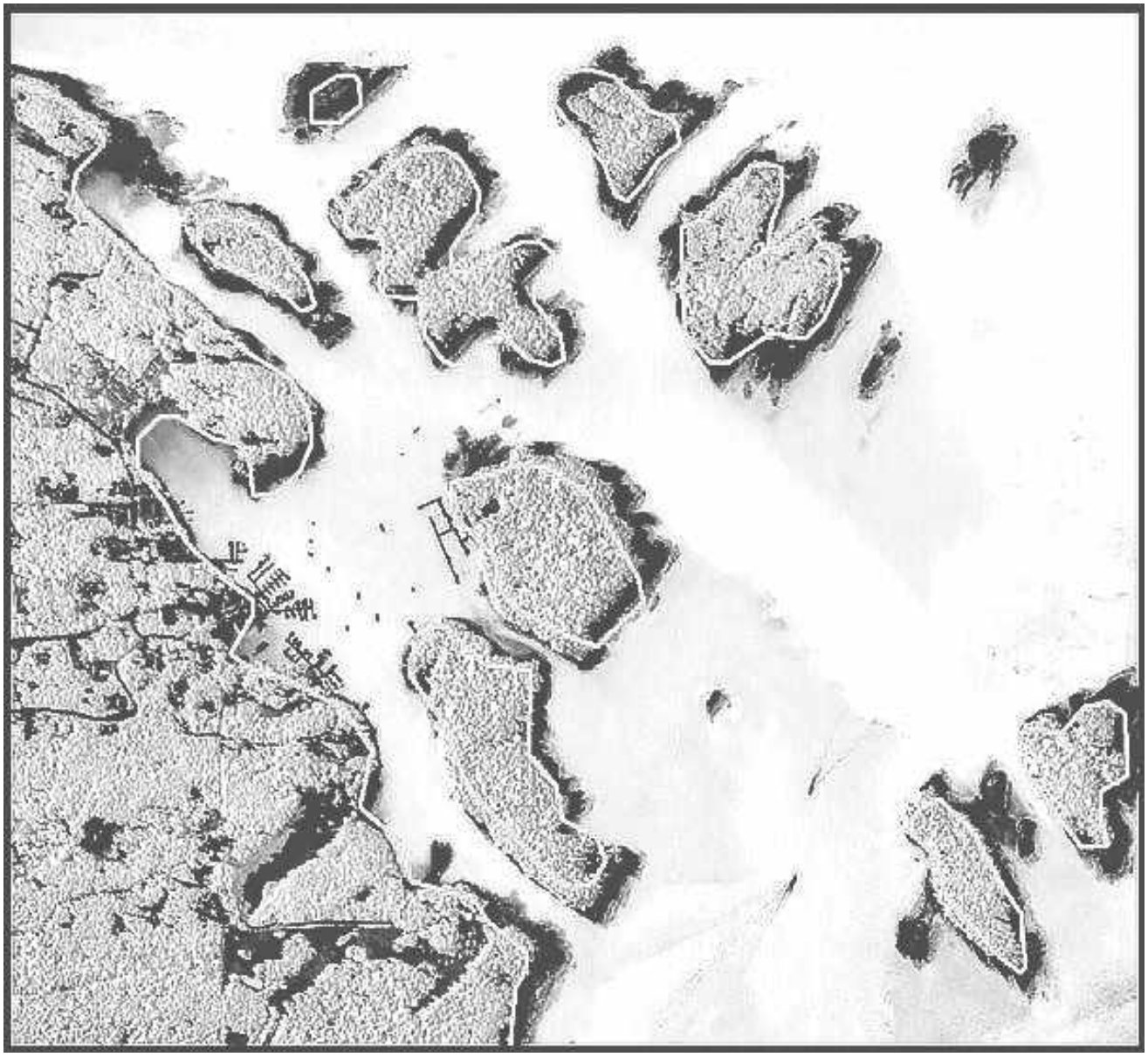


Рисунок 5 – Масштаб і генералізація просторових даних

Раніше відзначалося, що найменший об’єкт, який людина може дійсно розрізняти на паперовій карті, має розмір 0,5 мм. Це зазвичай також

стосується і відстеження ліній – можна очікувати, що оператор під час цифрування може відхилитися від лінії, яка цифрується, на таку ж відстань. Зазначені 0,5 мм є гарним емпіричним рішенням для величини очікуваної помилки у просторових даних винятково через обмеження, спричинені розрізнюванням і генералізацією під час уведення даних (Лонґлі, 2011 рік). У таблиці 1 узагальнено величину цієї помилки для багатьох широко розповсюджених масштабів. Це лише приблизна оцінка величини дійсної помилки, однак її просто обчислити оскільки вона надає користувачам даних ГІС простий спосіб розуміння імовірності помилок у їхніх даних, для яких відсутні детальні метадані.

Таблиця 1 – Очікувана помилка у просторових даних залежно від масштабу

Масштаб	Очікувана помилка
1:5 000	± 2,5 м
1:10 000	± 5,0 м
1:20 000	± 10 м
1:50 000	± 25 м
1:250 000	± 125 м
1:1 000 000	± 500 м

2.2 Помилки вимірювань

Дані в базах даних ГІС створено за допомогою певного виду вимірювань. Значення атрибутів, таких, як абсолютна висота, кислотність ґрунтів і температура води, вимірюються в польових умовах з використанням спеціального обладнання. Просторові дані створюються шляхом зняття координат з паперових карт або вимірювань на земній поверхні, використовуючи геодезичну зйомку або обладнання GPS. Усі ці способи вимірювань і спричинюють ті шляхи, якими помилки можуть "потрапити" до баз даних.

Інструментальна помилка

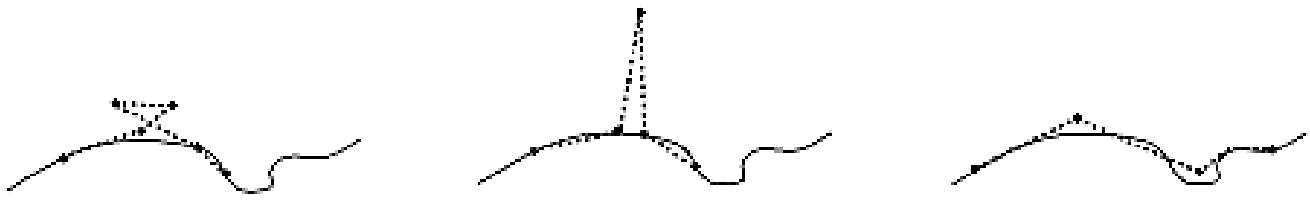
Обладнання для виконання вимірювань може бути джерелом помилки. Прилад, який використовується для вимірювання глибини океану та його солоності і є некоректно відкаліброваним, не може забезпечити надійність атрибутивних значень. При цифруванні паперових карт помилка може бути введена в бази даних через низьке розрізнювання графічного планшета (дигітайзера). Високоякісні дигітайзери можуть мати точність до 0,075 мм, тоді як дигітайзери більш низької якості можуть мати точність приблизно 0,25 мм (Крісман, 2001). Хоча щойно зазначена величина помилки є невеликою відносно інших джерел помилок, утім у такому дрібному масштабі, як 1:250 000, помилка у 0,075 мм може призвести до

позиційної помилки у приблизно 20 метрів. При цифруванні паперових карт додаткова помилка може бути спричинена станом самого джерела вихідної інформації (тобто власне карти). Паперові карти можуть бути занадто пересушеними або розтягнутими внаслідок користування ними та змін вологості. Копіювання, таке як фотокопіювання додатково спотворює карти. Якщо є можливість і якщо точність є принципово необхідною, замість паперових слід застосовувати більш стійкі карти, такі, як ті, що створено на пластиковій основі.

Помилка оператора

Однак названі вище джерела помилок є незначними порівняно з помилками, які вносить оператор. Уводячи в комп'ютер атрибути, оператори можуть неправильно інтерпретувати просторовий об’єкт і ввести неправильне значення, наприклад увести атрибут лісу “широколистяний”, коли в дійсності він є хвойним. Також можуть бути грубі помилки або топографічні помилки, що призводять до неправильних значень даних. Малі помилки при введенні адреси вулиці можуть призвести до переміщення кількох кварталів при виконанні географічної прив'язки даних, а топографічна помилка при введенні висоти лісового покриву може викликати помилку у декілька порядків від дійсної величини.

При виконанні цифрування помилки оператора містять надмірну для заданого масштабу даних генералізацію просторових об’єктів і помилки внаслідок неуважності або мимовільних м’язових рухів. На рисунку 6 показано приклади деяких проблем цифрування.



"Плутанина формотвірних точок" "Зубці" "Переліт" і "недоліт"

Рисунок 6 – Приклади помилок цифрування

Помилка складання

Навіть якщо оператори зможуть ідеально перевести паперову карту у цифрову версію, помилки в бази даних "потрапляють" через те, що вихідний (первинний) документ (наприклад, карта) сама по собі має помилки. Помилка у вихідному документі називається помилкою складання (карти тощо). Вибір картографічної проекції і сфероїда або системи координат буде видозмінювати характер і розподіл помилки на

паперовій карті. Помилки генералізації, які ми розглянули вище, також присутні на паперових картах, оскільки автор карти мав абстрагуватися від реального світу для розміщення інформації на карті. Також є додаткові проблеми з паперовими картами, якщо “картографічні угоди” застосувалися для покращення читабельності карти. Наприклад, там, де залізниця проходить паралельно автомобільному шляху або поруч з ним, картограф може "вручну" пересунути один або обидва із зазначених просторових об’єктів таким чином, щоб користувач карти міг чітко розрізнити два різновиди відповідних ліній, навіть якщо в дрібному масштабі ці дві лінії мали б фактично збігатися.

Помилки оброблення

Топологічна "чистка"

Оцифровані лінії та полігони часто мають невеликі помилки, такі, як різні прогалини та "перельоти", які перешкоджають формуванню належних полігонів або мереж ліній. Такі топологічні помилки можна виправити автоматичною обробкою, яка усуває подвійні лінії, закриває прогалини в лініях і видаляє осколкові полігони без ручного втручання. Для виконання такого оброблення програмне забезпечення ГІС оснащене значенням так званої толерантності, яка визначає величину відповідних змін, які мають бути здійснені. Наприклад, помилковий висячий вузол (що виник через ситуацію з дугою "недоліт" або "переліт"), який менший за значення толерантності, може бути усунутий автоматично, або прогалини в межі полігону, менші за значення толерантності, можуть бути автоматично закриті. На рисунку 7 показано два приклади такого топологічного оброблення.

До топологічного оброблення Після топологічного оброблення

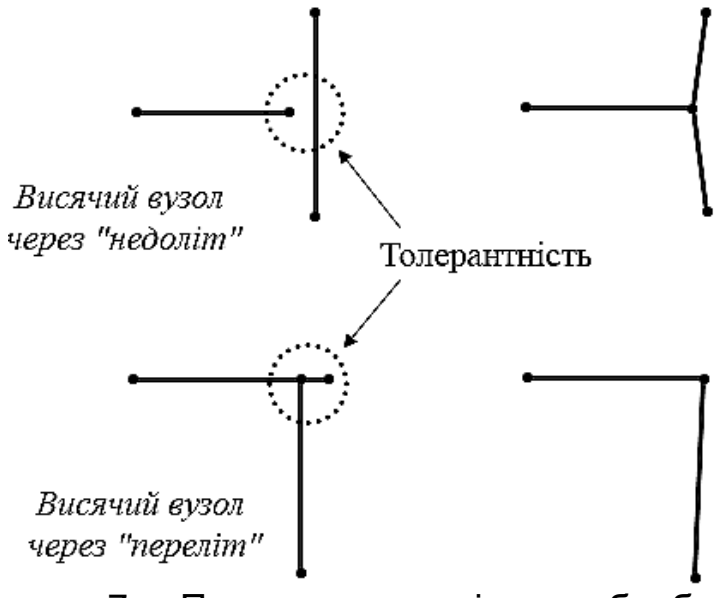


Рисунок 7 – Приклад топологічного оброблення

Вибір значення толерантності є вирішальним для якості оброблених даних. Занадто мале значення толерантності призведе до того, що багато топологічних помилок не буде автоматично виправлено і знадобиться значний обсяг ручного втручання оператора. Однак занадто велике значення толерантності призведе до значних проблем з даними. Ці проблеми можуть містити втрату малих полігонів, видалення прогалин і висячих вузлів, які були передбачені для створення, та надмірну генералізацію форми лінії чи межі (наприклад, видалення занадто великої кількості формотвірних точок).

Розглянемо такий приклад. Нехай останнім часом дані постачалися для ГІС-проекту урядовим підрозділом, який відповідає за створення та обслуговування базових карт. Дані, які постачаються, відповідають стандарту 1:20 000 масштабу для більшості ГІС-робіт у обраному регіоні. Вони містять такі просторові об’єкти, як гідрографічні та транспортні об’єкти, а також дані щодо рельєфу. Ці дані підлягають скрупульозному процесу контролю якості і повинні відповідати чинним деталізованим стандартам точності. Однак, судячи з усього цього, над наданими даними було виконано топологічне оброблення із занадто великим значенням толерантності, найбільш імовірно після того, як цей набір даних пройшов усі перевірки якості. Результатом застосування великого значення толерантності і став набір даних, який без сумніву є непридатним для аналітичних цілей у призначеному масштабі 1:20 000. На рисунку 8 показано передбачені для отримання дані та дані, які були надані.

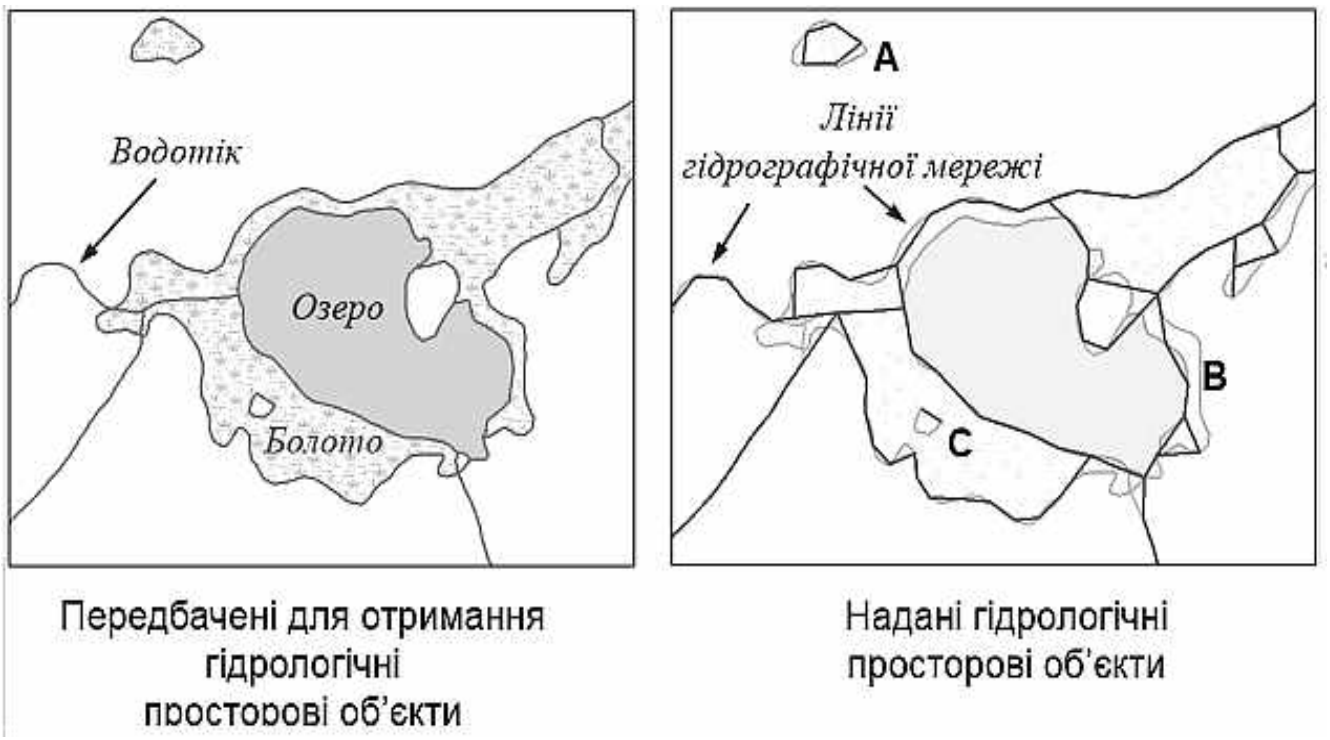


Рисунок 8 – Помилки внаслідок топологічного оброблення із застосуванням великого значення толерантності

Рисунок ліворуч (передбачені для отримання гідрологічні просторові об’єкти) показує очікуваний замовниками набір даних, які відображають озера, болота та водотоки. Однак дані, які були надані, значно відрізнялися від очікуваних. Вигляд на рисунку 8 праворуч показує лінії поданої гідрографічної мережі чорним кольором. Будь-хто може одразу помітити значні відмінності у даних унаслідок видалення значної кількості вершин (вертексів). Невелика заболочена ділянка у точці А є прикладом помилки, що виникає в результаті втрати точок форми. Оскільки зазначена ділянка як просторовий об’єкт є значущим, то він також залишається і топологічно релевантним для відображення, і тому він все ще подається замкненим полігоном. Також передбачалося, що вздовж усього периметра до озера у частині В прилягатиме значна заболочена ділянка. У цьому випадку надану топологію було значно видозмінено, і тепер вона відображає, що вздовж східного краю озера практично немає заболоченої ділянки, тому що вузький полігон цієї ділянки було повністю “стиснуто” до лінії. Крім того, невелика вільна ділянка посеред заболоченої території, позначена С, мала відображати піднесену незаболочену територію, або невеликий острів посеред болота. Топологічна ж обробка перетворила цей невеликий острівний полігон у коротку лінію. Знову ж таки, це є істотною помилкою, оскільки зазначений острівний полігон (як прогалина у більшому заболоченому полігоні) перетворився на лінійний об’єкт, що спричинить топологічні проблеми в майбутньому. Цей приклад засвідчує очевидну грубу помилку з боку того, хто обробляв ці дані перед їхнім випуском, він також ефективно демонструє небезпеки некоректної топологічної обробки. Щоб знизити втрату даних у ході такої обробки, слід обирати помірну величину толерантності. Он-лайн довідка програмного забезпечення ArcGIS радить застосування величини толерантності в розмірі 10 % від розрахункової точності даних. Наприклад, якщо набір даних мав точність +/- 25 метрів, то доречним значенням толерантності буде 2,5 метра, що забезпечить уникнення непередбачуваних змін у даних.

Перетворення векторних даних в растрові

При перетворенні векторних даних у растрову структуру також будуть виникати додаткові помилки. Просто через сутність растрової структури даних вона не може пристосуватися до точності, наявної у векторній структурі, хіба що розмір комірки растра буде нереально малим. Кодування дискретної точки як растрової комірки означає, що ми втрачаємо точне положення точки і знаємо лише те, що точка знаходиться десь у межах комірки. Подібно до цього окрема межа об'єкта, подана плавною векторною лінією, перетвориться на “ступінчасту” у растровій структурі і плавність форми лінії буде втрачено. Це є очікуваним і невід’ємним перетворенням у растровій структурі.

Класифікація даних

Подібно до зазначеного вище класифікація та перекласифікація даних може привносити втрату їхньої точності, або навіть помилку в них. Такий звичайний і простий процес, як класифікація числових даних для відображення їх на фоновій картограмі, знижує точність даних. Окрім того, спосіб, в який дані класифіковано, може змінити результувальні просторові візерунки зазначеної картограми і можливо сприймання даних для певної частини користувачів картограмою. На рисунку 9 показано приклад класифікації цифрових даних щодо стану територій для оселищ чорного ведмедя у два різні способи.

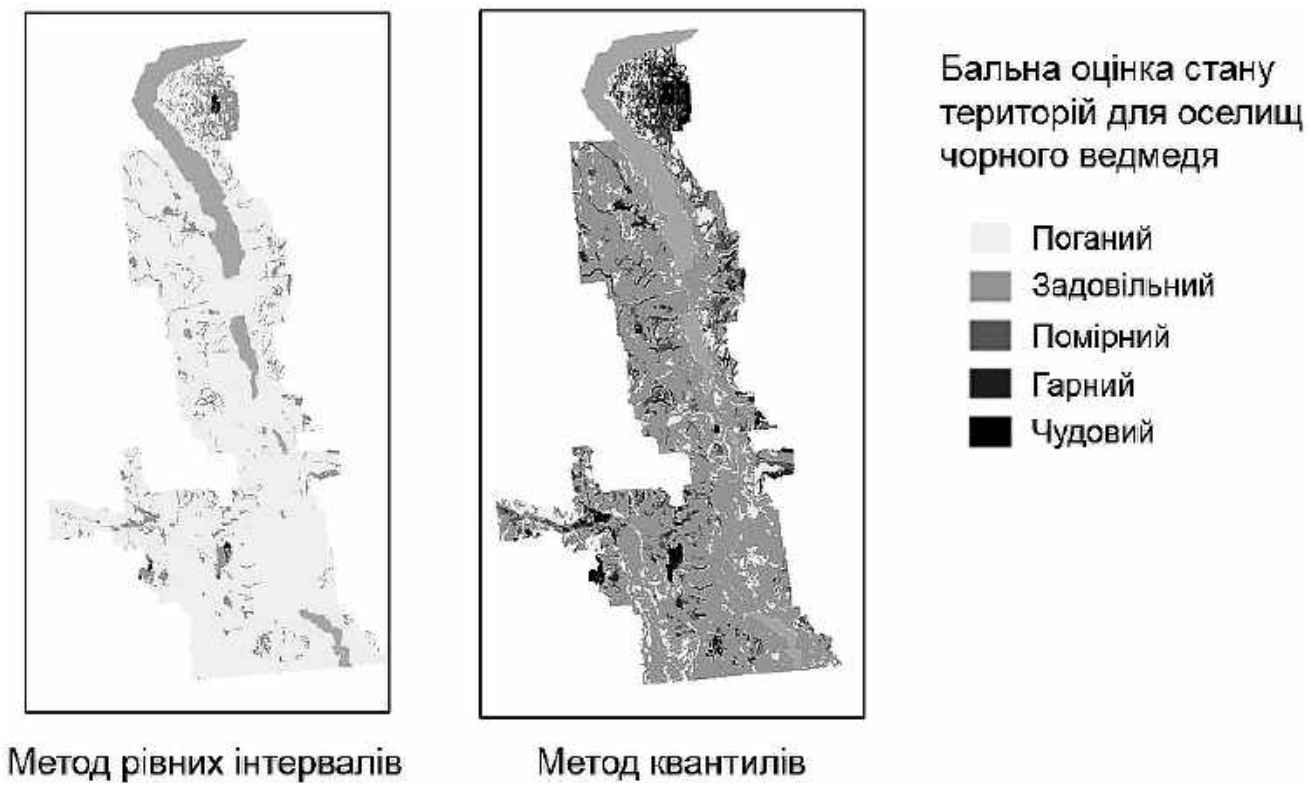


Рисунок 9 – Вплив методу класифікації на сприймання даних

Щоб виконати бажані картограми, числовий бал від 0 до 100, розрахований на основі параметрів землекористування та рослинних угруповань, було класифіковано на серію номінальних категорій, що відображають стан територій для оселищ чорного ведмеді (поганий, задовільний тощо). Класифікаційний метод рівних інтервалів передбачає однаковий розмір інтервалу кожного класу, тоді як метод квантилів забезпечує рівну кількість територіальних полігонів у кожному класі. Результати є помітно різними, а саме рівноінтервальна картограма свідчить про те, що стан усієї території є в цілому поганим для оселищ ведмедів, тоді як квантильна картограма показує, що в цілому цей стан є досить гарним. У доповнення до зміни сприймання даних класифікація

може змістити точність даних через застосування номінальної, а не, наприклад, інтервальної шкали категорій. Такі самі проблеми можуть виникнути під час класифікації растрових знімків дистанційного зондування.

Накладання полігонів

Операція накладання полігонів є іншим джерелом помилок у просторових даних. Коли два набори полігонів накладаються, дуже часто утворюються малі осколкові полігони (або фальшиві полігони). Це, зокрема, трапляється, коли межу полігона була оцифровано в обох полігональних наборах даних, однак через обмеження процесу цифрування межа не була оцифрована у однаковий за точністю такий же спосіб у кожному з наборів. Очевидно, що зі збільшенням масштабу (тобто у крупних масштабах) *площа* результувальних осколкових полігонів знижується, однак залишається висока ймовірність того, що *кількість* осколкових полігонів при цьому буде насправді зростати, оскільки дані крупнішого масштабу зазвичай мають збільшену кількість формотвірних точок, якими власне і задають криві лінії.

3 УПРАВЛІННЯ ЯКІСТЮ ДАНИХ

У даному розділі розглянуто характеристики та джерела помилок. З огляду на те що невизначеність завжди буде присутня у наших даних, цю тему буде присвячено шляхам управління такими помилками, тобто якістю даних із сподіванням знизити їхній вплив на бізнес-функції організації. Значущість помилки в даних залежить від застосування або призначення, для якого буде використано кожен з наборів даних. Наприклад, розгляньмо такі три сфери застосування даних.

Перша. Маркетингове дослідження може спиратися на адресне геокодування та встановлення певних зв’язків на основі соціально-економічних змінних. Неточна або неповна мережа вулиць зумовить неможливість розташування на карті усіх 100 % адрес, але ймовірно дослідження все ж може проводитися і з меншим розміром вибірки. Вплив обмеженості даних у такому випадку є досить мінімальним.

Друга. Застосування даних для цілей лісового господарства може потребувати виконання моделювання об’єму деревини, що буде заготовлена, на основі таких характеристик, як висота деревостану, видовий склад і діаметр дерев. Помилки у моделі такого виду можуть мати наслідки у вигляді фінансових втрат – робочі бригади можуть виконати рубку на хибних ділянках і фактичний обсяг заготовленої деревини може бути меншим за передбачуваний.

Третя. Прикладна програма маршрутизації руху спецавтомобіля екстреної допомоги використовує детальну мережу вулиць для відправки

поліції, пожежників чи швидкої допомоги до потрібних місцезнаходжень. Помилки в топології мережі вулиць або помилки в адресах можуть спричинити затримку надання екстреної допомоги, що може призвести до втрати життя. Тут помилка є неприпустимою.

Можна помітити, що вимоги до точності даних можуть відрізнятися. Маркетингова компанія може бути геть задоволена 70%-вою точністю вуличної мережі. Агентство з моделювання лісозаготівлі потребуватиме вищої точності, утім 90 % точність може бути прийнятним рівнем ризику для агентства. Маршрутизація руху спецавтомобіля екстреної допомоги, ймовірно, потребуватиме наближення рівня помилки до нуля. Визначення припустимих рівнів помилки є фундаментальним аспектом управління ризиками. Ризики мають бути відповідно оцінені для впровадження можливих заходів з послаблення їхніх наслідків.

3.1 Стратегії управління якістю

Деякі основоположні дії з боку організації можуть значно знизити вплив помилок даних.

Встановлення стандартів методик і даних

Ми повинні визнати, що помилки завжди властиві нашим даним, але через оброблення даних, яке ми виконуємо, помилки можуть збільшуватися через внесення нових помилок. Загалом стандарти можуть встановлювати спосіб створення даних, зміст кінцевого продукту, потрібні метадані та навіть умовні позначки, що використовуються для подавання набору даних. Стандарти можуть бути розроблені власне організацією (тобто внутрішні стандарти), зовнішньою урядовою установою або професійним об’єднанням. Такі організації, як Федеральний комітет з географічних даних (Federal Geographic Data Committee – FGDC) у Сполучених Штатах Америки розробляє стандарти для багатьох загальних застосувань даних, включаючи зміст кадастрових даних, класифікацію рослинного покриву, ґрунтові географічні дані та ризики для довкілля. Подібні організації існують у багатьох державах. Стандартні способи роботи забезпечують узгодженість продукту даних. Якщо дані мають використовуватися і внутрішньо, і зовнішньо, то вони мають створюватися і оброблятися в узгодженому порядку та у спосіб, який скорочує проблеми з даними. Наприклад, хтось може задокументувати найкращий досвід зі створення даних (зокрема, мінімальний розмір полігону, правила класифікації даних, прийнятні рівні генералізації при цифруванні) і проміжного оброблення (зокрема, величину толерантності для виправлення топологічних помилок, методики трансформації координат). Якщо зовнішні стандарти відсутні, то необхідно розробити внутрішні, встановити операційні правила та дотримуватись їх. З процесом

стандартизації та визначення продукції поєднане професійне навчання (підготовка). Забезпечення послідовного та правильного створення даних потребує, щоб усі оператори були ознайомлені зі стандартами. Короткий курс, пройдений усіма технічним персоналом, є вельми невеликою інвестицією порівняно з вартістю виправлення даних, які було неправильно створено або оброблено у неправильний спосіб. Метою ж є забезпечення узгодженого набору даних, який зводить до мінімуму зайві помилки. Узгодженість даних припускає і їхню відтворюваність, коли такий самий набір результатів буде отримано за умови застосування іншим суб'єктом тих самих методів уведення й аналізу даних у ГІС. Якщо ж уведення чи аналіз даних не може бути відтворено, то впевненість в отриманих результатах буде низькою.

Документування оброблення даних і його результатів

Документування слугує двом цілям у процесі оброблення та аналізу даних.

По-перше, воно є засобом відстеження процесу оброблення та забезпечення виконання всіх необхідних кроків для певного набору даних.

По-друге, воно потрібне для запису усіх дій з оброблення, які виконувалися над набором даних у минулому, та надання важливої інформації для споживачів кінцевих даних.

У середовищі комплексного оброблення, де, що цілком може бути, багато різних операторів працюють з багатьма різними наборами даних (з можливим розподілом великих наборів даних за аркушами карти), вдалим практичним рішенням є запис стану кожного набору даних під час його проходження через необхідні кроки оброблення. Це може мати форму "контрольного переліку" з визначеним кожним кроком оброблення (наприклад, реєстрація (запис значення середньоквадратичної помилки – RMS), цифрування, топологічна обробка, введення атрибутів, зміна проекції, контрольна ділянка тощо) і місцем для запису, коли кожен з кроків було виконано. Коли набір даних стає завершеним або як результат виконання аналізу, або як по-новому створений набір даних, необхідно записати хронологію зміни станів набору даних. Ця хронологія стосується послідовності дій щодо набору даних – від їхнього започаткування до поточного стану. При цьому сюди включаються всі джерела даних, методи їхнього введення, кроки оброблення, відомі помилки та проблеми, а також програмне забезпечення, що використовувалося для створення даних. Ця інформація дозволяє будь-яким майбутнім користувачам даними (або навіть Вам як їхньому розробнику, але через кілька років!) отримати уявлення щодо їхньої загальної якості та потенційних проблем.

Оцінювання та перевірка результатів

Управління якістю даних значною мірою потребує, щоб користувачі мали уявлення щодо проблем з даними. Вхідні дані, нові первинні набори даних і результати аналізу даних мають досліджуватися для визначення їхнього рівня помилки. У наступному розділі подано спеціальні методи вимірювання помилки.

У межах загальної програми контролю якості даних перевірку результувальних даних має бути вбудовано в кроки їхнього оброблення. Слід визначити ключові пункти в обробленні даних, в яких найбільш імовірнім є внесення помилки до даних, таких, як цифрування, зміна проекції або накладання полігонів. Далі зусилля щодо контролю якості може бути сконцентровано на цих конкретних пунктах у процесі контролю якості даних. На рисунку 10 показано приклад простого перебігу оброблення даних з вбудованими в нього перевірками якості даних.

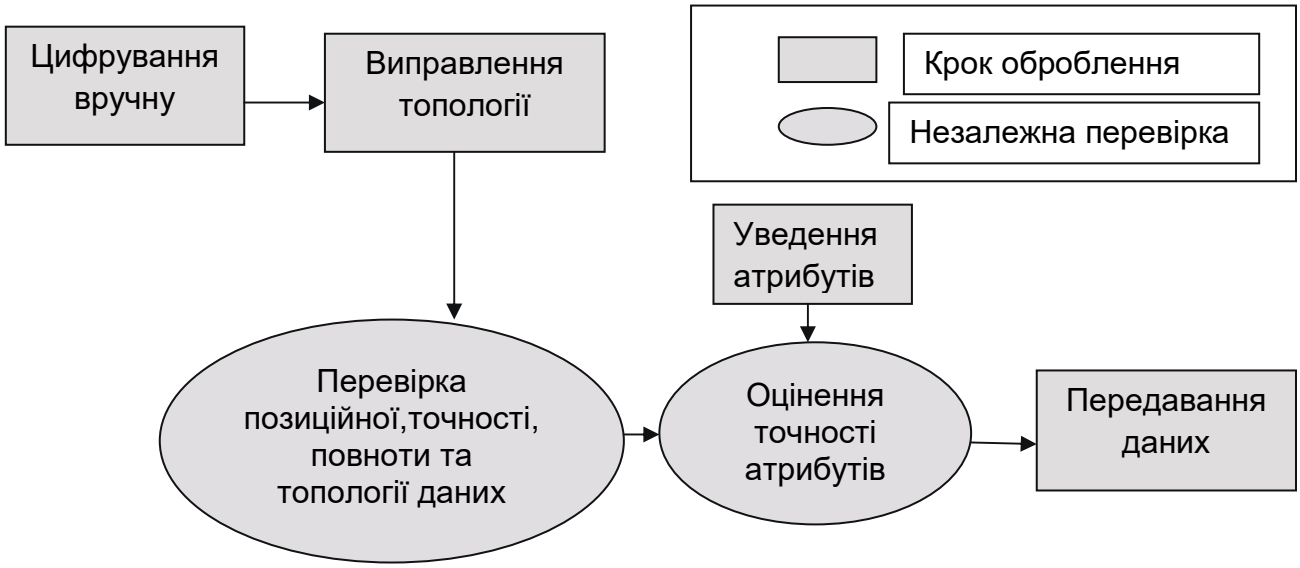


Рисунок 10 - Приклад перебігу оброблення даних з перевірками їхньої якості

Зазвичай ефективно мати незалежного інспектора з перевірки даних. Це може бути другий оператор, не залучений до оброблення даних, або зовнішнє агентство, залучене для виконання перевірки.

Складання звіту щодо невизначеності результувальних даних.

Насправді це є продовженням розглянутого вище в пункті (Документування), однак коли видається кінцевий набір даних або результат введення даних, їхнього перетворення чи аналізу, слід зазначити очікувану помилку даних. В ідеалі кожен шар ГІС матиме межі вірогідності як позиційних, так і атрибутивних значень. Окрім того, звіт має відображати невизначеність даних. Наприклад, якщо розраховано, що

площа полігону становить 1 426, 453 м^2, то мається на увазі, що дані є достатньо точними для підтримки розрахунків з точністю до тисячних. Якщо це не так, відображення трьох знаків після коми є несправжньою точністю. Така точність є просто відображенням результату у спосіб, що означає кращу точність, ніж вона є насправді. Площа 1 426,4 м^2 передбачає точність даних до 0,1 м, а площа 1 426 м^2 – до 1 м. Слід подавати результати у формі, яка збігається з точністю даних. Якщо наші розрахунки площі є точними лише до +/- 10 м, ми повинні округляти їх до найближчих 10 м, тобто площу має бути подано як 1430 м^2.

3.2 Кількісне оцінювання помилок

Щоб прийняти рішення щодо допустимих рівнів помилки, користувачам даних необхідно мати уявлення щодо якості кожного з наборів даних. Існує низка методик оцінювання якості даних, від простих і суто якісних до складних статистичних.

Якісні оцінювання

Найпростішим методом оцінювання даних є їхня візуальна перевірка, і вона має бути частиною будь-якого внесення змін до даних. Очевидні прорахунки або упущення мають бути помічені та виправлені. Для перевірки помилок цифрування можна роздрукувати підсумкові дані в тому ж масштабі, що й вихідний документ, і накласти обидва на "світловому столі". Знову ж таки, при цьому грубі помилки та упущення стануть помітними. Іншим простим методом перевірки оцифрованих даних є “подвійне цифрування” вибіркового аркуша карти. Маючи другого оператора, що цифрує ту саму карту, можна порівняти два результувальні набори даних. Можна зіставити кількість створених полігонів, загальну площу якогось класу атрибутів, довжини ліній та інші характеристики двох аркушів для визначення наявності значних розбіжностей. Цей метод є досить великозатратним способом виявлення помилок, оскільки певна робота має виконуватися двічі.

Можна вивчити розмір розглянутих вище "недольотів", прогалин і "перельотів" до виконання топологічних виправлень. Ці заходи створять у того, хто перевіряє, досить добре уявлення щодо міри відповідних помилок карти. Однак це не буде стосуватися її упущень і грубих помилок.

Атрибути можна швидко перевірити, використовуючи підсумкові статистичні побудови та гістограми. Зазвичай помилки друку атрибутів зумовлюють їхні значення, які не узгоджуються з дійсними величинами. Наприклад, зайва цифра призведе до помилки в значенні абсолютної висоти на порядок (скажімо, 144 м замість 14 м). Такі помилки будуть відображатися в аномальних значеннях гістограм або в узагальнювальних значеннях максимуму та мінімуму.

Кількісні оцінювання

Основою для кількісного оцінювання помилки координат є середньоквадратична помилка (Root Mean Square Error (RMSE) – RMS-помилка). RMS-помилка оцінює відмінності між значеннями координат на карті та координатами того ж просторового об'єкту, отриманими з незалежного джерела даних вищої якості (наприклад, більшого масштабу) або за результатами наземного обстеження. Значення RMS-помилки може бути обчислене за формулою на рисунку 11.

$$\text{RMS-помилка} = \sqrt{\frac{\Sigma\left[(x_i - x'_i)^2 + (y_i - y'_i)^2\right]}{n}}$$

Рисунок 11 – Обчислення середньоквадратичної помилки для координат

Тут *xi, yi* є координатами *i* точки , які мають бути перевірені з набору даних ГІС, *x'i, y'i* є координатами контрольної точки (дійсного місцезнаходження), а *n* є кількістю точок, що перевіряються. Обчислення середньоквадратичної похибки є стандартним методом обчислення горизонтальної позиційної помилки, а стандартні допустимі значення RMS-помилки для даних у певному масштабі публікуються такими організаціями, як, зокрема, Федеральний комітет з географічних даних у США. Наприклад, розгляньмо невеликий набір точок. Нехай ми оцифрували п'ять точок у нашій ГІС і визначили дійсні місцезнаходження цих точок (використовуючи дані крупномасштабного картографування або польові контрольні точки, виміряні із застосуванням GPS або геодезичного обладнання). На рисунку 12 показано такий набір точок.

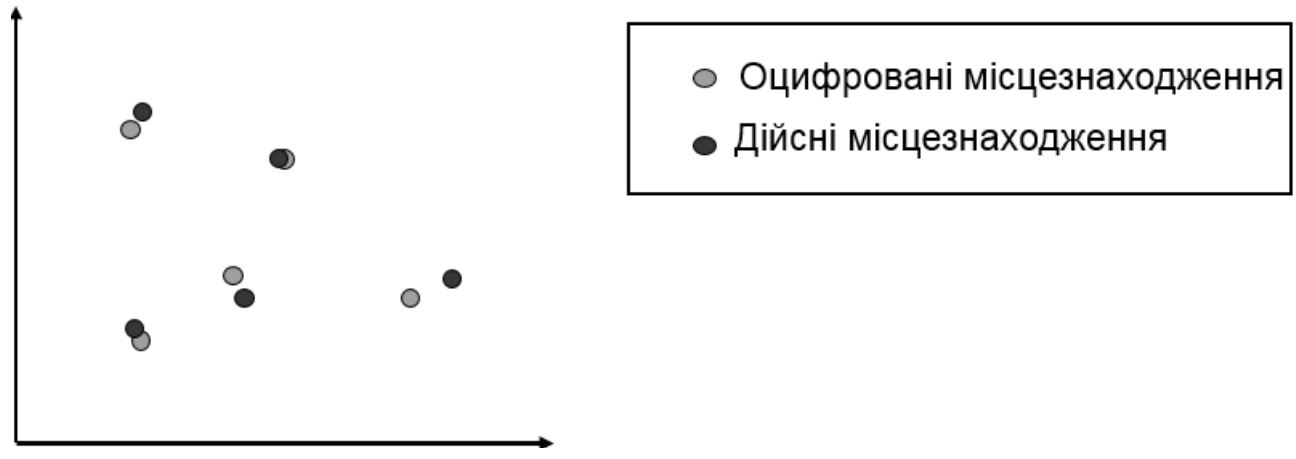


Рисунок 12 – Приклад точкових даних та їхніх оцифрованих і дійсних місцезнаходжень

На рисунку 13 показано розрахунок середньоквадратичної помилки для цих точок. Розрахунок відображено за допомогою електронної таблиці таким чином, щоб було видно кожен крок обчислення. Значення середньоквадратичної помилки 0,3 дає нам загальну міру точності нашого набору оцифрованих даних.

x	y	x'	y'	x - x'	y - y'	$(x - x')^2$	$(y - y')^2$	$(x - x')^2 + (y - y')^2$
1.1	1.0	1.0	1.1	0.1	-0.1	0.01	0.01	0.02
2.0	2.0	2.2	1.7	-0.2	0.3	0.04	0.09	0.13
0.9	3.1	1.1	3.3	-0.2	-0.2	0.04	0.04	0.08
1.4	2.8	1.3	2.8	0.1	0.0	0.01	0.00	0.01
3.2	1.8	3.6	2.0	-0.4	-0.2	0.16	0.04	0.20
							Sum:	0.44

$$\text{RMS Error} = \sqrt{\frac{0.44}{5}}$$

$$\text{RMS Error} = 0.30$$

Рисунок 13 – Приклад розрахунку середньоквадратичної помилки (RMS Error)

У цьому випадку *vi* та *v'i* є, відповідно, закодованими атрибутивними значеннями для *i* спостереження та дійсними атрибутивними значеннями для того ж спостереження.

Матриця неточностей

Рівні помилки для атрибутів, які використовують номінальні класи, такі, як землекористування або класи екологічних класифікацій, не можна розрахувати за допомогою методу середньоквадратичної помилки. Ці класи можуть бути перевірені при використанні перехресних таблиць закодованих і фактичних класів на вибіркових місцезнаходженнях. Це формує матрицю класифікації помилок, або матрицю неточностей. Таблиця 2 відображає приклад такої матриці для простої класифікації землекористування.

Таблиця – 2 Зразок матриці неточностей для класифікації землекористування

Землекористування	AG	BU	FO	FY	LOG	URB	Сума за рядком	Точність користувача
AG	21					1	22	95 %
BU		23	1				24	96 %
FO		3	42	3		1	49	86 %
FY			1	27			28	96 %
LOG					11		11	100 %
URB	2					36	38	95 %

Продовження таблиці 2

Землекористування	AG	BU	FO	FY	LOG	URB	Сума за рядком	Точність користувача
Сумма за стовпцем	23	26	44	30	11	38	172	
Точність виробника	91 %	88 %	95 %	90 %	100%	95 %		93 %

Заголовки рядків і стовпців – це коди класів землекористування. AG – сільськогосподарські території, BU – згарища, FO – старий ліс, FY – молодий ліс, LOG – недавно вирубаний ліс, а URB – урбанізовані території. Лівий стовпець з заголовками рядків відповідає введеним виробником даних класам землекористування. Заголовки у шапці таблиці відповідають правильним класам землекористування. Цифри в центральній частині таблиці відображають взаємовідношення між тим, як було виконане присвоєння класів землекористування, і правильними значеннями для певного класу. При цьому початково було обрано випадкову вибірку місцезнаходжень певних класів землекористування і для кожного з членів цієї вибірки було визначено значення атрибутів обох класифікацій – призначені та дійсні, із записом результатів у перетині відповідного рядка та стовпця таблиці. Таким чином, цифри потовщеним шрифтом уздовж діагоналі відображають правильність класифікацій, тобто кількість місцезнаходжень, де призначені класи та правильні (дійсні) класи збігаються.

Наприклад, розгляньмо рядок із заголовком AG (сільськогосподарські території). Сума за рядком (праворуч цього рядка) становить 22, отже оператор (виробник даних) визначив 22 місцезнаходження як сільськогосподарські території. Число 21 у рядку AG і стовпці AG свідчить, що 21 з цих 22 місцеположень було класифіковано правильно. 1 у стовпці URB (урбанізовані території) свідчить про те, що одне з 22 місцезнаходжень було класифіковане як сільськогосподарські території, тоді як правильним значенням його є урбанізовані території. З правого краю цього рядка у стовпці "Точність користувача" таку точність для місцезнаходжень, класифікованих як сільськогосподарські території, визначено як 95 %. Це розраховано шляхом поділу 21 правильних сільськогосподарських місцезнаходжень на 22 загальних, які виробник даних визначив як сільськогосподарські території. Подібно до цього ми можемо побачити, що виробник визначив 28 місцезнаходжень як молодий ліс (FY). З цих 28 правильними є 27, а одне мало б бути визначене як старий ліс (FO). Точність користувача для цих місцезнаходжень молодого лісу становить 27/28, або 96 %. Вивчивши дані таблиці 2 уже за її стовпцями, ми можемо з одного погляду побачити, яким чином дані мали б

бути класифіковані. Наприклад, з 38 місцезнаходжень, які мали б бути визначені як урбанізовані території (URB), 36 були визначено правильно, одне було неправильно визначено як сільськогосподарські території, а ще одне теж було неправильно визначено вже як старий ліс. Рядок "Точність виробника" знизу таблиці засвідчує, наскільки точно було визначено певний клас землекористування. Наприклад, урбанізовані території було визначено правильно на 36 місцеположеннях з можливих 38, тобто точність виробника становить 95 %.

Зрештою, загальну помилку для цього прикладу може бути розраховано шляхом додавання усіх чисел потовщеним шрифтом по діагоналі (правильні значення) і поділу цієї суми на загальну кількість перевірених місцезнаходжень. Таким чином, загальна точність цього набору даних становить 160/172, або 93 %.

Відмінність між точністю користувача та точністю виробника даних є тонкою та такою, що збиває з пантелику багатьох студентів. Ми можемо розглядати їх як точність відносно того, що є в натурі (дійсне значення), і того, що є на карті (значення, закодоване у базі даних виробником). Точність користувача є часткою землекористувань, визначених на карті, які є правильними. Наприклад, якби на карті було 100 полігонів сільськогосподарських територій, ми могли б припустити, що 95 з них є правильними, базуючись на матриці неточностей, наведеній вище. Точність же виробника даних є часткою землекористувань у натурі, які правильно визначено на карті. Наприклад, якби в дійсності було 100 сільськогосподарських місцезнаходжень, ми могли б припустити, що 91 з них буде правильно визначено на карті.

3.3 Поширення помилок

Однією з головних функцій ГІС є створення похідної (вторинної) інформації за результатами поєднання двох або більше наявних наборів даних. Ми вже обговорили наявність помилки у будь-якому окремому наборі даних, однак нам також необхідно розглянути, як помилки кожного з наборів даних взаємодіють під час поєднання шарів у ході типового ГІС-аналізу. Наприклад, розгляньмо модель ефективності оселищ. Цю модель призначено для визначення зниження ефективності заданого оселища певної тварини внаслідок антропогенних порушень ландшафту. Модель потребує два вхідних набори даних. Першим набором є шар оселищ, в якому растровій сітці або серії суміжних полігонів присвоєно бали від 0,0 (оселище не має жодної корисності) до 1,0 (оселище є оптимальним), відображаючи значення або придатність кожної території як оселища для обраної тварини. Другим набором є шар міри порушення ландшафту. Бал порушення 0,0 позначає максимальний рівень порушення, тоді як 1,0 свідчить, що порушення відсутнє взагалі. Обидва ці шари створено з використанням серії відповідних вхідних даних і дій з їхнього оброблення,

однак з метою спрощення ми просто розглянемо взаємодію саме між двома зазначеними шарами.

Для виконання моделювання ці два шари накладаються і значення щодо оселищ і щодо порушення ландшафтів перемножуються. У такий спосіб оселище отримує значення 0,0 (не має корисності) у всіх випадках множення на 0,0 (максимальне порушення). Множення на бал непорушених територій (1,0) не змінить ранжування корисності оселищ. А ось для територій з балом порушення 0,5 відповідне множення призведе до скорочення корисності оселищ удвічі. У такий спосіб ми й можемо кількісно обчислити загальний рівень деградації для оселищ. Ми дуже впевнені у нашій карті порушень і визначили, що на підставі польових перевірок або іншого незалежного оцінювання точність цієї карти становить до +/- 0,1 (10 % вірогідності). Карта оселищ, імовірно, є більш суб'єктивною, тому що була створена на основі дрібномасштабного картографування, у даних якого ми менш впевнені. Тому точність даних щодо оселищ становить до +/- 0,3 (30 % вірогідності).

Залишається запитання: наскільки упевненими ми можемо бути у наших об'єднаних кінцевих результатах? Як результувальний рівень помилки поєднано з рівнями помилок вхідних даних? Результат стає кращим, чи гіршим? Суттєві дослідження було проінвестовано для вирішення цих питань, і результат засвідчує, що спосіб, яким поширюються помилки, зумовлюється способом поєднання шарів. Наприклад, мультиплікативна модель, така як наша модель ефективності оселищ, наведена вище, спричинює помилку більшу за найгіршу помилку вхідних шарів. Таким чином, у нашій моделі ефективності оселищ результувальний набір даних матиме помилку *більшу* за +/- 30 %. А от адитивні моделі спричинюють рівні помилки менші, ніж у найбільш точного з вхідних наборів даних. Барроу та МакДоннелл (1998) математично оцінили помилку в результаті різноманітних способів поєднання двох просторових шарів. Результати узагальнено нижче в таблиці 3.

Таблиця 3 – Поширення помилки в результаті різних методів поєднання шарів

Метод поєднання шарів	Рівень помилки першого вхідного набору даних (шару)	Рівень помилки другого вхідного набору даних (шару)	Рівень помилки вихідного (результуючого) набору даних
Додавання	10 %	12,5 %	8 %
Віднімання	10 %	12,5 %	70 %

Продовження таблиці 3

Метод поєднання шарів	Рівень помилки першого вхідного набору даних (шару)	Рівень помилки другого вхідного набору даних (шару)	Рівень помилки вихідного (результуючого) набору даних
Множення або ділення	10 %	12,5 %	16 %
Піднесення до степеня (показник степеня 2)	10 %	н.д.	20 %
Операція AND (та)	10 %	12,5 %	≤ 12,5 %
Операція OR (або)	10 %	12,5 %	≤ 10 %

Окрім того, якщо є кореляція між двома вхідними шарами, то вона призводить до збільшення помилки в результуючому наборі даних. У випадку 100 % кореляції між двома вхідними шарами адитивна модель спричинить помилку в розмірі суми двох вхідних помилок. Тобто, наприклад, використовуючи значення з наведеної вище таблиці, бачимо, що цілковито корельовані вхідні шари збільшать помилку для адитивної моделі (методу додавання) від 8 до 22,5 %.

Є кілька способів моделювання поширення помилок. Наприклад, так звані методи Монте-Карло (які покладаються на випадковість, що й дає їхню назву) використовують статистичний розподіл помилки у кожному шарі карти, щоб об'єднати випадкові змінні у кожному шарі, які є пропорційними рівню помилки цього шару. Далі виконується такий аналіз. Нова випадкова змінна додається до кожного шару і аналіз повторюється, потенційно до 100 разів. Усереднення ста вихідних результатів дозволяє обчислити середнє і стандартне відхилення для кожної комірки растрової сітки або області карти. У такий спосіб можна обчислити загальну помилку карти, а також регіональні відмінності внаслідок рівнів помилок.

Нижче надано поради з моделювання поширення помилок, які допоможуть знизити загальну помилку результувальних даних:

- там, де це можливо, застосовуйте адитивні моделі;
- якщо не можете додавати, перемножуйте або діліть;
- уникайте віднімання або піднесення до степеня;
- працюйте з якомога меншою кількістю шарів; відкидання шарів, які роблять дуже малий внесок у результати моделі, може фактично підвищити якість результатів моделювання, оскільки буде менше поєднань шарів;
- уникайте взаємно корельованих випадкових величин;

- усвідомте зміст помилки Ваших вхідних шарів;
- виявіть джерело найбільшої помилки в моделі та попрацюйте над покращенням відповідного вхідного набору даних.

3.4 Аналіз чутливості моделі даних

Часто в ГІС-інструментарії застосовуються методики моделювання або послідовності оброблення, які потребують математичного або логічного поєднання шарів карти. Ми вже побачили, що таке поєднання може призводити до поширення існуючих помилок на аналітичні результати. Однак ми також повинні розуміти, як коливання у вхідних даних моделі будуть впливати на результат моделювання. Процес перевірки того, як модель реагує на такі коливанн, називається аналізом її чутливості. Чутливість моделі може розглядатися як її відповідь на очікувані зміни у вхідних даних або параметрах моделі. Ми можемо досліджувати чутливість щодо:

- Вхідних шарів: наскільки модель реагує, коли значення вхідного шару змінюються?
- Ваги: як модель реагує на зміну ваги, призначеної даному шару?

Під час дослідження вхідних шарів ідея полягає у відокремленні кожного з шарів і дослідженого, як модель реагує на їхні зміни. Це дозволяє нам визначити, який шар робить найбільший внесок у результати моделі, побачити, як впливає помилка певного вхідного шару та розглянути наслідки підвищення або зниження точності даних. Наприклад, розгляньмо комплексну модель, яка допомагає планувальникам визначати розташування нових промислових підприємств. Ми хотіли б обрати місце для таких підприємств, яке мінімізує водночас і вплив на довкілля, і вартість будівництва. Вхідні дані такої моделі можуть містити інформацію щодо ухилів поверхні, геології, природних оселищ, доступності до транспортних магістралей, рідкісних і тих, що зникають, рослинних угруповань, а також земель у приватній власності. Корисним способом виконання аналізу чутливості моделі у цьому випадку є застосування повного діапазону спостережень для перевірки цієї чутливості щодо кожного з шарів. Припустимо, що шар ухилів поверхні у нашому прикладі відображає три класи: рівна, полога та крута поверхні. У такій моделі рівна поверхня території є кращою для будівництва, оскільки вартість будівництва на крутих схилах є значно дорожчою. Ми можемо перевірити чутливість моделі до шару ухилів поверхні шляхом присвоєння усьому вхідному шару найкращого з можливих класів у цих даних, у даному випадку це “рівна поверхня”, і "перезапустити" модель. Надалі ми встановлюємо для всього шару похилів найгірший клас (тобто “крута поверхня”) і знову "запускаємо" модель. Відмінність між справжніми

модельними результатами і цими двома тестовими випадками засвідчує нам, наскільки великі зміни моделі зумовлюються її реагуванням на максимальне і мінімальне з можливих значень похилів поверхні в межах території дослідження. Якщо ми виконаємо цей же аналіз для кожного з вхідних шарів, то матимемо досить добре усвідомлення того, як поводиться модель відносно цього конкретного набору даних або просторового екстенту шару. Це важлива відмінність, оскільки, не зважаючи на те, що можна просто подивитися на модель і визначити, які вхідні дані *теоретично* мають найбільший вплив на зміни результатів моделі (наприклад, за вагою або математичними методами поєднання шарів), характеристики конкретного набору даних можуть засвідчити, що *на практиці* зовсім інший шар матиме домінантний вплив на модель.

Ще раз повернемося до нашої моделі. Якби її було структуровано таким чином, що найбільшу вагу мав би вхідний шар ухилів поверхні, то можна було б очікувати, що шар ухилів буде мати домінантне значення і найбільше впливатиме на модель. Однак якщо вся територія дослідження є рівною поверхнею, то функціонально ухил взагалі не матиме жодного впливу на модельні результати. Природно, що вплив конкретного шару зумовлено математичним апаратом моделі, однак мінливість значень цього шару за територією дослідження також має вплив. Шари з малою мінливістю значень за територією дослідження не роблять значного внеску у модель, незалежно від ваги цих шарів у математичному апараті моделі. На практиці лише невелика кількість шарів, можливо 2 або 3 з 8 або 10 вхідних до нашої моделі, матимуть значний вплив на модельні результати. При цьому важливо знати, які саме це шари, щоб ефективно підтримувати методику моделювання та бути здатним зрозуміти, де в моделі знаходяться головні джерела помилки.

Аналіз чутливості моделі також може бути виконано для дослідження впливів помилки або невизначеності вхідних даних. Наприклад, якщо цифрова модель рельєфу (ЦМР) була б одним з вхідних шарів нашої моделі і точність значень висот було відображено як +/- 5 м, то ми б могли застосувати це значення (5 м) для спостереження впливу невизначеності висот на модель. Ми б могли відняти від кожного значення висоти 5 м і "перезапустити" модель, а потім додати 5 м до усіх значень висот і знову "запустити" модель. Це показало б нам потенційні зміни в моделі на основі можливої помилки у ЦМР. Якщо ми виконаємо моделювання з пониженням значення кожного шару на розмір можливої помилки, а потім знову – з підвищенням значеннями кожного шару на розмір цієї помилки, ми зможемо визначити інтервал вірогідності для загального результату моделі.

Ще одним призначенням аналізу чутливості моделі є визначення найбільш рентабельного просторового розрізнювання вхідних даних. Ми можемо "запускати" модель, змінюючи його, для кількісного оцінювання впливу на результати моделі винятково змін розрізнювання. Надалі ми

можемо дати відповідь на запитання, яке розрізнювання є необхідним для виконання проекту. Чи буде підвищення витрат на отримання даних надзвичайно високого розрізнювання обґрунтованим істотним покращенням модельних результатів? Чи забезпечать дані недорогого дрібномасштабного картографування для певного шару достатню точність для результатів моделювання? Відповіді саме на ці запитання можна отримати, застосувавши аналіз чутливості моделі даних.

3.5 Стандартизація оцінювання та забезпечення якості даних

До системи оцінювання якості будь-якої продукції включають такі типові завдання: формулювання принципів оцінювання та обґрунтування номенклатури показників якості; розроблення методів і процедур визначення показників якості; оптимізацію типорозмірів параметричних рядів виробів; обґрунтування принципів побудови узагальнених показників і умов їхнього застосування у завданнях стандартизації та управління якістю. Ці завдання належать до сфери *кваліметрії* – (від латинського *quails* – який за якістю та ...*метрія*) – наукової дисципліни, що вивчає та розвиває методи кількісного оцінювання якості різних об'єктів. Загальним методичним базисом оцінювання і забезпечення якості геопросторових даних (ГД) є міжнародні стандарти ISO 19157: Якість даних (*Data quality)*, ISO/TS 19158: Забезпечення якості постачання даних (*Quality assurance of data supply*) та ISO 19115: Метадані (*Metdate*). У цьому підрозділі ми докладніше розглянемо застосування принципів, елементів, мір і процедури оцінювання якості наборів геопросторових даних (НГД) за стандартами комплексу ISO 19100 у процесі виробництва, постачання і використання геопросторових даних. Система комплексного управління якістю геоінформаційної продукції та надання геоінформаційних послуг (рисунок 14) ґрунтується на чотирьох основних концептуальних підходах, які відображують особливості змісту показників та організаційних аспектів забезпечення оцінювання якості ГД, а саме [Jakobsson, 2007]: виробничо-орієнтований (*production-centred)*; планувально- або проектно-орієнтований (*plann ing-centred)*; клієнт-орієнтований (*customer-centred)*; системно-орієнтований (*system-centred*) підходи.

Виробничо-орієнтований підхід належить до ключових у комплексній системі управління якістю географічної інформації, оскільки саме від забезпечення якості на етапах збирання інформації та формування НГД найбільшою мірою залежить у майбутньому рівень бездефектності й кондиційності геоінформаційної продукції та послуг.

Проектно-орієнтований підхід додатково передбачає розгляд питань якості ГД з точки зору їхньої придатності для створення та/або розвитку певних прикладних систем, зокрема, з урахуванням придатності даних для використання в середовищі інструментальних ГІС,

що плануються до застосування в цих прикладних системах.

При *клієнт-орієнтованому підході* якість геопросторових даних оцінюється як складова комплексної геоінформаційної продукції (прикладної ГІС), яку отримують кінцеві користувачі для вирішення своїх прикладних задач. Варто зауважити, що кінцевими користувачами якість прикладної ГІС, як правило, сприймається інтегрально, при цьому невизначеність даних може бути причиною недоліків програмних засобів і навпаки.

Рисунок 14 – Структура концептуальних підходів до забезпечення якості ГД у комплексній системі управління якістю надання геоінформаційних послуг

При *системно-орієнтованому підході* якість НГД оцінюється в контексті його інтероперабельності та придатності для розміщення й використання в технологічному середовищі SDI, а також разом з іншими наборами даних та/або уніфікованими геоінформаційними сервісами. Важливою компонентою при оцінюванні якості набіру геопросторових даних з точки зору системно-орієнтованого підходу є повнота, конкретність і достовірність метаданих про набір даних загалом і про його якість зокрема.

Якість може мати різну інтерпретацію залежно від стадії (фази) життєвого циклу (ЖЦ) НГД. У таблиці 4 узагальнено концепти якості для трьох основних стадій ЖЦ НГД [Карпінський, 2012].

Таблиця 4 – Інтерпретація концептів якості для різних стадій (фаз) життєвого циклу НГД

Фаза ЖЦ НГД	Документація про якість	Мета заходів системи управління якістю	Процедури визначення якості	Рівень компонентів НГД
Підготовка виробництва	Технічні вимоги та вхідна модель якості	Визначення вимог якості	Вивчення вимог користувача	Рівень класів об’єктів
Виробництво	База даних з документами історії процесу	Оцінювання відповідності специфікації Зазначення очікуваної якості в базі даних	Ретельний контроль	Екземпляри об’єктів (дата, точність координат тощо)
Використання	Метадані, звіт про контроль якості	Вимірювання відповідності вимогам якості	Оцінювання Звітування	Рівень бази даних

Концептуально процес визначення якості ГД можна розглядати як оцінювання відмінності реально вироблених даних від певного ідеального еталонного набору, в якому немає будь-яких помилок, що виникають у ході формування баз геопросторових даних.

Розподіл помилок розглядається за фазами ЖЦ НГД:

- *збирання даних*: неточності польових вимірювань, неточність приладів, неточність ведення записів, помилки при аналізі даних, отриманих дистанційно;
- *уведення даних*: помилки цифрування, нечіткість природних контурів об’єктів;
- *збереження даних*: числова неточність, просторова неточність (для растрових даних);
- *оброблення даних*: неправильні класифікаційні інтервали, помилки при створенні полігонів;
- *формування кінцевої продукції*: помилки масштабування (для растрових даних), обмеження кінцевого формату даних;

- *використання даних*: неправильне розуміння структури та вмісту, неправильне використання даних.

Більшість дослідників цієї теми розрізняють внутрішню та зовнішню якість ГД. Під *внутрішньою якістю* розуміють рівень відповідності між створеним та «ідеальним» набором, який мав би бути виготовлений (тобто дані, виготовлені без помилок). Такий набір є відображенням реального світу на певну дату відповідно до специфікації на продукцію, що встановлює набір правил і вимог переходу від реального світу до моделі даних. Специфікація містить, наприклад, перелік об'єктів, що мають бути відображені, тип геометрії для кожного класу об'єктів, атрибути, що мають описувати ці об'єкти, а також допустимі значення для атрибутів. Описати внутрішню якість можна різними способами, але в основному використовують критерії, визначені у міжнародному стандарті ISO 19113, а саме: повнота, логічна узгодженість, позиційна, часова й тематична точність. Під *зовнішньою якістю* мають на увазі рівень відповідності готового продукту потребам чи очікуванням користувача. Така якість не є абсолютною, а тому один і той же НГД може мати різну зовнішню якість для різних користувачів. Зовнішню якість часто визначають як придатність продукту для використання. Оскільки саме поняття зовнішньої якості відрізняється для різних користувачів, то не існує єдиних стандартизованих критеріїв для її опису. І. Бедард і Д. Вальєре [Bedard, 1995] виділяють шість характеристик для опису зовнішньої якості ГД:

- призначення: цільове призначення набору даних;
- охоплення: період часу й територія, на яку створено дані;
- походження: методи та процеси оброблення, використані для отримання кінцевих даних;
- точність: відповідність тематичної, часової та просторової точності вимогам користувача;
- легітимність: відповідність створених даних стандартам; ґарантії від постачальника даних;
- доступність: зручність для користувача в отриманні даних (вартість, формат, конфіденційність, авторські права і т. д.).

Загальна структура концепції для визначення якості геопросторових даних за ISO 19157 (рисунок 15) відображує ролі виробників і користувачів у процесі створення та оцінювання якості НГД. Набор геопросторових даних створюється для певного застосування. Якість набору може бути оцінена на основі знань про елементи якості даних, а в деяких випадках – й опосередковано на підставі некількісних елементів, наприклад, призначення та походження даних, що описуються в метаданих згідно з ISO 19115. Елементи якості даних оцінюють рівень відповідності набору даних предметній сфері, що є частиною геопростору, а фактично – технічним умовам (специфікації виробника на створення НГД для певної мети.

Користувач даних оцінює якість НГД як рівень відповідності даних вимогам застосування для геоінформаційного моделювання певної предметної сфери, яка може не збігатися з цільовим призначенням набору даних. А отже, оцінка якості виробника даних може не відповідати оцінці якості потенційного користувача даних. Важливо, щоб вона в контексті мети оцінювання достовірно відображала рівень відповідності НГД специфікації виробника або вимогам користувача. Природно, що при зміні технічних умов виробника або вимог користувача має бути проведено нове оцінювання якості набору даних. Також важливо із застереженням порівнювати різні оцінки якості НГД, якщо його цільове призначення за специфікацією виробника не відповідає предметній сфері потенційного користувача.

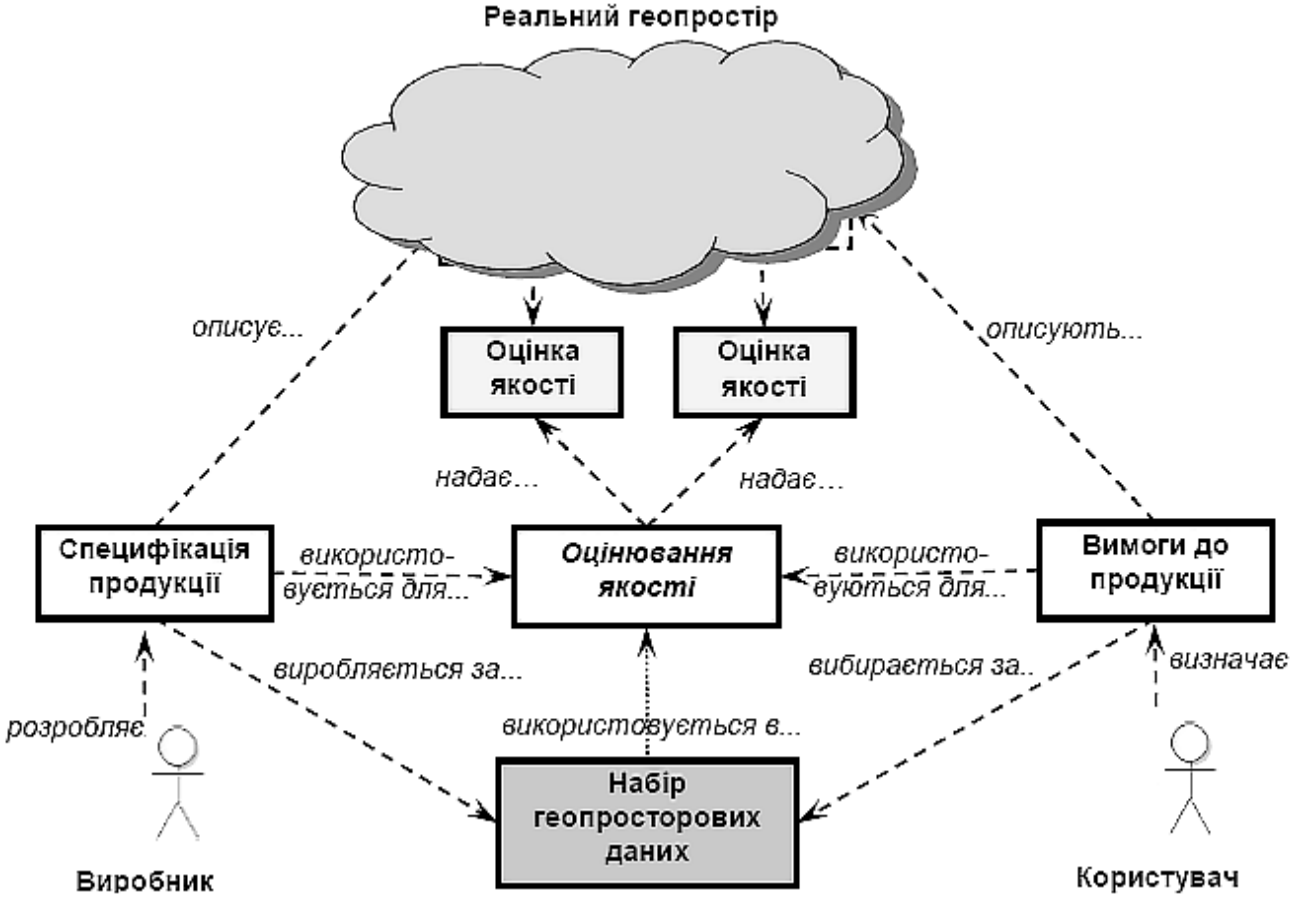


Рисунок 15 – Загальна структура концепції оцінювання якості ГД за ISO 19157

За ISO 19157 передбачається розроблення в складі технічних вимог вхідної моделі якості НГД (рисунок 16), яка в подальшому використовується в процедурах оцінювання переважно внутрішньої якості продукції. Згідно з цією моделлю якість ГД подається за допомогою елементів якості. Для кожного елемента визначаються міри та методи його

оцінювання. Міра якості дає кількісну характеристику елемента якості. Методи оцінювання якості визначають підходи до перевірок даних та обчислення мір. Окрім цього, кожен елемент якості описується елементами метаякості, що характеризують ступінь довіри до результату оцінювання і містять обґрунтування доцільності застосування обраної міри якості та методу оцінювання для конкретного елемента. Інформація про якість даних подається в спеціальному звіті про оцінювання якості та в метаданих. Звіт про якість складається обов'язково як результат оцінювання за процедурами та методами, визначеними у вхідній моделі якості. Найчастіше такий звіт використовується в подальшому для виправлення помилок, виявлених в НГД під час контролю та оцінювання якості. Інформація про якість також описується у відповідних розділах та елементах метаданих, уніфікованих за ISO 19115. Метадані включаються до НГД як кінцевого продукту та можуть використовуватись споживачами наборів у процесі оцінювання зовнішньої якості даних.

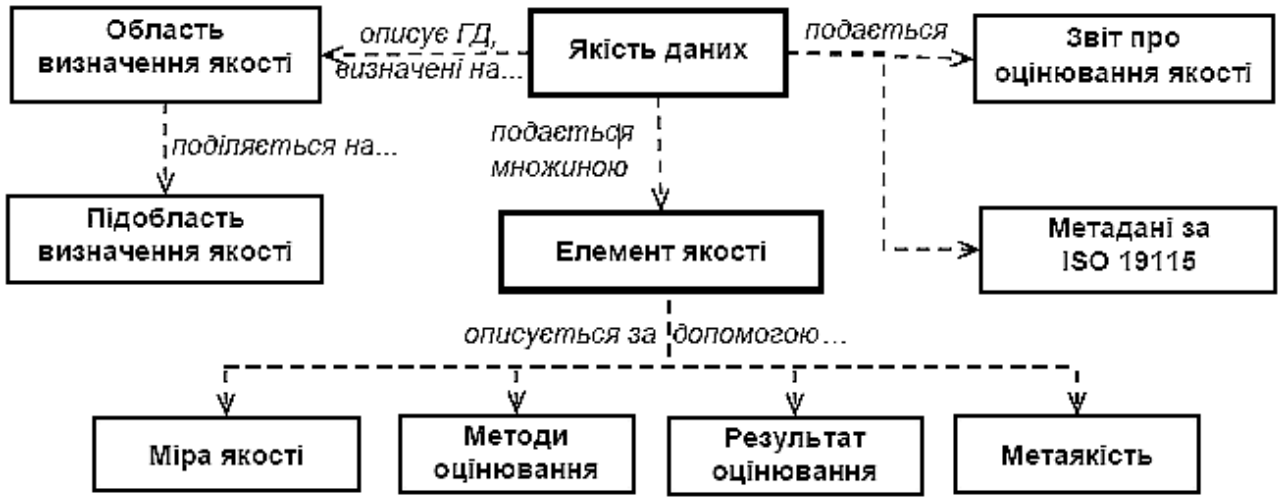


Рисунок 16 – Концептуальна модель якості геопросторових даних за ISO 19157

Поєднання елементів якості та меж їх застосування встановлює область визначення якості. Область описує просторові та часові рамки застосування кожного елемента. Нею може бути набір даних, серія наборів або фрагмент даних, обмежений однією або кількома характеристиками: клас об'єктів, екземпляр об'єкта, територіальне охоплення, часове охоплення. Загалом набір даних розглядається як такий, що містить велике, хоча й скінченне число менших груп даних. Припускається, що менші групи даних, об'єднані за належністю до того самого типу об'єкта, типу атрибуту об'єкта або відношення, критерію збору або до однієї і тієї ж географічної області, мають однакову якість. Менша група даних може складатися з одного екземпляра типу об'єкта, значення атрибута або відношення. Теоретично за концепцією оцінювання якості

даних допускається, що кожен екземпляр типу об'єкта, значення атрибута й вид відношення набору даних може мати окремі елементи якості. Якість менших груп даних може не збігатися з якістю іншої частини набору даних, до якого вони належать. Концепція оцінювання якості даних допускає видачу інформації про якість набору загалом і додатково інформацію про якість менших груп, що визначається специфікою профілю ГД. Згідно з ISO 19157 якість НГД вказується із використанням *кількісних* та *описових* елементів.

Кількісні елементи якості геопросторових даних (таблиця 5) дають змогу оцінити, наскільки той або інший набір даних відповідає критеріям, указаним у специфікації на відповідну продукцію. Аспекти елементів якості даних називають піделементами. Піделементи оцінюють або перевіряють різними способами. Як уже зазначалося, за концепцією оцінювання якості даних не всі елементи й піделементи, а також не всі засоби їхнього оцінювання та перевірки можуть підходити конкретному набору даних. Крім того, деякі піделементи можуть застосовуватися до всього набору даних, можуть бути оцінені або перевірені для нього, а інші застосовуються до менших груп даних, що належать до набору даних вищого порядку та можуть бути оцінені або перевірені для них.

ISO 19157 визначає елементи якості даних безпосередньо як засоби виявлення і вироблення різних видів інформації про якість. Як правило, піделементи якості даних взаємозалежні. Наприклад, помилка в координатах може викликати, як мінімум, помилки двох типів – позиційну і топологічну. Значення піделементів, передбачене в угоді на продукт, та спосіб їх трактування належать до компетенції розробника специфікації на продукт.

Таблиця 5 – Кількісні елементи і піделементи якості за ISO 19157

Назва елемента та піделемента якості та їх ідентифікатори в метаданих	Опис змісту
Повнота даних *DQ_Completeness*	Наявність чи відсутність об’єктів, їх атрибутів і відношень
Надлишковість *DQ_Commission*	Надлишкові дані, наявні в наборі даних
Відсутність *DQ_Omission*	Дані, відсутні в наборі даних
Логічна узгодженість даних *DQ_LogicalConsistency*	Ступінь відповідності логічним правилам структури даних, атрибутики та відношень
Концептуальна узгодженість *DQ_ConceptualConsistency*	Відповідність правилам концептуальної схеми

Продовження таблиці 5

Назва елемента та піделемента якості та їх ідентифікатори в метаданих	Опис змісту
Доменна узгодженість *DQ_DomainConsistency*	Відповідність значень домену
Форматна узгодженість *DQ_FormatConsistency*	Ступінь відповідності зберігання даних фізичній структурі набора даних
Топологічна узгодженість *DQ_TopologicalConsistency*	Правильність експліцитно зако-дованих топологічних характе-ристик набора даних
Позиційна точність об’єктів *DQ_PositionalAccuracy*	Точність місцеположення об’єктів
Абсолютна чи зовнішня точність *DQ_AbsoluteExternalAccuracy*	Близькість значень координат значенням, вказаних у звіті, прийнятим як правильні
Відносна чи внутрішня точність *DQ_RelativeInternalAccuracy*	Близькість відносних місце-положень об’єктів, вказаних у наборі даних, відповідним місцеположенням, прийнятим як правильні
Точність місцеположення даних у комірках *DQ_ GriddedDataPositionAccuracy*	Близькість значень місце-положення, зазначених для да-них у комірках, значенням, що прийняті як правильні
Тематична точність даних *DQ_ThematicAccuracy*	Точність числових атрибутів, правильність нечислових атри-бутів, класифікації об’єктів та їх відношень
Правильність класифікації *DQ_ClassificationCorrectness*	Порівняння класів об’єктів та їх атрибутів з певною предметною сферою (тобто відповідність базовим концептам предметної сфери або концептам еталонного набору даних)
Правильність нечислових атрибутів *DQ_NonQuantitativeAttributeCorrectness*	Правильність нечислових атри-бутів
Точність числових атрибутів *DQ_QuantitativeAttributeAccuracy*	Точність числових атрибутів
Часова точність даних *DQ_TemporalQuality*	Точність часових атрибутів і часових відношень об’єктів

Продовження таблиці 5

Назва елемента та піделемента якості та їх ідентифікатори в метаданих	Опис змісту
Точність часового вимірювання *DQ_AccuracyOfATimeMeasurement*	Правильність часових зв'язків елемента
Часова узгодженість *DQ_TemporalConsistency*	Правильність упорядкованих по-дій або послідовностей, якщо про них звітують
Часова відповідність DQ_TemporalValidity	Коректність даних з плином часу

Описові елементи якості даних забезпечують загальну некількісну інформацію про якість. Вони дозволяють додатково оцінити придатність набору даних для конкретного застосування, містять інформацію про його призначення, використання й походження. Елементи "Призначення" і "Використання" описують сферу застосування набору даних. Використання набору даних визначається його розроблювачем або користувачами даних. Елемент "Походження" описує історію формування набору даних і певною мірою його життєвий цикл, починаючи з процесів збирання, наступного кодування й перетворення у поточний формат даних.

Міра якості даних є кількісною характеристикою ГД. Уніфікація мір якості здійснюється з метою досягнення сумісності та порівнюваності кількісної інформації про якість різних наборів даних. Однією з основних вимог до мір якості є однозначність їх визначення та коректність методів обчислення. В ISO 19157 пропонується набір стандартизованих мір якості, які дозволяють оцінювати практично всі кількісні елементи та піделементи якості ГД. Кожна міра якості (DQM_Measure) описується такими компонентами (рисунок 17):

- ідентифікатором міри якості *(measureIdentifier)*;
- назвою міри якості (*name*);
- псевдонімом *(alias)*;
- назвою елемента якості *(elementName)*, до якого застосовується міра;
- базовою мірою якості *(basicMeasure)*;
- визначенням *(definition)*: фундаментальним концептом міри якості (якщо міра базується на одній з базових мір, то дається її визначення);
- описом *(description)*: описом міри якості, включаючи всі методи обчислень і формули, необхідні для застосування міри;

- параметром *(parameter)*: змінною, яку використовує міра якості (містить назву, визначення і тип даних параметра);
- типом значення *(valueType)*: одним із типів даних, що використовується для отримання результату міри (визначається за ISO/TS 19103:2005);
- структурою значення *(valueStructure)*, якщо результат містить декілька значень;
- посиланням на джерело *(sourceReference)*: посиланням на документацію міри (якщо для міри якості додаткова інформація міститься у зовнішньому джерелі, то вказується посилання на це джерело);
- прикладом застосування міри *(example).*

В описі базових мір якості виділяють два класи:

1) міри, що базуються на підрахунку помилкових або правильних об’єктів;

2) міри, що базуються на моделюванні невизначеності вимірювань статистичними методами.

У першому класі розрізняють шість базових мір якості (таблиця 6), що ґрунтуються на різних методах підрахунку кількості помилкових чи правильних об’єктів. Числові дані, отримані в результаті вимірювань, мають певну точність, а тому для оцінювання ступеня невизначеності якоїсь виміряної величини рекомендується використовувати статистичні методи.

Таблиця 6 – Базові міри якості, основою яких є підрахунок кількості помилок

Назва міри якості	Визначення	Приклад	Тип значення
Індикатор помилки	Показник того, що об’єкт виявлено як помилковий	False	Логічне
Індикатор правильності	Показник того, що об’єкт виявлено як правильний	True	Логічне
Кількість помилок	Загальна кількість помилкових об’єктів у наборі	11	Ціле

Продовження таблиці 6

Назва міри якості	Визначення	Приклад	Тип значення
Кількість правильних об’єктів	Загальна кількість правильних об’єктів у наборі	15 571	Ціле
Відсоток помилок	Відношення кількості помилкових об’єктів до їх загальної кількості	0,0189	Дійсне
Відсоток правильних об’єктів	Відношення кількості правильних об’єктів до їх загальної кількості	0,9811	Дійсне

У міжнародному стандарті ISO 19157 застосування базових мір якості конкретизовано для усіх кількісних елементів і піделементів якості ГД, що в підсумку дозволило ідентифікувати понад 80 окремих мір якості. Однак через специфіку якості даних цей список не може бути повним. З розвитком ГІС вимоги до якості постійно зростають, тому природним є розроблення додаткових мір якості.

Процес, процедури та методи оцінювання якості. Оцінювання якості геопросторових даних здійснюється на різних стадіях життєвого циклу продукції. Воно має різні цілі для кожної стадії. До основних стадій життєвого циклу набору геопросторомих даних можна віднести: розроблення технічних вимог (специфікації), виробництво, постачання, використання та оновлення.

Процес оцінювання якості – це послідовність етапів, операцій та процедур, виконання яких дозволяє отримати результат як сукупність елементів якості для визначеної області (набору даних, окремих екземплярів об’єктів, їх атрибутів або відношень). Загалом процес складається із шести основних етапів (рисунок 17).

Наголосимо на важливості перших трьох етапів, результатом яких є фактично специфікація моделі якості для процесу оцінювання якості конкретного набору даних з обґрунтуванням вибору

елементів/піделементів, мір якості для них, і процедур і методів оцінювання. Ця модель має узгоджуватися із вхідною моделлю якості, що зазначається у специфікації на продукцію або в технічних вимогах користувача.

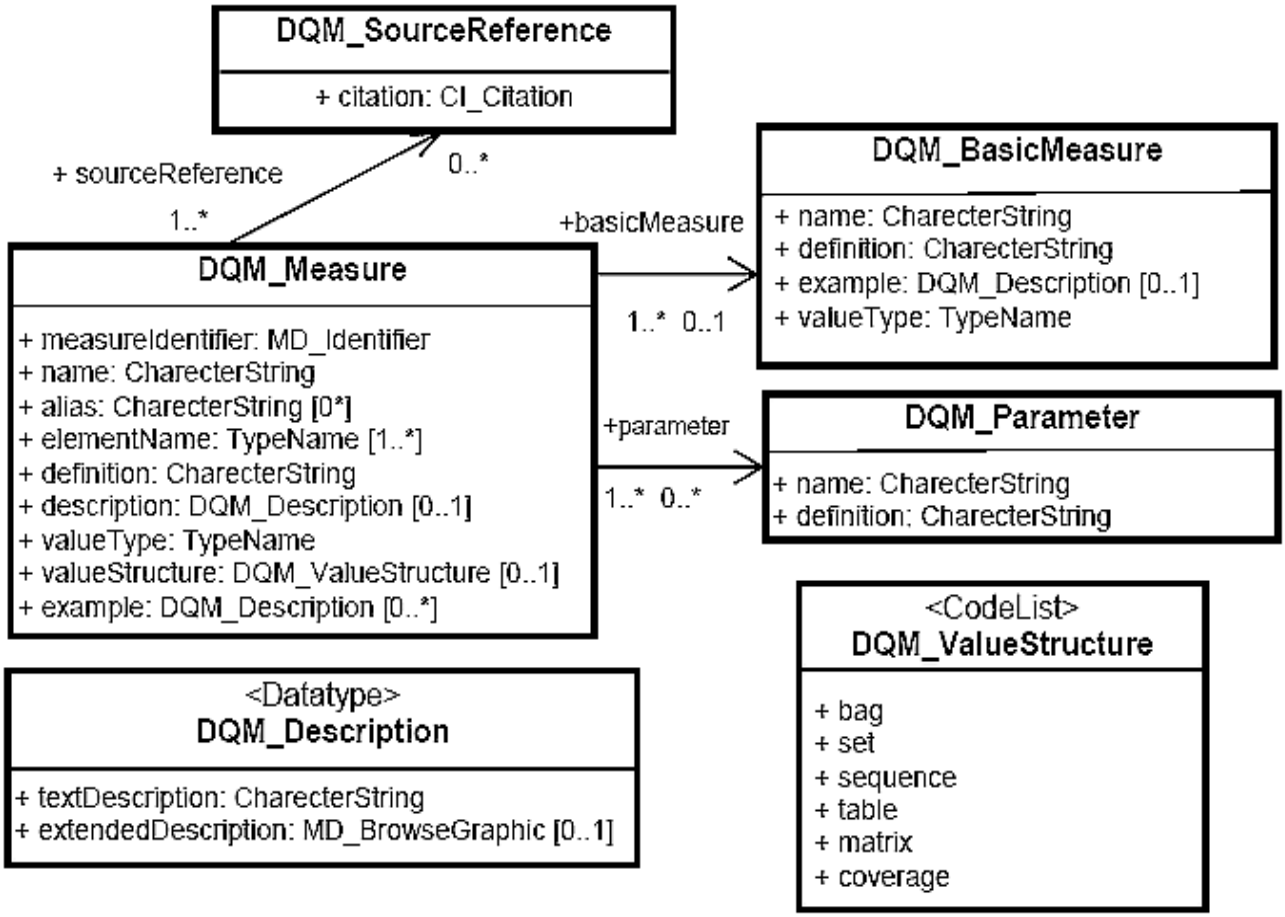


Рисунок 17 – Структура опису мір якості геопросторових даних за ISO 19157

Процедури оцінювання якості визначають порядок застосування одного або кількох методів оцінювання для певного елемента/піделемента якості даних. Методи оцінювання якості поділяються на два основних класи: прямі та непрямі.

Прямі методи ґрунтуються на порівнянні оцінних даних з еталонними, *непрямі* – на використанні довідкової інформації про дані, наприклад, про їх походження. Прямі методи за джерелом інформації, необхідної для проведення порівняння, в свою чергу, поділяються на внутрішні та зовнішні. Усі дані, необхідні для виконання внутрішнього прямого методу, є внутрішніми відносно оцінного набору даних. Наприклад, для перевірки топологічної узгодженості меж земельних ділянок потрібні лише дані про межі. Зовнішні методи потребують еталона

для порівняння. Наприклад, для перевірки повноти атрибутивної інформації про назви населених пунктів потрібно мати додаткове джерело інформації про ці назви, а для перевірки позиційної точності – координати еталонних (контрольних) точок. Схему процесу оцінювання якості геопросторових даних за ISO 19157 зображено на рисунку 18.

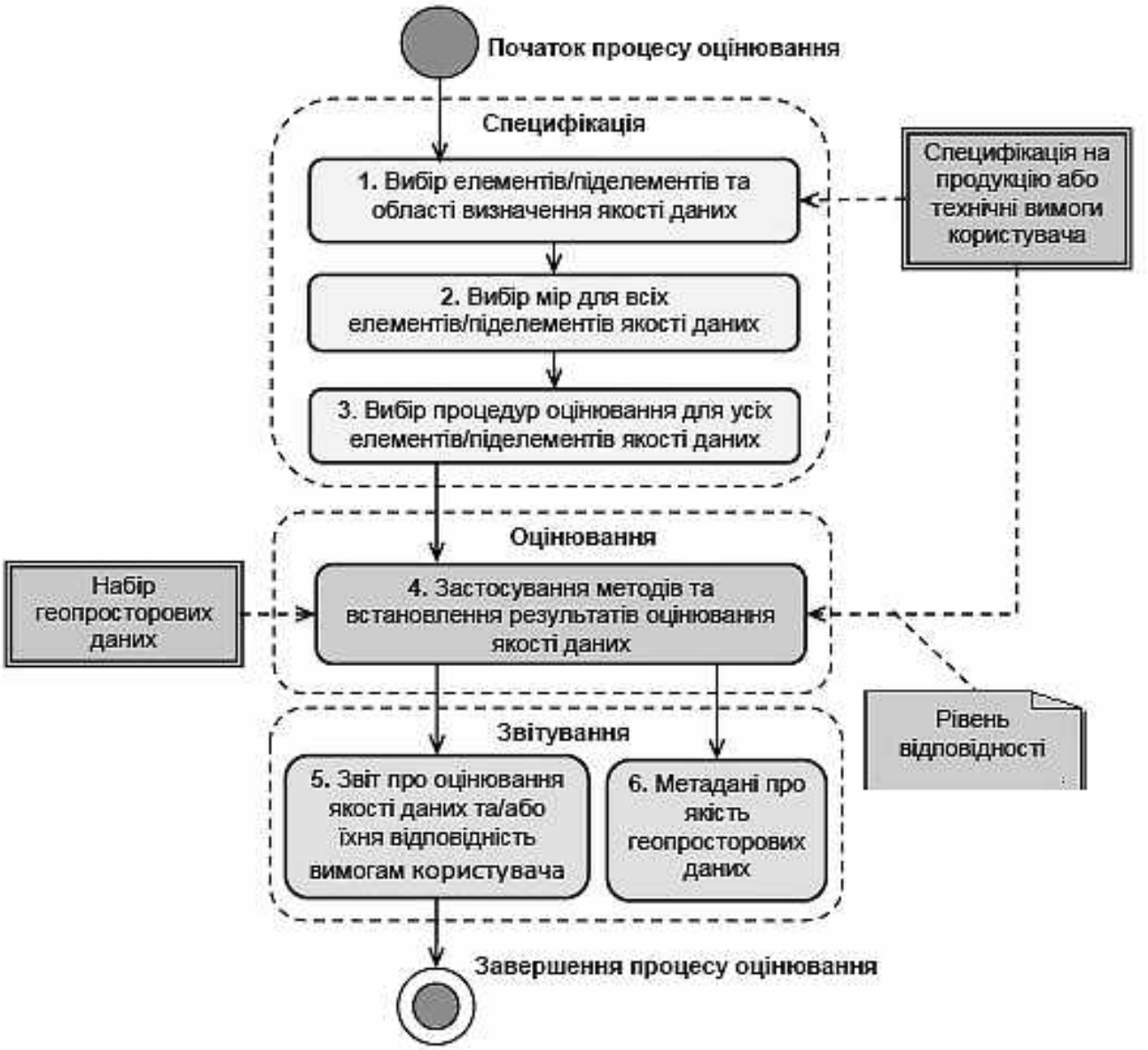


Рисунок 18 – Загальна схема процесу оцінювання якості геопросторових даних за ISO 19157

Непрямі методи ґрунтуються на зовнішніх знаннях. Наприклад, інформацію про позиційну точність можна одержати, якщо відомо методи вимірювання координат точок. Проте методи непрямого оцінювання рекомендується застосовувати лише у випадках, коли не можна використати прямі методи.

Як для внутрішніх, так і для зовнішніх методів можливе застосовування вибіркового або повного контролю. Повний контроль

передбачає перевірку всіх об’єктів з набору даних. Його доцільно використовувати для невеликих наборів або у випадках, коли контроль можна здійснити автоматизованими засобами за допомогою комп’ютерних програм. В інших випадках доцільно проводити вибірковий контроль порівнянням певної частини набору даних з еталоном. Так, як правило, перевіряється близько 5 % даних.

На завершальних етапах процесу оцінювання якості зазвичай виводиться узагальнений показник якості даних – ADQR (Aggregated Data Quality Result), в якому агрегуються окремі результати оцінювання елементів/піделементів якості.

Система забезпечення якості геопросторових даних. Відомо, що контроль якості продукції на кінцевій стадії виробничого процесу дає неадекватну оцінку. Він не може ефективно впливати на поліпшення якості продукції, а тому необхідна система наскрізного контролю якості всіх елементів на всіх етапах виробничого процесу. Наскрізний контроль охоплює загалом стан виробництва та визначається за ISO 8402 як підхід “Повна якість”. Реалізація даного підходу потребує використання системи управління якістю QMS (Quality Management System), зокрема, визначення повноважень, відповідальності та порядку взаємодії усіх учасників та усіх основних процесів і процедур, важливих для досягнення якісного виконання завдань організації та проведення довгострокової сталої політики якості. Систему забезпечення якості геопросторових даних можна визначити як сукупність організаційно-технологічних заходів, методик і спеціального програмного забезпечення для контролю якості даних на усіх етапах збирання, уведення, оброблення і використання даних, що реєструються, накопичуються і оновлюються в базі даних та експортуються у зовнішні системи як НГД та/або цифрові й електронні карти, створені на основі вмісту БД. Контроль якості даних є обов’язковою складовою технологічної ланки в реалізації будь-якого способу отримання, введення, реєстрування та виведення ГД.

ISO/TS 19158 Забезпечення якості постачання даних визначає рамки для виробника і споживача в їх виробничих відносинах. Він ліквідує розрив між системами управління якістю, концепцію яких визначено в стандартах комплексу ISO 9000, і технічно-орієнтованими стандартом якості ISO 19157. Положення, визначені в ISO/TS 19158, дозволяють клієнту переконатися в тому, що постачальники як внутрішні, так і зовнішні, здатні поставляти геопросторові дані необхідної якості, виходячи з дотримання процедур системи управління якістю за ISO 9000 у процесі виробництва/оновлення ГД з урахуванням вимог ISO 19157 до якості геопросторових даних. В ISO/TS 19158 визначено принципи та обов'язки у взаємовідносинах між замовником і постачальником даних, включаючи розподіл відповідальності за процедури оцінювання якості геопросторових даних між замовником і постачальником.

4 Практична робота

Цю практичну роботу буде зосереджено на формуванні вмінь з кількісного оцінювання помилок в географічних даних. Студенти набувають практичного досвіду щодо:

- оцінювання правильності класифікації землекористування;
- створення матриці неточностей для узагальнення помилок атрибутів.

У цій практичній роботі буде обчислено тематичну або атрибутивну точність набору даних. Розглянуто набір даних щодо землекористування масштабу 1:50 000, який було скориговано для цієї практичної роботи (наприклад, було додано помилки тощо).

Ціль: необхідно створити матрицю неточностей для дев'яти класів землекористування, наявних на території дослідження, та обчислити загальну атрибутивну точність.

Для виконання цієї практичної роботи знадобляться такі файли даних:

Шейп-файлы:

- Landuse.SHP
- Sample_Points.SHP

Ортофотознімки Вільнюського регіону:

- 7530, 7531, 7532, 7533
- 7630, 7631, 7632, 7633, 7634, 7635
- 7730, 7731, 7732, 7733, 7734, 7735
- 7831, 7832, 7833, 7834, 7835
- 7932

Файли знаходяться в компьютері в робочій папці «Організація і управління геодезичними та земельно-кадастровими роботами» (ОУГЗКР).

Набір даних щодо землекористування містить класи, наявні на території дослідження (таблиця 7).

Таблиця 7 – Набір даних щодо землекористування

Код класу	Назва класу	Опис класу
15 11 110	Агроугіддя	Територія охоплює луки, пасовища, ріллю та інші відкриті території, які не включено до інших класів. Дороги без значних прилеглих до них урбанізованих структур розглядаються як агроугіддя

Продовження таблиці 7

Код класу	Назва класу	Опис класу
		Крім того, якщо певні землі не відповідають жодному з поданих далі класів землекористування, то їх слід класифікувати теж як агроугіддя (тобто це є класом землекористування “за замовчуванням”)
90 50 120	Забудова	Урбанізовані території міст, селищ міського типу та сіл, а також інші урбанізовані території (фабрики, електростанції, злітно-посадкові смуги, землі військового призначення, нафтові сховища та інші нежитлові забудовані території за межами міст) Цей шар містить усі житлові урбанізовані території понад 50х50 м (2500 м^2). Усі садиби подано незалежно від розміру їхньої площі
90 50 150	Цвинтарі	Мають бути більшими 70х70 м (5000 м^2)
15 10 120	Ліси	Усі ліси, лісові території (лісонасадження, чагарники, лісопарки) і тимчасово незаліснені території (ділянки вирубок) понад 150х150 м (22 500 м^2), а також невеликі лісосмуги та просіки, що виконують межову роль (наприклад, на територіях з низькою густотою заселення, заболочених землях та ін.) з розмірами, більшими 100х100 м (10000 м^2).
14 10 120	Озера, ставки	Мають бути більшими 50х50 м (2500 м^2).

Продовження таблиці 7

Код класу	Назва класу	Опис класу
15 12 000	Піски, торфовища, кар'єри, звалища	Піски мають бути більшими за 200х200 м (40 000 м^2). Смуги піску не можуть бути меншими 50 м. Торфовища, кар’єри, звалища повинні бути більшими 150х150 м (22 500 м^2)
14 10 210	Водотоки, ширші 30 м	
15 11 120	Дачі, фруктові сади	Шар містить ділянки дач, фруктові та ягідні насадження. Площі мають бути більшими 10 000 м^2. За відсутності насаджень дачі інколи може бути відділено від забудованих територій за їхньою більш дрібною структурою та більш щільним деревним покривом на їхній території
13 35 200	Болота, водно-болотні угіддя	Території постійно вологих ґрунтів більші 250х250 м (50 000 м^2). Болота, які правлять за межі, мають бути ділянками більше 150х150 м (22 500 м^2)

Полігони зазначених вище класів землекористування створено на основі литовського набору даних LTDBK50000-V. Для мети практичної роботи зроблено припущення, що ці полігони було створено вручну на основі дешифрування зображень з відносно низьким розрізнюванням. Для перевірки правильності цієї класифікації полігонів за землекористуванням необхідно скористатися більш детальними кольоровими ортофотознімками розміром піксела 0,5 м. Треба проаналізувати землекористування у серії заданих (контрольних) точок, використовуючи ортофотознімки, і зіставити результати дешифрування зображень (очевидно більш точних, оскільки отриманих зі знімків з більш високим розрізнюванням) з вихідною класифікацією полігонів за землекористуванням.

4.1 Етапи виконання роботи

1. Створіть таблицю, використовуючи програмний інструментарій електронних таблиць або текстового редактора, з дев’ятьма класами землекористування за обома осями, як показано нижче. У цьому прикладі

наведено повні коди та назви класів, однак можливе застосування коротших записів або абревіатури за умови їхньої відповідності вихідним.

Таблиця 8 – Класи землекористування

Класи землекористування	151111 Агроугіддя	9050120 Забудова	9050150 Цвинтарі	1510120 Ліси	1410120 Озера, ставки	1512000 Піски, торфовища, кар'єри	1410210 Водотоки, ширші 30 м	1511120 Дачі, фруктові сади	1335200 Болота, водно-болотні угіддя	Сума за рядком	Точність виробника
151111 Агроугіддя											
9050120 Забудова											
9050150 Цвинтарі											
1510120 Ліси											
1410120 Озера, ставки											
1512000 Піски, торфовищакар'єри											
1410210 Водотоки ширші 30 м											
1511120 Дачі, фруктові сади											
1335200 Болота, водно-болотні угіддя											
Сума за стовпцем											
Точність виробника											

2. Запустіть ArcMap.

3. Додайте шейп-файли Sample_Points.SHP і Landuse.SHP з папки ОУГЗКР (див. дод.1).

4. Виконайте відповідне панорамування та/або масштабування щодо кожної контрольної точки у шарі Sample_Points. У кожній такій точці цього файла Ви проаналізуєте ортофотознімок і визначите, який клас землекористування Ви вважаєте правильним для цієї точки, а потім порівняєте це з вихідною класифікацією, яку знайдете у шарі Landuse. Надалі Ви оновите Вашу матрицю неточностей для відображення

отриманих Вами відомостей щодо певної точки. Після аналогічної перевірки усіх точок Ви й обчислите точність шару Landuse.

Нижче наведено приклад процесу створення матриці неточностей.

– Виконаємо панорамування та/або масштабування щодо вашої першої контрольної точки (Sample_Point). Для такої першої точки відтворення на екрані монітора має бути подібним до наступного (рисунок 19).

– Визначимо кодоване значення землекористування для цієї точки, тобто клас землекористування, який було початково застосовано ГІС-техніком (виробником даних), який створював полігони землекористування Landuse. У наведеному вище прикладі ці полігони землекористування було подано як напівпрозорі з розміщенням їх вище кольорового ортофотознімка (див. дод. 2). Ми можемо побачити за кольором полігону (жовтий), у межах якого знаходиться контрольна точка № 1, що клас, початково закодований для цієї точки, це *Агроугіддя*.

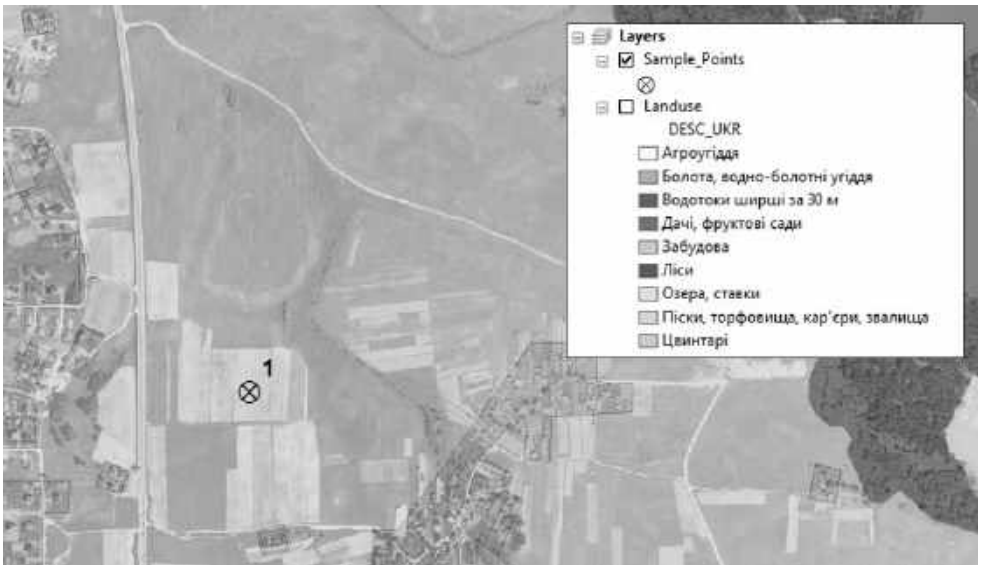


Рисунок 19 – Відображення контрольної точки № 1 на екрані

– Визначимо правильне (дійсне) значення землекористування для зазначеної точки. У нашому прикладі кодоване значення землекористування у точці № 1 видається правильним, тобто з огляду на наші можливості дійсним значенням землекористування є також *Агроугіддя*. Для того щоб зробити таке визначення Вам слід ураховувати описи класів землекористування, наведені у таблиці 7 на початку цієї практичної роботи. Вам може також знадобитися використання інструмента *Вимірювання* (*Measure Tool*) для визначення площі певного полігона або ширини.

– Додамо щойно проаналізовану точку до Вашої матриці неточностей. Тобто для точки № 1 на зображенні вище ми оберемо в нашій таблиці рядок, позначений як *Агроугіддя* (рядки містять кодовані

значення або значення виробника даних) і знайдемо його перетин зі стовпцем, позначеним як *Агроугіддя* (стовпці містять дійсне (правильне) значення землекористування). Додамо одиницю до значення, яке було у цій комірці перетину. Якщо ж це наша перша контрольна точка (що так і є у цьому прикладі), то ми просто розмістимо 1 у комірці перетину, як показано на рисунку 20.

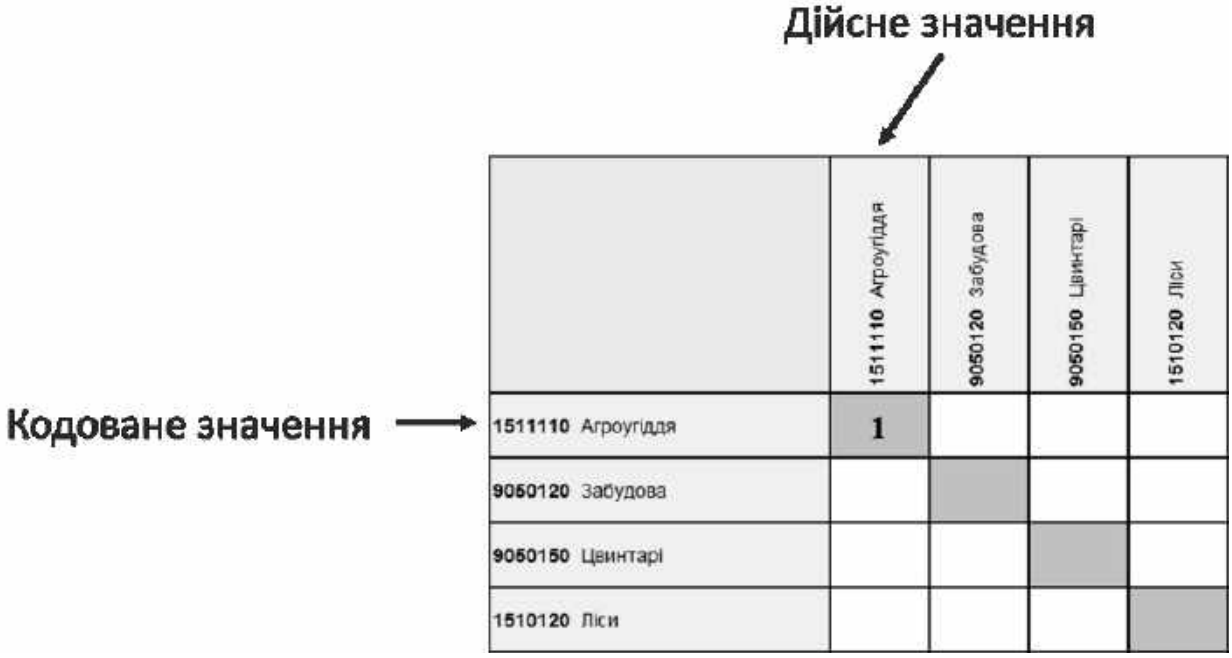


Рисунок 20 – Приклад заповнення матриці для першої контрольної точки

– Проаналізуємо другу точку. Припустимо, цю точку початково було кодовано як *Агроугіддя*, однак Ви розумієте, що вона в дійсності є *Забудовою* (рисунок 21).

Рисунок 21 – Відображення другої контрольної точки на екрані

– Знову оновимо нашу матрицю неточностей, знаходячи кодоване значення землекористування у заголовках рядків, а дійсне (правильне) значення землекористування (як ви його визначили) у заголовках стовпців. Надалі додамо одиницю до значення, яке було у комірці перетину. У даному випадку, результат додавання нашої другої контрольної точки у таблицю виглядав би наступним чином (рисунок 22).

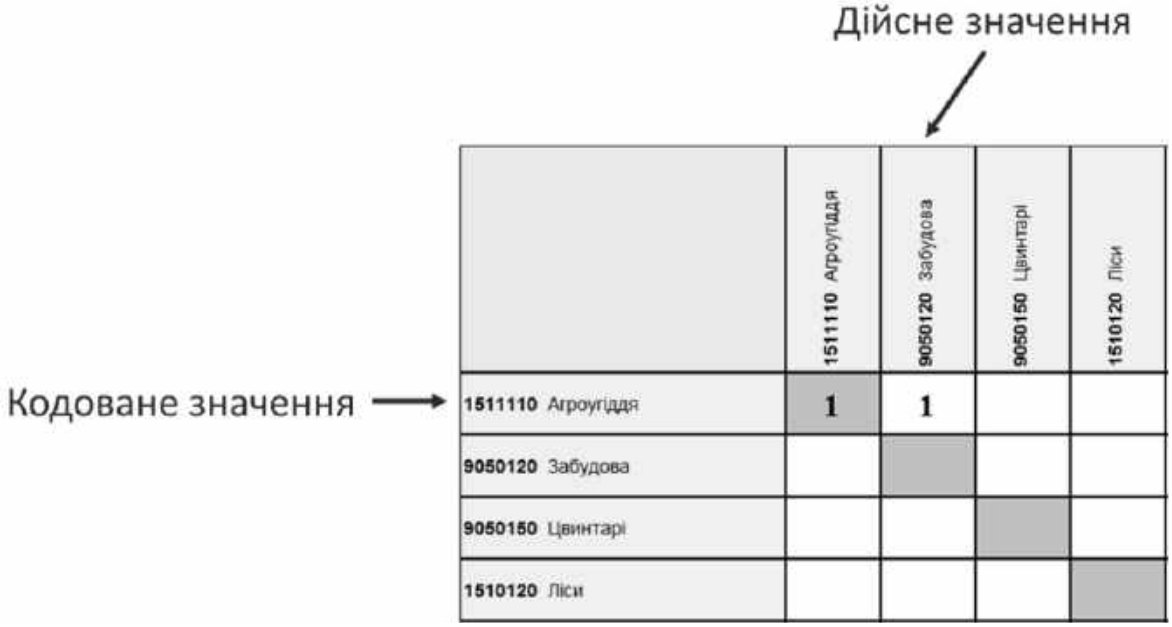


Рисунок 22 – Приклад заповнення матриці для другої контрольної точки

– Припустимо, що ми проаналізували третю та четверту точки й визначили, що їх правильно кодовано як *Агроугіддя* та *Забудову* відповідно. У цьому випадку ми відповідним чином додамо й значення у нашій матриці неточностей. Результат виглядатиме як на рисунку 23.

	1511110 Агроугіддя	**9050120** Забудова	**9050150** Цвинтарі	**1510120** Ліси
1511110 Агроугіддя	**2**	**1**		
9050120 Забудова		**1**		
9050150 Цвинтарі				
1510120 Ліси				

Рисунок 23 – Приклад заповнення матриці для третьої та четвертої контрольних точок

– Нарешті оновимо (розрахуємо) відповідні суми за рядками та стовпцями (у таблиці це "Сума за рядком" і "Сума за стовпцем"). Результат для нашої простої матриці неточностей за чотирма контрольними точками буде виглядати таким чином (таблиця 9).

Таблиця 9 – Матриця неточностей за чотирма контрольними точками

Класи земле-користу-вання	151111 Агроугіддя	9050120 Забудова	9050150 Цвинтарі	1510120 Ліси	1410120 Озера, ставки	1512000 Піски, торфовища, кар'єри	1410210 Водотоки ширші 30 м	1511120 Дачі, фруктові сади	1335200 Болота, водно-болотні угіддя	Сума за ряд-ком	Точність користувача
151111 Агроугіддя	2	1								3	
9050120 Забудова		1								1	
9050150 Цвинтарі										0	
1510120 Ліси										0	
1410120 Озера, ставки										0	
1512000 Піски, торфовища, кар'єри										0	
1410210 Водотоки, ширші 30 м										0	
1511120 Дачі, фруктові сади										0	
1335200 Болота, водно-болотні угіддя										0	
Сума за сто-впцем	2	2	0	0	0	0	0	0	0		
Точність виробника											

– Оновимо (розрахуємо) значення точності користувача та точності виробника даних. Значення (числа) за діагоналлю, які затінено сірим у нашому прикладі, є кількістю тих контрольних точок, які було кодовано правильно, тобто для них кодоване та дійсне значення землекористування є однаковими. Щоб обчислити кожне із згаданих вище значень точності, слід просто розділити кількість дійсних значень на загальну кількість точок у рядку або стовпці з певним кодуванням. Так, для отримання точності

користувача щодо *Агроугідь*, відображеній у правому верхньому кутку таблиці нижче, ділимо кількість правильних кодувань або дійсних значень (2) на загальну кількість точок у рядку *Агроугіддя* (3), отримуючи 67 % як точність користувача. Так сам для точності виробника щодо *Забудови* (див. значення відповідного стовпця наступної таблиці) ділимо кількість правильних кодувань або дійсних значень (1) на загальну суму за стовпцем *Забудова* (2), отримуючи 50 % як точність виробника. Такі результати обчислення точності записуємо в таблицю 10.

Таблиця 10 – Матриця зі значеннями точності виробника і користувача

Класи землекористування	151111 Агроугіддя	9050120 Забудова	9050150 Цвинтарі	1510120 Ліси	1410120 Озера, ставки	1512000 Піски, торфовища, кар'єри	1410210 Водотоки, ширші 30 м	1511120 Дачі, фруктові сади	1335200 Болота, водно-болотні угіддя	Сума за рядком	Точність користувача
151111 Агроугіддя	2	1								3	67 %
9050120 Забудова		1								1	100 %
9050150 Цвинтарі										0	0 %
1510120 Ліси										0	0 %
1410120 Озера, ставки										0	0 %
1512000 Піски, торфовища, кар'єри										0	0 %
1410210 Водотоки, ширші 30 м										0	0 %
1511120 Дачі, фруктові сади										0	0 %
1335200 Болота, водно-болотні угіддя										0	0 %
Сума за стовпцем	2	2	0	0	0	0	0	0	0		
Точність виробника	100 %	50 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %		

– Обчислимо загальну точність нашого набору даних. Загальна точність є загальною кількістю правильних кодувань (дійсних значень), розділеною на загальну кількість проаналізованих контрольних точок. Загальна кількість правильних кодувань є сумою значень комірок за діагоналлю (затінених сірим), тобто три. Загальна ж кількість проаналізованих контрольних точок є сумою значень за всіма стовпцями (за рядком "Сума за стовпцем") або за всіма рядками (за стовпцем "Сума за рядком"), тобто чотири. Звідси загальна точність набору даних є відношенням трьох до чьотирьох і становить 75 %.

Завершена матриця неточностей для наших чотирьох контрольних точок виглядатиме наступним чином (таблиця 11).

Таблиця 11 – Завершена матриця неточностей

Класи землекористування	151111 Агроугіддя	9050120 Забудова	9050150 Цвинтарі	1510120 Ліси	1410120 Озера, ставки	1512000 Піски, торфовища, кар'єри	1410210 Водотоки, ширші 30 м	1511120 Дачі, фруктові сади	1335200 Болота, водно-болотні угіддя	Сума за рядком	Точність користувача
151111 Агроугіддя	2	1								3	67 %
9050120 Забудова		1								1	100 %
9050150 Цвинтарі										0	0 %
1510120 Ліси										0	0 %
1410120 Озера, ставки										0	0 %
1512000 Піски, торфовища, кар'єри										0	0 %
1410210 Водотоки ширші 30 м										0	0 %
1511120 Дачі, фруктові сади										0	0 %
1335200 Болота, водно-болотні угіддя										0	0 %
Сума за стовпцем	2	2	0	0	0	0	0	0	0	4	
Точність виробника	100 %	50 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %		75 %

Примітки:

Якщо немає явного доказу, що кодоване значення є помилковим (наприклад, може бути складно відрізнити класи *Дачі, Фруктові сади* та *Забудова*), Вам слід припустити, що початково визначений клас землекористування є правильним (дійсним).

Однією з проблем, з якою можна стикнутися під час побудови матриці неточностей, є проблема “втрати Вашого місцезнаходження”, або проблема забування, які точки вже було проаналізовано, а які ще ні. Така проблема виникає найчастіше, коли Ви виконуєте цю вправу впродовж кількох днів на відміну від виконання її за одноразовий неперервний проміжок часу. Нижче надано кілька порад, які мають допомогти Вам виконати цю практичну роботу без пропуску контрольних точок або аналізу якихось із них двічі:

- Працюйте з Sample_Points послідовно ідентифікаторам атрибутів контрольних точок. За таких умов Ви зможете визначити, які точки було проаналізовано, оскільки сума за всіма стовпцями або рядками у матриці неточностей буде такою, що дорівнює ідентифікатору останньої проаналізованої точки. Це знизить імовірність помилок у процесі аналізу.
- Додайте атрибут до Sample_Points, до якого будете вносити певне значення після аналізу кожної точки. У поле цього атрибута може записувати, наприклад, “Y”, “Yes”, “Так” або “Готово” тощо, коли Ви проаналізували відповідну точку. Утім при цьому слід взяти до уваги, що заповнення поля зазначеного атрибута кожного разу після аналізу чергової точки збільшить час, потрібний на виконання цієї практичної роботи.
- Додайте два атрибути до Sample_Points, в яких Ви зберігатимете кодоване та дійсне значення землекористування для кожної з контрольних точок. Якщо Ви це зробите, Ви зможете швидко відновити матрицю неточностей у будь-який час за змістом цих двох атрибутів. Це найбільш досконалий та безпечний спосіб виконання аналізу контрольних точок, однак заповнення значень відповідних полів кожного разу після аналізу певної точки збільшить час, необхідний для виконання цієї практичної роботи.

Питання щодо матриці неточностей (до 10 балів):

Питання 1: Подайте матрицю неточностей, яку Ви створили за результатами аналізу 50 контрольних точок у Sample_Points.SHP. Це може бути таблиця, створена безпосередньо у текстовому файлі, або зовнішня електронна таблиця, вставлена до текстового документа як таблиця MS Word (до 4 балів).

Питання 2: Який клас (класи) землекористування було початково визначено оператором (виробником даних) найбільш неправильно? (до 1 бала).

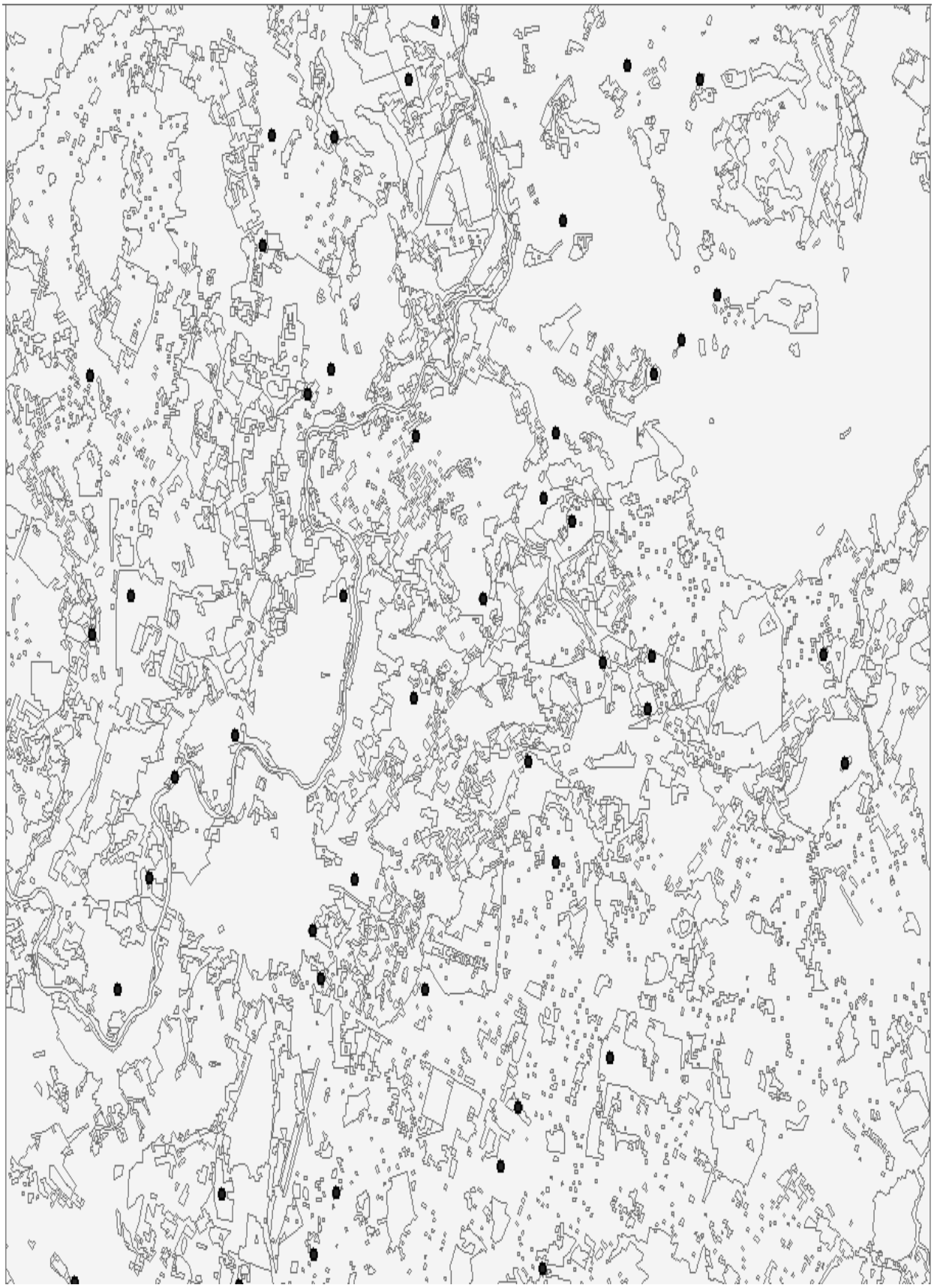
Питання 3: Який відсоток ділянок, що в дійсності є *Агроугіддями*, ми можемо оцінити як той, що кодовано правильно у наборі даних Landuse? (до 1 бала).

Питання 4: Визначіть типову помилку, яку Ви знайшли декілька разів, або під час аналізу 50 контрольних точок, або як просто щось таке, що Ви помітили, пересуваючись за даними (наприклад, озера помилково кодовано як болота). Поясніть, чому може трапитись така помилка і які зміни Ви зробили б у процесі створення даних, що могло б знизити ймовірність виникнення такої помилки в майбутньому (до 4 балів).

Ці методичні рекомендації створено на основі даних навчального курсу «Питання менеджменту в ІПД», наведеного *Vancouver Island University* з використанням програмного продукту *ESRI ArcGIS 10.3.1.*

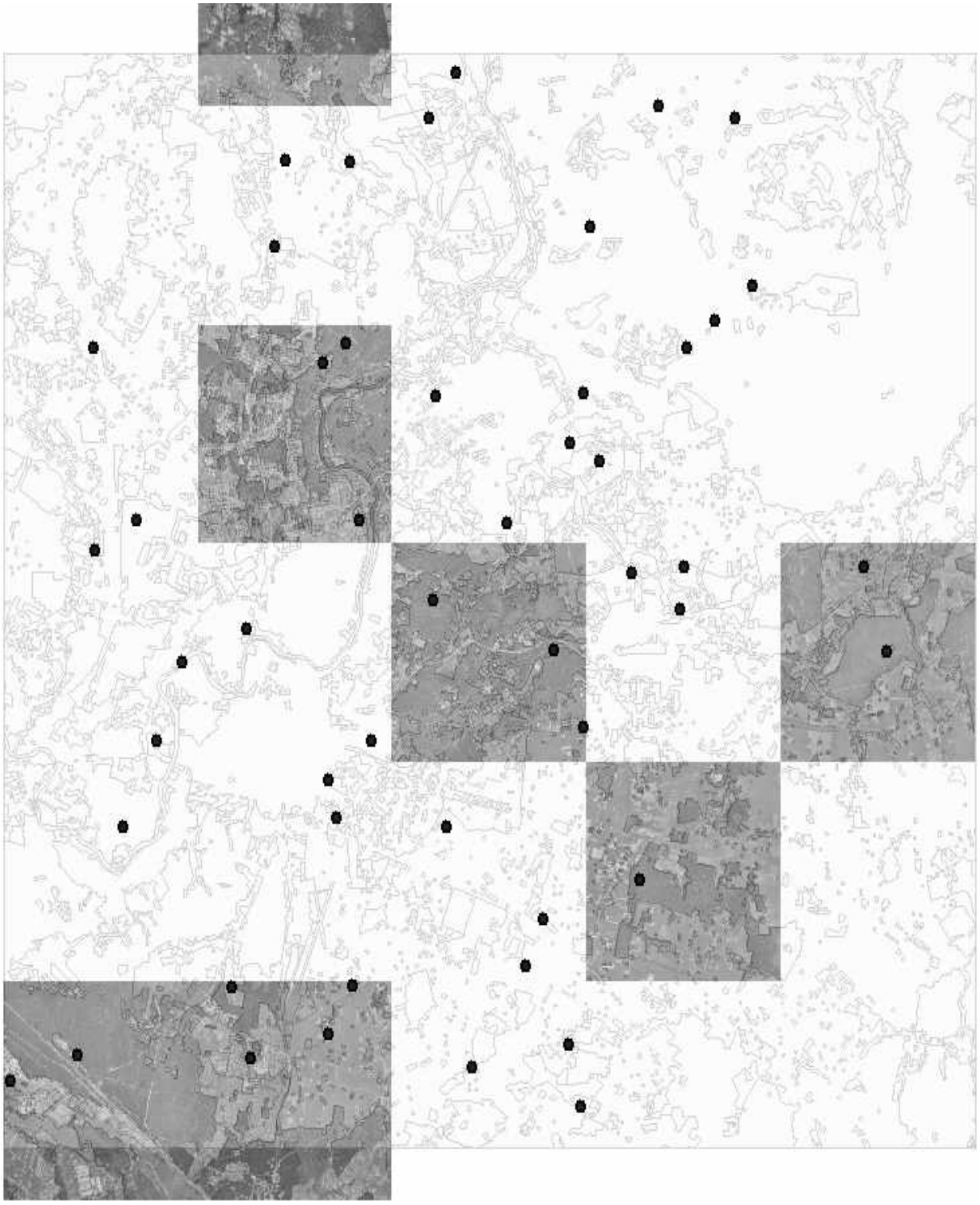
Додаток 1

Зображення файлів Sample_Points.SHP і Landuse.SHP



Додаток 2

Зображення файлів Sample_Points.SHP і Landuse.SHP з додаванням ортофотознімка



БІБЛІОГРАФІЧНИЙ СПИСОК

1. Карпінський, Ю. О. Концептуальні засади оцінювання та забезпечення якості геопросторових даних Ю. О. Карпінський // Вісник геодезії та картографії. – 2012. – № 4. – с. 33 – 41.

2. Геоинформатика: учеб. пособие для студентов вузов / Е. Г. Капралов, А. В. Кошкарев, В. С. Тикунов и др. — М. : Издательский центр «Академия», 2005. — 480 с.

3. Лурье, И. К. Геоинформационное картографирование. Методы геоинформатики и цифровой обработки космических снимков: учебник / И. К. Лурье. – М. : КДУ, 2008. – 424 с.

4. ISO 19157: Geographic information — Data quality. – International Organization for Standardization, 2013.

5. *http://www. eurogeographics.org/documents/ Guidelines_ ISO 19100_Quality.pdf.*

6. ISO 19157: Geographic information — Data quality. – International Organization for Standardization, 2013.

7. Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies. EuroGeographics Expert Group on Quality. Edited by Antti Jakobsson, Jørgen Giversen. EuroGeographics – 2007. – 68 pp. – http://www. eurogeographics.org/documents/ Guidelines_ ISO 19100_Quality.pdf.

Зміст

Навчальне видання

Красовська Інеса Григорівна

Ковальова Віра Олександрівна

УПРАВЛІННЯ ЯКІСТЮ ДАНИХ ГІС

Редактор С. П. Гевло

Зв. план, 2019
Підписано до друку 29.05.2019
Формат 60×84 1/16. Папір офс. № 2. Офс. друк
Ум. друк. арк. 3,6. Обл.-вид. арк. 4. Наклад 75 пр.
Замовлення 172. Ціна вільна

Видавець і виготовлювач
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»
61070, Харків-70, вул. Чкалова, 17
http://www.khai.edu
Видавничий центр «ХАІ»
61070, Харків-70, вул. Чкалова, 17
Izdat@khai.edu

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготовлювачів і розповсюджувачів
видавничої продукції сер. ДК № 391 від 30.03.2001