Ученые записки Таврического национального университета имени В.И.Вернадского
Серия «География». Том 25 (64). 2012 г. № 1, С. 129-135.

УДК 004.942

МОДЕЛЮВАННЯ ПРОСТОРОВО РОЗПОДІЛЕНИХ ДИНАМІЧНИХ СИСТЕМ ІЗ ЗАСТОСУВАННЯМ ГЕОІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ESRI

Ковгар В. Б.[1], Філозоф Р.С.[2]

[1]ПрАТ «ЕСОММ» Со, Київ, Україна
[2]Київський національний університет імені Тараса Шевченка, Київ, Україна

Наведено огляд методів моделювання просторово розподілених динамічних систем для вирішення проблеми «Big Data». Обґрунтовано застосування геоінформаційної технології Esri для підвищення ефективності роботи з великими обсягами накопичень зазвичай просторово розподілених даних. Розглянуто способи опрацювання значних масивів просторово розподілених даних. Запропоновано методику вирішення поставленої задачі. Наведено перелік предметних областей, в яких можливе застосування даної методики.
Ключові слова: моделювання, статистичні дані, динамічні системи, просторово розподілені дані, ГІС.

ВСТУП

Оперування значними масивами інформації є невід'ємною ознакою сучасності і одним із факторів формування інформаційної кризи. Перед суспільством постає проблема ефективного зберігання та управління даними, їх оптимального використання. В ужиток досить швидко увійшли такі поняття, як Big Data та Data Mining і проблеми, пов'язані із цими напрямками досліджень обговорюються все ширшим колом користувачів корпоративних інформаційних систем, зокрема – проблема видобутку корисних відомостей аналітичного характеру з наявного корпоративного інформаційного ресурсу. З огляду на таку ситуацію надзвичайно **актуальною** є необхідність розробки ефективного інструменту роботи з масивами просторово-розподілених даних, що накопичуються із плином часу. **Метою** даної роботи є розробка методичної схеми моделювання просторово розподілених динамічних систем на основі аналізу наукових методів, що застосовуються у комп'ютерних, математичних та географічних науках, та синтезу відповідних інформаційних технологій, що ґрунтуються на цих методах.

ВИКЛАДЕННЯ ОСНОВНОГО МАТЕРІАЛУ

Сьогодні ми є свідками активного розвитку технологій автоматизації інтелектуального аналізу даних, поява яких пов'язана насамперед з необхідністю аналітичної обробки надвеликих об'ємів даних, накопичуваних в інформаційних сховищах даних (Data Warehousing). Це зумовлено головним чином потоком нових ідей, які витікають із сфери комп'ютерних наук, що утворилася на перетині штучного інтелекту, статистики та теорії баз даних. Дану область позначають як

KDD (Knowledge Discovery in Databases – виявлення знань в базах даних). Нині відбувається зростання кількості програмних продуктів, в яких застосовані технології KDD, а також типів задач, де використання даних технологій дає вагомий економічний ефект. Елементи автоматичної обробки та аналізу даних стають невід'ємною частиною концепції електронних сховищ даних. Основним кроком KDD щодо опрацювання вмісту електронних сховищ даних є Data Mining (дослідження даних і, як наслідок, видобуток додаткових відомостей). Інтелектуальний аналіз даних і KDD вважають, в загальному випадку, синонімами Data Mining, але фактично Data Mining є основним, та не єдиним елементом в множині KDD.

За визначенням Григорія Піатецького-Шапіро, одного із засновників цього напряму, Data Mining — це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності [1, 2]. Цей процес включає три основних етапи: дослідження, побудову моделі або структури та її перевірку. В його основі лежить статистичний аналіз, що використовувався до цих пір в ролі практичного інструмента, а також такий, що приваблював математиків-теоретиків. Але до недавнього часу процес видобутку «прихованих» відомостей аналітичного характеру був достатньо довготривалим, проводився вручну. Точність цього процесу істотно залежала від того, хто його виконував. Нині з'явилися засоби, які можуть автоматизувати цей процес, що дає можливість використовувати видобуток відомостей аналітичного характеру широкому колу фахівців – користувачів цих засобів автоматизації.

Інформаційна технологія Data Mining використовує складний статистичний аналіз і методи моделювання для знаходження відношень (кореляційних залежностей або моделей), захованих у Big Data (корпоративному банку даних) – таких моделей, які не можуть бути знайдені звичайними методами. Отже, методи видобутку «прихованих» відомостей аналітичного характеру набувають все більшої популярності в ролі інструменту для аналізу різноманітних даних, особливо в тих випадках, коли передбачається, що із наявних даних можна буде витягнути знання (відомості) для прийняття рішень в умовах невизначеності. З іншого боку, запорукою успішного застосування цих методів є не просто вибір алгоритму, а майстерність людини, яка проводить конструювання моделі та можливості програми проводити саме процес моделювання. Тобто ми підходимо до проблеми створення автоматизованої аналітичної системи, яка керуючись певним алгоритмом, застосовуючи нові ефективні методи здатна моделювати складні процеси. Слід детальніше зупинитись на тому, які саме методи можуть бути використані в подібному моделюванні, які дані мають використовуватись, яким є алгоритм побудови та роботи подібної системи і де її можна застосувати.

Інтелектуальні засоби аналізу даних використовують наступні основні методи:

- нейронні мережі;
- дерева рішень;

- індукцію правил.

Крім цих головних методів існують ще декілька допоміжних:

- системи міркування на основі аналогічних випадків (прецедентів);
- нечітка логіка;
- генетичні алгоритми;
- алгоритми встановлення асоціацій і послідовностей;
- аналіз із виборчою дією;
- логічна регресія;
- еволюційне програмування;
- візуалізація даних.

В складних аналітичних системах найчастіше застосовують комбінацію перерахованих методів. Тим не менше, основним методом сучасної математичної статистики по праву можна назвати регресійний аналіз [3-6]. Типова процедура регресійного аналізу випливає із передумови, що всі необхідні дані для побудови математичної моделі вже зібрані.

Для будь-яких задач із кількісними змінними інтерес становить дослідження впливу (дійсного чи підозрюваного) одних змінних на інші. Таким впливом, зазвичай, може бути простий функціональний зв’язок між змінними; проте у багатьох фізичних процесах це швидше виняток, ніж правило. Часто, швидше за все, існує функціональний зв’язок, що є занадто складним для розуміння чи для опису в простих термінах. У такому випадку можуть прагнути підібрати апроксимацію цього функціонального зв’язку за допомогою якої-небудь простої математичної функції (скажімо, такої, як поліном), яка включає відповідні змінні, і згладжувати «істинну» функцію в певній обмеженій області зміни цих змінних. Досліджуючи таку згладжену функцію, більше дізнаються про розглядувану «істинну» залежність та оцінюють окремі чи сукупні ефекти зміни деяких важливих змінних.

Навіть тоді, коли за змістом не існує фізичного зв’язку між змінними, ми можемо прагнути відобразити його за допомогою математичного рівняння даного виду. Якщо рівняння фізично не має сенсу, то воно тим не менше може виявитися достатньо цінним для передбачення значень ряду змінних за невідомими значеннями інших змінних, можливо, за певних обмежень. Саме для досліджень такого роду послуговуються апаратом регресійного аналізу.

Будуючи функціональну залежність, розрізняють два основних типи змінних. Перший тип називають предикторами, або незалежними змінними (факторами, сигналами) і другий – залежними змінними, або змінними-відгуками. Під предикторами, або факторами розуміють такі змінні, для яких, як правило, можна встановити бажані значення, або ті, що їх можна лише спостерігати, але не управляти ними. В результаті навмисних змін, чи змін, що сталися із незалежними змінними випадково, з’являється ефект, який передається на інші змінні, на відгуки. Тобто інтерес становить те, як зміни предикторів впливають на значення відгуків.

В залежності від явища або процесу, що моделюється, та відповідно від пре дикторів (факторів впливу на шукані змінні) може бути застосований лінійний або нелінійний регресійний аналіз. Однією з переваг застосування саме регресійного аналізу в пропонованій методиці є те, що експлораторна регресія вже вбудована в інструментальну платформу Esri ArcGIS, яка починаючі з 10-ї версії підтримує темпоральну складову даних, що зберігаються.

Зазначимо одразу, що і візуалізація даних не даремно вказана серед аналітичних методів. Не зважаючи на те, що даний метод не має математичної основи, він, все ж, відіграє надзвичайно важливу роль у зв'язці «аналітична система – оператор», особливо у випадках із надвеликими об'ємами даних (Big Data). Можливість обробити значні масиви даних, обрахувати залежні змінні та подати їх у зручному для розуміння і прийняття рішення вигляді – основна функція описуваної аналітичної системи. Форма візуалізації даних буде описана далі, а поки розглянемо самі дані. Зокрема, якими вони мають бути, що може слугувати джерелом таких даних, як вони мають оброблятися перед їх візуалізацією?

Як випливає з назви даної статті, ми акцентуємо увагу по-перше на моделюванні динамічних системам, а по-друге таких із них, що мають просторово розподілений характер. Такі системи мають подвійну природу – вони можуть еволюціонувати в часі та в просторі. В цьому випадку обчислення залежних змінних може виконуватись за двома критеріями: визначення кількісної зміни показника, що має просторову прив'язку та зміни місцеположення (або конфігурації) просторового об'єкта. Моделювання просторово розподілених динамічних систем може виконуватись на базі одного або декількох основних підходів видобутку відомостей аналітично характеру, а саме:

- класифікація;
- регресія;
- прогнозування часових послідовностей (рядів);
- кластеризація;
- асоціація;
- послідовність.

Перші три використовуються, головним чином, для передбачення, в той час, як останні зручніші для опису існуючих закономірностей у статистичних вибірках даних.

Джерелами даних для подібного моделювання в ідеальному випадку є значні масиви статистичних даних спостереження за станом певних об'єктів дослідження (дані про забруднення, температури) або за просторово розподіленими фізичними процесами (селеві потоки, атмосферні фронти тощо). При цьому, чим більше факторів впливу на залежні змінні буде виявлено, тим точнішою буде модель оцінки їх динаміки. Тобто в даному випадку наявність надмірних даних перетворюється із недоліку в перевагу при їх обробці за пропонованою методикою. В такому ідеальному випадку етап попередньої підготовки даних значно спрощується за рахунок того, що відпадає необхідність геокодування – дані вже мають просторову

прив’язку. Втім, сфера застосування даної методики може бути значно розширена завдяки тому, що значна кількість статистичних даних може бути геокодована на етапі підготовки даних, не будучи від початку просторово прив’язана. В такому разі об’єктами моделювання можуть виступати динамічні системи, що описуються даними спостереження соціальних, економічних, природних процесів [7].

Просторово розподілений характер даних передбачає і особливий підхід до їх зберігання та структуризації. Не залежно від того, чи були дані просторово розподіленими від початку, чи – були прив’язані до певних просторових об’єктів (кластерів), окрім характеристик, що визначаються статистичними даними такі об’єкти мають просторові характеристики. Тобто на даному етапі виникає потреба застосування геоінформаційної технології. Адже формування бази даних, що містить такі об’єкти з їх характеристиками відбувається за правилами створення геобаз даних з відповідною структурою. З огляду на це, корисним є використання в даній методиці інструментів, що запропоновані геоінформаційною технологією Esri – ArcGIS [8].

Окрім можливості формування геобази даних шляхом інтегрування різнорідних даних із різних джерел, ArcGIS надає широкі можливості з просторового моделювання та візуалізації отриманих результатів. В пропонованій методиці ArcGIS є інструментальним середовищем, в якому інтегруються математична та географічна компоненти моделювання.

Візуалізація відіграє чималу роль у тому наскільки швидким та ефективним буде процес прийняття рішення, що ґрунтується на основі адекватної просторово часової моделі. Перевагою застосування геоінформаційної технології Esri є можливість картографічного відображення динаміки просторово розподілених систем як в просторі, так і в часі. Тобто, за рахунок створення часових (темпоральних) класів просторових об’єктів можливість технологія передбачає відображення динаміки подій у часовому вимірі з визначеним періодом часу. В поєднанні з методами регресійного аналізу та прогнозування часових послідовностей такий інструмент стає потужним засобом візуалізації динаміки досліджуваних явищ або окремих величин.

Отже, в загальному випадку пропонована методична схема моделювання просторово розподілених динамічних систем (Рис. 1) складається з таких етапів:

- виокремлення залежних і незалежних змінних та збір статистичних даних;
- первинна обробка даних: формування статистичної бази даних, за необхідності просторова прив’язка (геокодування);
- визначення ступеню кореляції між змінними, виділення ключових факторів, застосування регресійного аналізу;
- побудова моделі та обрахунок значень залежних змінних;
- картографічна візуалізація результатів моделювання.

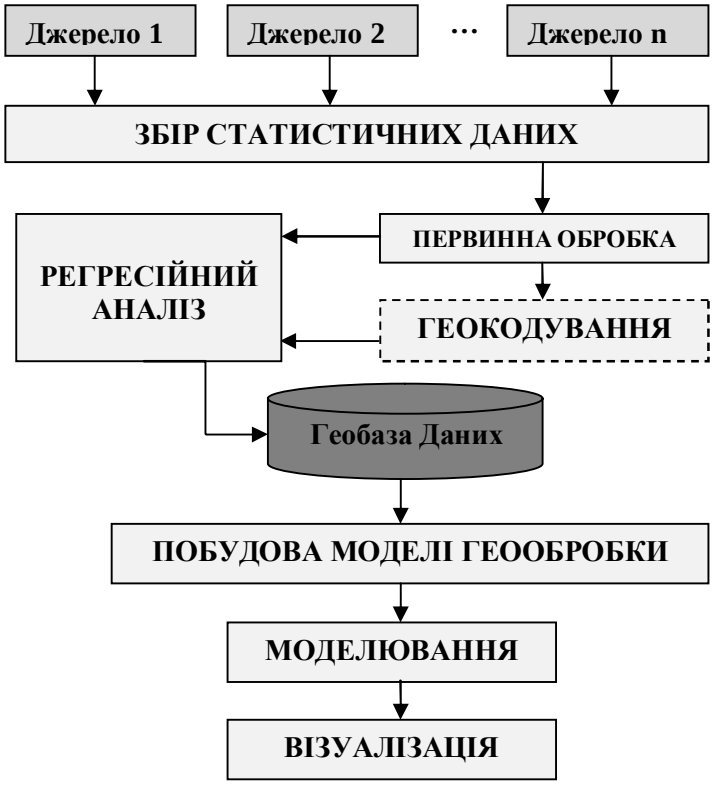

Рис. 1. Загальна методична схема моделювання просторово розподілених динамічних систем

ВИСНОВКИ

Застосування методів регресійного аналізу при моделюванні динамічних систем, сформованих з надвеликих об'ємів статистичних даних дозволяє уникнути проблеми Big Data, оцінити кореляцію даних (в тому числі і приховану), виокремити найважливіші фактори та обрахувати прогнозні значення шуканої величини за умов, закладених в моделі. Робота з даними стає ефективнішою, коли є можливою інтеграція наступних компонентів: картографічна візуалізація, графічний інструментарій, засоби формування запитів, оперативна аналітична обробка, які дозволяють зрозуміти дані та інтерпретувати результати моделювання і, нарешті, самі алгоритми, які будують моделі. Отже, просторова та темпоральна динаміка модельованих систем має бути візуалізована в зручній для сприйняття формі (зокрема – картографічній). З огляду на геопросторовий характер даних просторово розподілених динамічних систем, необхідним є формування геобази даних. Для цієї мети та для візуалізації найкращим чином підходить геоінформаційна технологія Esri, яка інтегрує в собі весь інструментарій, необхідний для ефективної роботи з такими даними. Запропонована методика за умов належної підготовки даних (та

геокодування, в разі необхідності) може бути застосована в найрізноманітніших предметних областях: в торгівлі, сфері фінансів, банківській справі, сфері телекомунікацій, медицині, демографії, різних галузях економіки, для оцінки екологічного стану навколишнього середовища, для моделювання природних явищ тощо. Перспективи застосування даної методики вбачаються досить широкими з огляду на накопичення значної кількості даних в корпоративних інформаційних системах (Big Data) та найрізноманітніших областях досліджень.

Список літератури

1. Data Mining – интеллектуальный анализ данных / Информационные Технологии. [Электронный ресурс] – Режим доступа – http://www.inftech.webservis.ru/it/database/datamining/ar2.html. – 10.04.2012.
2. Дюк В.А. Data Mining / В.А. Дюк, А.П. Самойленко – Санкт-Петербург: Изд-во «Питер», 2001. – 368 с.
3. Демиденко Е.З. Линейная и нелинейная регрессии. / Е.З. Демиденко – М.: Финансы и статистика, 1981. – 302 с.
4. Дрейпер Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит // Пер. с англ – В 2-х кн. Кн. 1 – М.: Финансы и статистика, 1986. – 366 с.
5. Ивахненко А.Г. Долгосрочное прогнозирование и управление сложными системами. – К.: Техника, 1975. – 312 с.
6. Ивахненко А.Г. Моделирование сложных систем по экспериментальным данным. / А.Г. Ивахненко, Ю.П. Юрачковский – М.: Радио и связь, 1987. – 120 с.
7. Кравченко. Ю.А. Информационное геомоделирование: модели и методы: [монография] / Ю.А. Кравченко – Новосибирск: СГТА, 2008. – Книга 2, Часть 2 – 316 с.
8. Цейлер М. Моделирование нашего мира: пособие Esri® по проектированию баз геоданных : Пер. с англ. / М. Цейлер. – К. : ECOMM, 2003. – 254 с.

Ковгар В.Б. Моделирование пространственно распределенных динамических систем с применением геоинформационной технологии Esri / В.Б. Ковгар, Р.С. Филозоф // Ученые записки Таврического национального университета имени В.И. Вернадского. Серия: География. – 2012. – Т. 25 (64). – № 1− С.129-135.

Приведен обзор существующих методов моделирования пространственно распределенных динамических систем. Обосновано их применение для повышения эффективности работы с пространственно распределенными данными. Рассмотрены способы накопления и хранения значительных массивов пространственно распределенных данных. Предложена методика решения поставленной задачи. Приведен перечень предметных областей, в которых возможно применение данной методики.

Ключевые слова: моделирование, статистические данные, динамические системы, пространственно распределенные данные, ГИС.

Kovgar V.B. Modeling of spatially distributed dynamic systems using GIS-technology Esri / V.B. Kovgar, R.S. Filozof // Scientific Notes of Taurida National V. I. Vernadsky University. – Series: Geography. – 2012. – Vol. 25 (64). – № 1 − P. 129-135.

Provides an overview of existing methods for modeling spatially extended dynamical systems. This justified their use to improve performance with spatially distributed data. It provides an overview of the methods of storage of large arrays of spatially distributed data. The method of solving this problem is proposed. A list of subject areas, which may use this technique, is given.

Keywords: modeling, statistics, dynamical systems, spatially distributed data, GIS.

Поступила в редакцию 18.04.2012 г.