Бідюк П.І., Терентьєв О.М., Просянкіна-Жарова Т.І., Савастьянов В.В.
ННК «Інститут прикладного системного аналізу» Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

Застосування інструментів SAS BASE для дослідження ефективності методів обробки пропусків у вибірках даних з метою підвищення якості прогнозування показників продовольчої безпеки країни

Вступ. В умовах суспільно-політичних перетворень проблема забезпечення населення продовольством в достатній кількості і за доступною ціною є не лише економічним чинником, а й передумовою соціальної стабільності. Саме тому у даному дослідженні, виконаному в рамках проекту NUKR.SFPP G4877 «Modeling and Mitigation of Social Disasters Caused by Catastrophes and Terrorism» [1], як одні із пріоритетних, розглядаються питання розробки та впровадження ефективних інструментів аналізу та прогнозування кількісних параметрів продовольчої безпеки. Однією із практично значущих проблем при розв'язуванні таких задач є проблема розкриття невизначеностей різного роду, зокрема інформаційної, зумовленої пропусками даних.

Мета дослідження. Аналіз існуючих методів заповнення пропусків та розроблення рекомендацій щодо вибору оптимального підходу до оброблення неповних вибірок даних з метою підвищення якості прогнозного моделювання показників продовольчої безпеки.

Викладення основного матеріалу. У роботах багатьох вітчизняних та закордонних вчених, пов'язаних із прогнозним моделюванням соціально-економічних процесів, значна увага приділена питанням, пов'язаним із заповненням пропусків у відповідних часових рядах. Очевидно, що вибір методики залежить від типу змінних, що містять пропуски, особливостей предметної області дослідження, виду аналізу, що в подальшому буде використаний для побудови прогнозних моделей [2]. Враховуючи специфіку часових рядів, що описують досліджувані процеси, для заповнення пропусків доцільно використовувати адаптивні моделі з трендом та сезонною компонентою, зокрема, моделі із затухаючим трендом, адитивні та мультиплікативні моделі Хольта-Вінтерса, авторегресійні моделі тощо. Зазначимо, що методи, які використовують заміну пропусків сталими значеннями є взагалі неприйнятними. Саме тому, особливе значення має ефективна організація робіт з накопичення необхідної інформації, попередньої обробки даних для їх подальшого опрацювання у системі підтримки прийняття рішень (рис.1).

Рис. 1. Послідовність обробки інформації на етапі попереднього аналізу даних для прогнозування продовольчої безпеки

Не менш складною проблемою є вибір комп'ютерної аналітичної системи, яка б дозволяла реалізувати якнайширший спектр методів дослідження часових рядів та інтелектуального аналізу даних. Тому у даному дослідженні використано середовище SAS Base [3], зокрема, модулі SAS/ETS (Econometrics and Time Series Analysis) та SAS/STAT (Statistical analysis).

В рамках дослідження було виконано декілька обчислювальних експериментів із використанням вибірок різних статистичних даних, що описують продовольчу безпеку. Зокрема, були досліджені часові ряди показників, вимірюваних у натуральних одиницях (дані щодо валового збору основних сільськогосподарських культур в Україні за період 1940-2014 рр.) та у грошових одиницях (дані про виробництво валового регіонального продукту однієї з

областей України за 2003-2014 рр.) [4]. Досліджувана вибірка даних щодо валового збору основних сільськогосподарських культур України за 1940-2014 рр. у ряду «фрукти та ягоди» містить пропуски даних за окремі роки, додатково 10 відсотків вхідних даних були заповнені пропусками у довільному порядку. Заміна пропусків даних здійснювалась на: середнє, медіанне значення, попереднє заповнене значення, середнє сусідніх значень, значення, отримане в результаті побудови моделі з трендом. Крім того, запропоновано власний метод, який передбачає заповнення пропусків усередненим середнім, розрахованим на основі попередніх п'яти відомих значень ряду із використанням вагових коефіцієнтів. Вагові коефіцієнти розраховувалися за тим же принципом, що використовується у методі експоненційного згладжування. Прогнози найкращої якості одержані за методом заміни пропусків на попереднє заповнене значення (MSE=172,94) та за методом заміни пропусків на усереднене за попередні п'ять років значення (MSE=160,55).

Враховуючи особливості формування вибірок даних фінансово-економічних показників продовольчої безпеки, а саме, неможливість одержання досить довгого часового ряду співставних статистичних даних за наявності значної кількості показників (використано 26 показників за 2003-2014 рр. на мезорівні та 32 показники за 2003-2013 рр. на макрорівні), для відновлення пропущених значень були застосовані методи імпутування із використанням регресійних моделей, дерев рішень, зокрема методу Кааса і методу Stepwise. Для прогнозування показників продовольчої безпеки на мезорівні побудовані такі моделі: нейронна мережа зі структурою багатошарового персептрона з трьома прихованими вершинами, з критерієм оптимізації моделі – мінімізація середньої похибки моделі (MSE – 9589,711, SSR – 86307,4); нейронна мережа із структурою багатошарового персептрона з трьома прихованими вершинами, використанням кластерного аналізу та дерев рішень і критерієм оптимізації – мінімум середньої похибки моделі нейронної мережі (AIC – 114,6902, MSE – 112672,7, BSC – 115,676, SSR – 1014054); регресійна модель, у якій кластери, побудовані для попередньої моделі, використано у якості вхідних змінних (AIC – 23,67, MSE – 31,925, BSC – 31,03, SSR – 287,322); пропущені значення часового ряду в окремих кластерах доповнені за методом дерев рішень. Для прогнозних моделей, побудованих для макрорівня (відсутні значення заповнювалися у часових рядах чотирьох показників із 32, що використані у прогнозному моделюванні), прийнятні результати заповнення пропусків у всіх випадках одержані з використанням регресійних моделей, значення критеріїв якості знаходяться в межах: AIC: 5,67-20,68, MSE: 0,13-0,91, BSC: 5,71-20,73 (відповідно). Для перевірки якості прогнозування використано порівняння прогнозних показників із відповідними статистичними показниками за 2012-2014 рр. Розбіжність значень реальних та прогнозованих показників складає від 3 до 5%, що вказує на високу якість прогнозу.

Висновки. Як показало виконане дослідження, для заповнення пропусків цілком прийнятні результати можна одержати використовуючи не лише ітеративні алгоритми, а й навіть такі методи, як просте усереднення (для випадків, коли відсутня лише незначна частина спостережень), регресійне моделювання, заповнення пропусків шляхом генерації прогнозованих значень на основі наявних, генерації відсутніх (втрачених) даних на основі форми розподілу та параметрів, що визначені використовуючи наявну частину даних, оптимізаційні методи, зокрема EM-алгоритми, експоненційне згладжування, тощо. Однак, найкращі результати можна одержати лише за умови якісної попередньої обробки вхідної інформації.

Література. **1.** Офіційний сайт «Україна-НАТО». (2016) Програма НАТО «Наука заради миру та безпеки» [Електронний ресурс]. Режим доступу: http://ukraine-nato.mfa.gov.ua/ua/about-nato/science-for-peace. **2.** N. V. Kusnietsova and P. I. Bidyuk, "Business intelligence techniques for missing data imputation", Наукові вісті НТУУ «КПІ», No. 5, 2015, pp. 47–56. **3.** Multiple imputation in SAS / Institute for digital research and education. (2016) [Електронний ресурс]. Режим доступу: http://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1. **4.** Офіційний сайт Державної служби статистики України. (2016) Комплексні статистичні публікації [Електронний ресурс]. Режим доступу: http://www.ukrstat.gov.ua.