МЕТОДИ ІМПУТАЦІЇ ПРОПУЩЕНИХ ДАНИХ

І.В. РОЗОРА, С.О. ОЛАСЮК, А.О. МЕЛЬНИК

В сучасному світі перед нами постає багато задач, що вимагають обробки і дослідження масивів даних. Поширеною проблемою, яка виникає при роботі з ними є відсутність частини значень. Задачі обробки пропущених значень присвячено багато наукових робіт, одні з найбільш популярних методів, які розглянуті в даній роботі: метод середніх, метод k-найближчих сусідів, метод гарячого набору.

Найпростішим підходом, очевидно, є метод середніх, але заміна всіх відсутніх значень середнім може викривити дисперсію та коваріацію. Метод k-найближчих сусідів, натомість, індентифікує k найбільш схожих значень з масиву та заповнює порожні значення шляхом усереднення "сусідніх хоча такий спосіб не підходить якщо відсутніх значень забагато. Існують декілька варіантів методу гарячого набору, їх суть полягає в знаходженні патернів і заміни відсутнього значення з групи даних на значення зі співставного набору, перевагою способу є збереження внутрішньої структури даних.

В даній роботі застосовано три вищеописані підходи до одного набору даних, CAC 40 Index, з якого випадковим чином вилучено 5 відсотків даних. Вірогідність видалення однакова для кожного значення масиву, з масиву вилучаються значення з індексами, отриманими за допомогою генератора випадкових чисел. Для роботи з даними використано програмне середовище R.

Література

[1] 5. Little R., Rubin D. *Statistical Analysis with Missing Data (3rd Edition)*. 2020 by John Wiley and Sons, Inc

[2] Stef van Buuren. *Flexible Imputation of Missing Data (1st Edition)*. 2012 by Taylor and Francis Group, LLC

Київський національний університет ім. Т. Шевченка, Київ, Україна
Email address: irozora@knu.ua

НТУУ "КПІ ім. І. Сікорського Київ, Україна
Email address: nameisneedfull@gmail.ua

Київський національний університет ім. Т. Шевченка, Київ, Україна
Email address: melinik2011@gmail.com