ВИКОРИСТАННЯ СУЧАСНИХ СТАТИСТИЧНИХ ПРОГРАМ R, SPSS І SAS ДЛЯ ОБРОБКИ МАСИВІВ ДАНИХ

Бойко Я.А.

IV курс, фізико-математичний факультет

Терентьєв Олександр Миколайович, канд.тех.наук, старший викладач

Уманський державний педагогічний університет імені Павла Тичини

Умань

До найбільш вживаних засобів аналізу даних належить проект R, SPSS та SAS. Мова програмування R була розроблена співробітниками статистичного факультету Оклендського університету Россом Айхекой і Робертом Джентлменом. Середовище R - це безкоштовний пакет для обробки даних з відкритим кодом, який конкурує безпосередньо з комерційними аналітичними інструментами, а також доповнює їх [4].

SAS призначений для забезпечення універсального доступу до даних і забезпечує дружній інтерфейс. Продукти SAS, широко відомі як модулі, в основному використовуються в сфері аналізу соціальних явищ. Ці модулі дозволяють їм виконувати різні типи функцій, як аналіз електронних таблиць, доступ до даних, статистичний аналіз, побудова додатків та управління. Продукти SAS можна придбати окремо або в комплекті. Рішення SAS пропонують ряд методів і процесів для осмисленого прийняттям рішень [2, с.1-2].

Всі комерційні аналітичні засоби містять графічний інтерфейс користувача. З часом акцент змінився від кодування до використання платних послуг. До переваг графічного інтерфейсу користувача відносимо надійність та безпомилковість, що дозволяє використовувати аналітичні можливості із швидкістю, що, як правило, дорівнює чи перевищує кодування. Графічний інтерфейс користувача допомагає фахівцям бути більш ефективними, а також скористатись заощадженим часом для концентрації на самих методах аналізу замість використання його на написання коду. Сильна і слабка сторони співіснують в інтерфейсі користувача. З одного боку, легко генерувати код за

допомогою інтерфейсу користувача, а з іншого – така легкість тісно пов’язана з формуванням недосконалого коду. Це передусім характерно для недосвідчених користувачів, що в кінцевому рахунку призводить до результату, який повністю відрізняється від запланованого.

Дані, які отримані від аналізу, повинні бути презентованими у формі, якою може скористатися замовник. Засоби візуалізації дають можливість професіоналам створити інтерактивну, візуальну аналітику. Експерт має пояснити складні аналітичні результати замовнику у бізнес сфері. Для багатьох людей більш зрозумілішим є візуальне представлення дерева рішень, ніж довгий список бізнес правил і настанов.

Отже, аналітичні інструменти R, SAS і SPSS відрізняються один від одного призначенням і можливостями, а тому їх необхідно дослідити відповідно певних критеріїв.

До першого критерію відносимо інтерфейс користувача. SAS має найбільш інтерактивний та зручний інтерфейс. За ним слідує SPSS, який містить досить інтерактивний графічний інтерфейс. Інструмент R має найменш інтерактивний аналітичний інструмент, але для даного програмного забезпечення можна використовувати редактори, які підтримують графічний інтерфейс і програмування в R. Проте для вивчення та практики середовище R є відмінним інструментом, оскільки це дійсно допомагає програмістам опанувати різні етапи та команди аналітики.

Другий критерій тісно пов’язаний із прийняттям рішень. Статистика IBM SPSS має певну перевагу над SAS не тільки за нижчою ціною, але також і за можливістю отримати Answer tree (Дерева класифікації) для дерева прийняття рішень без необхідності придбання пакету Data Mining. Для побудови дерева прийняття рішень в SAS необхідно придбати Enterprise Miner. В IBM SPSS дерева прийняття рішень також відзначаються більшою конкурентоспроможністю, ніж в середовищі R, оскільки тут не пропонується багато алгоритмів дерев. Більшість пакетів використовують лише CART, а їх інтерфейс не є зручним для користувача.

Третім чинником, що відіграє особливо важливу роль у виборі інструменту аналітичної обробки даних, є управління даними. У керуванні даними інструмент SAS має перевагу над IBM SPSS і дещо кращий, ніж середовище R. Основним недоліком R є те, що більшість його функцій повинні завантажувати всі дані в пам'ять перед виконанням, що встановлює обмеження на обсяги їх обробки. Однак, деякі пакети починають звільнятися від такого обмеження. Одним з прикладів є biglmpackage для лінійних моделей.

Останній критерій оцінки можливостей статистичного програмного забезпечення – це документація. Середовище R має багато доступних файлів документації. В той час як SPSS не має цієї особливості через обмеженість використання. SAS має комплексну технічну документацію, яка становить понад 8000 сторінок.

Оскільки інструмент SAS більш широко використовується у великих підприємствах, ніж SPSS Statistics IBM, то він містить більше таких джерел та ресурсів, як форуми, користувацькі клуби, тренери, веб-сайти, макро бібліотеки та книги. Однак спільнота R є однією з найпотужніших структур відкритих джерел. SAS пропонує більше попередньо визначених математичних і фінансових функцій, ніж в IBM SPSS Statistics. До них відносяться амортизація, складні відсотки, грошовий потік, гіперболічні функції, факторіали, комбінації та механізми тощо.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бідюк, П.І. Прикладна статистика / П.І. Бідюк, О.М. Терентьєв, Т.І. Просянкіна-Жарова. – Вінниця: ПП «Едельвейс і К», 2013. – 288 с.
2. Приступая к программированию в SAS Studio 3.2 [Electronic resource]. – URL https://support.sas.com/documentation/cdl_alternate/ru/webeditorgs/67431/PDF/default/webeditorgs.pdf
3. Терентьев А.Н. SAS BASE: Основы программирования / Терентьев А.Н., Домрачев В.Н., Костецкий Р.И. – К: Эдельвейс, 2014. – 304 с.

4. R vs SAS vs SPSS – Top 3 Data Analytics tools Comparison [Electronic resource]. – URL : https://data-flair.training/blogs/r-sas-spss-data-analytics-tools-comparison/

Відомості про авторів	
Прізвище, ім'я, по батькові **студента** (повністю), курс, факультет	Бойко Яків Анатолійович, IV курс, фізико-математичний факультет
Прізвище, ім'я, по батькові **керівника** (повністю), вчене звання, посада	Терентьєв Олександр Миколайович, канд.тех.наук, старший викладач
Повна назва навчального закладу	Уманський державний педагогічний університет імені Павла Тичини
Поштова адреса з індексом (домашня)	вул.Леніна 70, кв.132, Умань, Черкаська обл., 20301
Контактний телефон	0973894719
E-mail	yakivboyko@meta.ua
Тема доповіді	Використання сучасних статистичних програм R, SPSS і SAS для обробки масивів даних
Напрям:	Інформаційно-комунікаційні технології в освіті та науці