ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

УДК 004.43

Яків Бойко

(Умань, Україна)

ВИКОРИСТАННЯ СТАТИСТИЧНИХ МОВ ПРОГРАМУВАННЯ SAS, SPSS, R І PYTHON ДЛЯ АНАЛІЗУ ДАНИХ

У даній статті розглядається питання використання статистичних мов програмування SAS, SPSS, R і PYTHON для аналізу даних. Використання статистичних мов програмування SAS, SPSS, R і Python для обробки даних залежать від конкретних поставлених цілей, а тому вони сильно відрізняються за такими критеріями, як статистичні методи та прийоми, доступність у вивченні, візуалізація, підтримка та витрати.

Ключові слова: *статистичні мови програмування, статистичні методи, візуалізація, пояснювальна та прогностична моделі, інтерфейс користувача.*

The article deals with the problem of using statistical programming languages SAS, SPSS, R and PYTHON. In this article, we compare the four languages on methods and techniques, ease of learning, visualization, support and costs. We explicitly focus on the languages, the user interfaces SAS Enterprise Miner and SPSS Modeler are out of scope. The use of four languages depends on the goal.

Key words: *statistical programming languages, statistical methods, visualization, explanatory and predictive models, user interface.*

До основних статистичних мов програмування слід віднести SAS, SPSS, R і Python. З них найбільш часто використовуються SAS і SPSS. Проте зацікавленість мовами з відкритим кодом R та Python зростає. В останні роки відбулася певна міграція користувачів від програмування на SAS або SPSS до використання R та / або Python. В даному досліджені наше завдання полягає в порівнянні чотирьох мов відповідно до таких критеріїв: статистичні методи та прийоми, доступність у вивченні, візуалізація, підтримка та витрати.

SAS був розроблений у Державному університеті штату Північна Кароліна і його основне завдання полягало в аналізі великої кількості даних про сільське господарство. Скорочення «SAS» означає система статистичного аналізу. У 1976 році була заснована компанія SAS як наслідок на попит на відповідне програмне забезпечення [3]. Статистичний пакет для соціальних наук (Statistical Package for the Social Sciences) (SPSS) був розроблений саме для соціальних наук і став першою статистичною мовою програмування для ПК. Вона була розроблена в 1968 році в Стенфордському університеті, і через вісім років була заснована компанія SPSS Inc., куплена IBM в 2009 році.

У 2000 році Університет Окленду випустив першу версію мови програмування R, яка орієнтована на статистичне моделювання і відкрита за ліцензією GNU. Python – єдина мова програмування, яка була розроблена за межами університету. Python був створений голландцем, який є великим шанувальником Монті Пайтон (звідки походить назва). Python - це багатоцільова мова, така як C ++ та Java, з однією великою перевагою – Python легкий у вивченні. Програмістами було створено безліч модулів для Python, і тому на сьогоднішній день він має широкий спектр можливостей статистичного моделювання.

До першого критерію відносимо статистичні методи та прийоми. Аналіз даних передбачає два підходи: пояснювальний та прогностичний. Використання підходу залежить від вашої кінцевої мети. Прогностична сторона аналізу даних тісно пов'язана з такими поняттями, як Data Mining і Machine Learning.

Коли ми розглядаємо SPSS та SAS, то обидві ці мови походять з пояснювальної сторони аналізу даних (Data Analysis). Вони були розроблені в навчальному середовищі, де тестування гіпотез грає важливу роль. Це означає, що вони мають значно менше методів і прийомів у порівнянні з R і Python. Сьогодні у SAS та SPSS є інструменти для обробки даних (data mining) (SAS Enterprise Miner і SPSS Modeler), однак це різні інструменти, і вам знадобляться додаткові ліцензії [4].

Одним з основних переваг засобів або програм з відкритим вихідним кодом є те, що спільнота постійно покращує і підвищує його функціональність.

R був створений розробниками, які хотіли зробити свої алгоритми якомога доступнішими. Ergo R має найширший діапазон алгоритмів, що робить R сильним як на пояснювальній стороні, так і на прогностичній (Data Analysis).

Python був створений з акцентом на бізнес використання, а не на академічні чи статистичні підходи. Основна сильна сторона Python полягає у використанні алгоритмів. Python в основному використовується в Data Mining або Machine Learning, де аналітика даних не потребує втручання. Ще однією сильною стороною Python є ефективний аналіз зображення та відео. Python також відноситься до найпростіших мов, які можуть бути використанні для фреймворків даних, таких як Spark [7].

Наступним чинником виступає легкість та простота у вивченні. Інструменти SPSS і SAS мають зручний інтерфейс користувача, де не обов'язково прописувати код. Крім того, SPSS містить paste-функцію, яка створює синтаксис з кроків, що були виконані у інтерфейсі користувача, а SAS має Proc SQL, що робить SAS-кодування значно простішим для людей, які знають мову запитів SQL. В SAS та SPSS коди синтаксично далекі один від одного, а також дуже відрізняються від інших мов програмування.

Незважаючи на те, що існують альтернативи графічному інтерфейсу R, як, наприклад, Rattle, з точки зору його функціональності він і близько не наближається до SAS або SPSS. Мова R легка у вивченні програмістам, проте багато аналітиків не мають досвіду в програмуванні. R має найскладніший процес навчання з усіх. Але як тільки ви отримаєте основи, тоді такий процес значно спрощується. Python, який заснований на ABC, розроблювався з єдиною метою навчання не-програмістів як програмувати. Читабельність є однією з ключових особливостей Python. Це робить Python однією з найпростіших мов для вивчення. Python не містить GUI. Таким чином, якщо брати до уваги простоту у вивченні, то SPSS та SAS є найкращим варіантом для початку аналізу, оскільки вони надають інструменти, де користувачеві не потрібно програмувати.

Третій основний фактор, який впливає на вибір програмного забезпечення, полягає в офіційній підтримці. Як SAS, так і SPSS є комерційними продуктами, тому вони мають офіційну підтримку [5]. Це спонукає деякі компанії вибирати

ці мови програмування: якщо існують певні труднощі – то можна отримали підтримку.

Існує неправильне уявлення стосовно підтримки інструментів з відкритим вихідним кодом. Це правда, що немає офіційної підтримки від творців або власників, але існує велика спільнота для обох мов, які готові допомогти вам вирішити вашу проблему. Великі шанси, що ваше запитання вже було задано і відповідь була дана на таких сайтах, як Stack Overflow. Крім того, є численні компанії, які надають професійну підтримку R і Python. Таким чином, хоча офіційної підтримки для R і Python немає, на практиці у вас є шанси отримати відповідь швидше стосовно R або Python, ніж у випадку SAS або SPSS.

Четвертим основним критерієм є візуалізація. Графічні можливості SAS і SPSS є чисто функціональними; хоча можна внести незначні зміни в графіки для точного налаштування графіків. Візуалізація в SAS і SPSS може бути дуже громіздким інструментом або навіть неможливим. R і Python пропонують набагато більше можливостей для налаштування та оптимізації графіків завдяки широкому діапазону доступних модулів. Найбільш широко використовуваним модулем для R є ggplot2, який має широкий набір графіків, де ви можете налаштувати практично все. Ці графіки також легко створити інтерактивними, що дозволяє користувачам грати з даними за допомогою програм, таких як shiny.

Python і R запозичили (і все ще роблять) багато один від одного. Одним з кращих прикладів цього є те, що у Python є також модуль ggplot, який має практично ту саму функціональність і синтаксис, що й у R. Ще одним широко використовуваним модулем для візуалізації в Python є Matplotlib [7].

Останнім критерієм виступає врахування витрат при виборі програм для аналізу. Інструменти R і Python є програмами з відкритим вихідним кодом, що робить їх доступними для всіх. R і Python – це мови, які важче вивчати в порівнянні з графічним інтерфейсом SAS або SPSS. В результаті, аналітики, які володіють R та / або Python, мають більш високі зарплати, ніж інші аналітики. Навчання працівників, які наразі не знайомі з R та / або Python, також коштує грошей. Отже, на практиці мови програмування з відкритим вихідним кодом можуть передбачати в собі витрати, але коли ви порівнюєте його з ліцензійними

витратами за SAS або SPSS, бізнес рішення легко прийняти: R і Python - це набагато дешевші варіанти.

Отже, використання інструментів R чи Python залежать від конкретної ситуації. Ці мови можна використовувати без необхідності придбання ліцензій і очікування на них. Крім ліцензій, однією з головних причин вибору є широкий спектр статистичних методів: існує можливість використання із великої кількості алгоритмів саме той, який підходить найкраще для вирішення даної проблеми.

Сильна сторона Python полягає в його застосуванні в машинному навчанні. Найкраще використовувати Python, наприклад для програм Face або Object Recognition або Deep Learning. Використовувати R має сенс для цілей, які мають відношення до поведінки споживачів, де також пояснювальна сторона відіграє важливу роль. Ці дві мови значною мірою взаємодоповнюють один одною. Є бібліотеки для R, які дозволяють запускати код Python (reticulate, rPython), і є модулі Python, які дозволяють запускати R-код (rpy2). Це робить комбінацію двох мов ще сильнішою.

До найбільш вживаних статистичних мов програмування слід віднести SAS, SPSS, які містять зручний інтерфейс і багато вбудованих функцій. SAS Advance Analytics добре підходить для візуальних даних, де використовуються кілька графіків та моделей. Однак його не варто застосовувати для описових даних. SAS містить багато позитивних сторін: надзвичайно універсальне поєднання доступних функцій, наявність критичних аналітичних можливостей і використання R для нових підходів.

Основний недолік SPSS полягає в тому, що оновлення не відбуваються дуже часто. Оскільки інтерфейс SPSS не зазнав багато змін з моменту випуску, то, відповідно, старі статистичні методи аналізу все ще присутні в ньому. Також виникають певні труднощі через обмежені можливості імпорту даних, що може призвести до великої витрати часу на виправлення таких помилок. В SPSS важко працювати з великими файлами або текстовими даними, які перевищують певну кількість символів.

Використання статистичних мов програмування SAS, SPSS, R і Python для аналізу даних залежать від конкретних поставлених цілей, а тому вони сильно відрізняються за такими критеріями, як статистичні методи та прийоми, доступність у вивченні, візуалізація, підтримка та витрати.

табл. 1

Основні переваги та недоліки статистичних мов програмування SAS, SPSS, R і Python для аналізу даних

	SAS	**SPSS**
Переваги	**1.** Широке використання в промисловості 2. Потоко-орієнтований інтерфейс (інтерфейс з потоком даних) з drag and drop 3. Офіційна підтримка 4. Можливість обробки великих масивів даних 5. Використання PROC SQL	1. Широке використання в навчальному середовищі та університетах 2. Зручний інтерфейс з великою кількістю документації 3. Функція Click and Play 4. Полегшення написання коду за допомогою кнопки Paste 5. Офіційна підтримка
Недоліки	1. Ліцензійні послуги 2. Необхідність написання коду для нестандартних опцій 3. Повільний перехід до нових методів 4. Різне програмне забезпечення для візуалізації та Data Mining	1. Ліцензійні послуги 2. Необхідність різних ліцензій для функціоналу 3. Обмежений синтаксис 4. Повільний перехід до нових методів 5. Повільність в обробці великих масивів даних
	R	**Python**
Переваги	1. Велика кількість бібліотек 2. Безкоштовність 3. Достатньо повно представлені пояснювальне та прогностичне моделювання	1. Масштабованість 2. Універсальна багатоцільова мова 3. Доступність у вивченні 4. Ефективність в машинному навчанні

	4. Можливість підключення до джерел даних, включаючи NoSQL і webscraping 5. Встановлення IDE (RStudio) і необхідних пакетів обробки даних максимально спрощена	5. Підтримка великої спільноти 6. Безкоштовність
Недоліки	1. Повільність в обробці великих масивів даних (необхідність використовувати бібліотеки data.table and dplyr) 2. Специфічність в порівнянні зі стандартними мовами програмування, так як мова вузькоспеціалізована (наприклад, індексація векторів починається замість нуля з одиниці) 3. Труднощі у вивченні 4. Відсутність офіційної підтримки 5. Відсутність інтерфейсу користувача	1. Недостатньо представлена в пояснювальній моделі 2. Вибір версії: 2.7 чи 3.5 3. Відсутність інтерфейсу користувача 4. Відсутність офіційної підтримки 5. Python - це мова з динамічною типізацією. Це істотно прискорює розробку програм, але також ускладнює пошук деяких важко відслідковуваних помилок, пов'язаних з невірним присвоюванням різних даних одним і тим же змінним

ДЖЕРЕЛА ТА ЛІТЕРАТУРА

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. Explanatory vs. Predictive Models in Machine Learning 2 [Electronic resource]. – URL https://velotio.com/blog/2017/4/22/machine-learning-choice-of-models
5. IBM SPSS vs SAS Advanced Analytics [Electronic resource]. – URL : https://www.trustradius.com/compare-products/ibm-spss-vs-sas-advanced-analytics
6. Python и R: что выбрать для Data Science в 2018? 2 [Electronic resource]. – URL https://proglib.io/p/python-vs-r/
7. Python & R vs. SPSS & SAS [Electronic resource]. – URL : https://www.r-bloggers.com/python-r-vs-spss-sas/

Бойко Яків Анатолійович – кандидат педагогічних наук, доцент кафедри англійської мови та методики її навчання; тел. 0973894719; yakivboyko@meta.ua; сертифікат: ТАК; про конференцію дізнався від колег