МИСЬКІВ Т.Й.,
старший викладач,
Самбірський технікум економіки та інформатики,
м. Самбір, Україна

ВИБІР ПРОГРАМНОГО ІНСТРУМЕНТАРІЮ ДЛЯ ВИВЧЕННЯ СТАТИСТИЧНИХ ДИСЦИПЛІН

На сьогодні існує безліч комп'ютерних програмних пакетів для статистичного аналізу даних (ПЗСА). Вибір ПЗСА для вивчення статистичних дисциплін є нетривіальною задачею. При виборі пакетів використані релевантні критерії: поширеність на нашому ринку; вартість ліцензії; функціональність; локалізація для вітчизняних користувачів. В наших умовах витрати для користувачів особливо значимі. Ми врахували цей чинник в нашому дослідженні. Для цілей навчання рейтинги найбільш розповсюджених пакетів R, SAS та SPSS приблизно рівні.
Ключові слова: *статистика, програмне забезпечення, навчання.*

В настоящее время существует множество программных пакетов для статистического анализа данных (ПОСА). Выбор ПОСА для изучения статистических дисциплин является нетривиальной задачей. При выборе пакетов использованы релевантные критерии: распространенность на нашем рынке; функциональность; стоимость лицензии; локализация для отечественных пользователей. В наших условиях стоимость для пользователей особенно значима. Мы учли этот фактор в нашем исследовании. Рейтинги трех наиболее распространенных пакетов R, SAS и SPSS для целей обучения примерно равны.
Ключевые слова: *статистика, программное обеспечение, обучение.*

Today there are many software packages for statistical data analysis (SDAS). In the teaching process the students must to master the skills necessary to operate in the most popular packages. The choice of the SDAS is not a trivial task. For selecting of packages is

 © МИСЬКІВ Т.Й., 2016

relevant criteria: prevalence in our market; functionality; license cost; localization for our users. In our conditions, the cost to users is particularly important. We have taken into account this factor in our research. The ratings of the three most common packages R, SAS and SPSS is about equal. Depending on the students contingent there may be some advantage in choosing one of them. The trend in the market of the SDAS is the movement toward open source technology.

Key words: *statistics, analysis, computing, software, teaching.*

Постановка проблеми. Останнім часом намітилися стійкі тенденції поширення використання статистичного аналізу даних. Фахівці статистичної аналітики бажані не тільки в українських філіях зарубіжних компаній, де це вже традиційно, але і на суто вітчизняних підприємствах, що прийняли європейські стандарти сталого розвитку. Аналітика, як науковий метод, виявляє нове розуміння накопичених даних, яке створює конкурентну перевагу, коли використовується при прийнятті рішень. З'явились історії успіху компаній, які за допомогою статистичного аналізу знайшли свою нішу в сучасному світі. Наприклад, компанія IBM (США) у зв'язку з втратою ринку великих обчислювальних машин не змогла б вижити в 1990-ті роки. Використання аналітики привнесло наукову обгрунтованість в процес прийняття бізнес-рішень, стало фундаментом для трансформації компанії [1]. Сьогодні IBM випускає програмне забезпечення для статистичного аналізу (ПЗСА), в яких реалізований в тому числі і успішний досвід компанії в цій сфері.

Статистичний аналіз даних вже давно ведеться виключно з застосуванням комп'ютерних технологій. Вивчення курсів статистичних дисциплін обов'язково передбачає освоєння пакетів ПЗСА. За даними Forbes [2] світ вже зіштовхується з дефіцитом випускників за напрямками статистичного аналізу даних. Тому актуальність підготовки сучасних фахівців в цій предметній галузі сумнівів не викликає.

На сьогодні існує безліч пакетів ПЗСА. Випускники повинні отримати в процесі навчання як універсальні навички роботи в ПЗСА, так і конкретику найбільш поширених пакетів, орієнтуватися в програмних інструментах відомих виробників. Конкурентні переваги надає вміння аналізувати дані за допомогою ПЗСА, що знаходиться в експлуатації у потенційних роботодавців. У зв'язку з цим, задача вибору програмного інструментарію для вивчення статистичних дисциплін не є тривіальною.

Аналіз актуальних досліджень. Останнім часом суттєво поглибилось застосування ПЗСА від традиційної державної та медичної статистики до

емпіричної соціології і політології [3] та маркетингу [4, 5], кризового менеджменту [6] тощо. Внаслідок цього отримати універсальні показники для порівняння пакетів ПЗСА не видається можливим.

Найвідоміші рейтинги, наприклад, компанії Gartner [7], порівнюють функціональність ПЗСА у відриві від їх вартості, яка проте особливо актуальна саме для вітчизняних умов і є одним з ключових чинників застосовності. Необхідно розглянути весь спектр критеріїв функціональності ПЗСА та найбільш розповсюджені серед безкоштовних пакетів: статистичну мову R, графічні пакети на її основі Rstudio, Revolution R тощо [8].

Відомі зарубіжні дослідження функціональності останніх версій пакетів ПЗСА, як в якісному [9], так і в рейтинговому [10, 11] плані. Проте умови їх проведення не відповідають нашим реаліям і, відповідно, результати можуть бути використані з істотними обмеженнями.

Останні поліпшення ПЗСА направлені на розширення можливостей пакетів розпорядчої аналітики [12], розгортання її сервісів у хмарах [13], інтеграцію з системами інтелектуального аналізу даних (Data Mining) [14] в контексті четвертої парадигми наукових досліджень [15]. Вказані фактори теж в певній мірі впливають на рейтинг пакетів і частково враховані в нашій роботі.

Мета статті. Метою цієї роботи є дослідження найбільш поширених пакетів ПЗСА на предмет їх використання в навчальному процесі з дисциплін статистичних спрямувань. При виборі програмного забезпечення для вивчення статистичних курсів скористаємось релевантними критеріями: поширеність ПЗСА на нашому ринку; функціональність для задач, що стануть перед випускниками — майбутніми фахівцями; вартість ліцензії або ПЗСА з відкритим кодом, локалізація для вітчизняних користувачів.

1. Поширеність ПЗСА на ринку праці

Доступними відкритими джерелами для оцінки поширеності ПЗСА [16] є: оголошення про вакансії статистичного напрямку; наукові статті про програмні продукти та їх застосування; звіти дослідницьких організацій в IT-сфері; огляди використання; книги в предметній галузі; продажі і завантаження пакетів на сайтах виробників тощо.

Одним з кращих способів для вимірювання популярності або частки на ринку пакетів ПЗСА є підрахунок кількості згадувань у вимогах (або бажаності) для вакансій кожного з них. Оголошення про роботу підкріплені реальними коштами роботодавців, тому вони на сьогодні, можливо, є найкращим показником популярності пакетів. Аналіз найбільших українських сайтів вакансій work.ua та rabota.ua, показує, що існує значний попит на фахівців зі статистичного аналізу даних (більше ніж 300 вакансій за 30 днів). Серед вимог та побажань до претендентів на вакансії вказане зокрема володіння ПЗСА SAS, SPSS, R, MATLAB, Python тощо. Але ці сайти, на жаль, не надають

повнотекстового пошуку за деталями вакансій, що дозволило б виявити кількість фахівців, до яких у вимогах наведено володіння конкретними пакетами. Тому проведемо дослідження на сайті пошуку роботи в Україні [17], який надає таку можливість та пропонує більше 130 вакансій статистичного напрямку.

Для порівняння скористаємось результатами дослідження [18] по найбільшому сайту пошуку роботи в США Indeed.com, а також дослідженням [19], яке проводилось без розгляду пакета SPSS. Порівняння проведемо за розподілом між собою трьох найбільш розповсюджених пакетів програмного забезпечення SPSS, SAS та R. Введемо свій рейтинг пакетів, як експертну оцінку поширеності пакетів з урахуванням зазначених джерел та власних досліджень. Результати дослідження наведено в таблиці 1.

Таблиця 1. Відносна поширеність пакетів ПЗСА на ринку праці

Пакет	Вакансії України	Вакансії США [18]	Вакансії США [19]	Книги [18]	Наукові статті [18]	Найвідоміші користувачі [20]	Рейтинг
SAS	30	52	39	576	3200	HP, Paribas (Укрсиббанк)	4
SPSS	55	19	-	339	7600	Canon, ABN	5
R	25	29	42	240	3500	Google	4

Відмінність розподілу популярності пакетів ПЗСА в Україні можна пояснити тим, що пропріетарні пакети надто вартісні для середніх українських компаній, а зарубіжні компанії, що працюють в Україні, представляють лише окремі галузі ринку. Особливим попитом у нас користуються соціологічні дослідження, в яких стандартом є SPSS.

В графі "Книги" таблиці наведено кількість книг, виданих у світі, які містять найменування відповідного ПЗСА в їх назві. Написання книги потребує значних коштів та зусиль, тому видавці роблять своє власне дослідження частки ринку, перш ніж приймати ризик публікації. Ми опосередковано користуємося цими дослідженнями. Чим популярніший програмний пакет, тим більша ймовірність, що він з'явиться в наукових публікаціях як інструмент аналізу або навіть об'єкт дослідження [18]. В графі "Наукові статті" наведено кількість статей, опублікованих у 2015 році, зі згадуванням відповідних пакетів.

Останнім часом для статистичних досліджень поряд з R набуває популярності і мова програмування Python, яка розроблялась спочатку для WEB-

застосувань і вже набула в інших сферах високого статусу. На жаль дослідження її популярності в статистичному аналізі зіштовхуються з труднощами неможливості виділити в рейтингу Python частину її використання в статистиці. Наприклад, в рейтингу [21] за 2016 рік однієї з найавторитетніших організації Python посідає третє місце, а R п'яте. Але для Python це використання у всіх галузях, в той час як використання R орієнтоване лише на статистичну аналітику.

2. Функціональність найпоширеніших пакетів ПЗСА

При аналізі функціональності інструментів для статистичного аналізу даних, необхідно враховувати багато факторів: можливість легкого встановлення пакету на комп'ютері користувача; охоплення програмним забезпеченням всіх методів, що потрібні користувачу, якщо не всіх, то наявність можливості оновлень та розширень; використання для розширюваності вбудованої мови (або зовнішньої мови, наприклад, R); достатність для користувача засобів візуалізації результатів аналізу; можливість експорту результатів в пакети презентацій; обмеження на обсяги наборів даних, що підлягають аналізу; поширеність серед колег в предметній галузі, щоб можна було обмінюватись даними і програмами; вартість програмного забезпечення; наявність навчальних версій.

Будемо розглядати в цій роботі програмні пакети статистичного аналізу в розрізі повноти підтримки введених вперше компанією IBM рівнів аналітики [12]: описового (descriptive), прогностичного (predictive) та розпорядчого (prescriptive). Описова аналітика є найбільш простою та найчастіше використовуваною і містить стандартну звітність, спеціальну звітність та аналіз частоти і надає єдине уявлення про минуле, дозволяє краще контролювати та управляти процесами в державі, регіоні, бізнесі. Прогностична аналітика є наступним рівнем і містить: інтелектуальний аналіз даних (кореляція в даних), розпізнавання образів для попередження про необхідність вжиття заходів, аналіз основних причин, прогнозне моделювання. Вона використовує розуміння минулого для ймовірнісного передбачення на майбутнє. Інтелектуальний аналіз допомагає виявити приховані закономірності в даних з метою прогнозування майбутніх подій. Розпорядча аналітика, як найвищий рівень, надає методи оптимізації для моделювання та вибору найкращих рішень для майбутнього з багатьох альтернативних варіантів, які містять в тому числі і оптимізацію при невизначеності деяких даних (стохастичну оптимізацію). Вона оцінює для особи, що приймає рішення, ймовірнісний результат та ризики альтернативних варіантів рішень, а також підтримує систему зворотного зв'язку, що відстежує результат прийнятих рішень.

Excel є найпопулярнішою електронною таблицею у світі і в Україні зокрема. Але він знаходиться на рівні описової аналітики [9] і слугує, в основному,

для складання стандартних звітів, тому не може бути універсальним інструментом для вивчення курсу статистики.

Якщо не враховувати надто високу вартість ПЗСА, пакети SAS і IBM SPSS є безперечними лідерами за найвідомішими рейтингами, наприклад, компанії Gartner [7] та мають найкращу функціональність серед всього сучасного програмного забезпечення. Найбільш розповсюдженим серед безкоштовних пакетів є статистична мова R та графічні пакети на його основі. Пакети SPSS, SAS, як правило, надають користувачам, які не мають достатнього фаху в галузі статистики, можливість аналізу даних, і розробки аналітичних моделей практично без кодування за рахунок інтуїтивно зрозумілого користувацького інтерфейсу для управління ходом аналітики в процесі підготовки даних, аналізу, валідації тощо [9]. Використання R відразу потребує програмування.

SAS є світовим лідером на корпоративному ринку (десятки тисяч робочих місць [7]). SAS та SPSS широко використовуються великими організаціями і корпоративними компаніями, а системи на R є кращим вибором для бюджетними компаній, які оптимізують ефективність витрат, та таких, що тільки розпочинають свою діяльність.

Відволікаючись від несуттєвих нововведень, дослідження [11] показують, що всі три пакети мають достатні можливості оброблення даних, включаючи паралельні обчислення. Аналогічна рівність пакетів спостерігається і в методах аналізу та моделювання. Всі вони підтримують базові методи: кореляційний, багатовимірний регресійний, факторний, дисперсійний та кластерний аналіз; аналіз часових рядів (експоненціальне згладжування, сезонна декомпозиція, спектральний аналіз); автоматичну класифікацію; дерева рішень; розпізнавання образів; нейронні мережі. Пропонують широкий набір інструментів для підтримки всіх етапів статистичного дослідження, для побудови описових і прогнозних моделей для всіх рівнів аналітики від описової до розпорядчої, для прогнозування, оптимізації, імітаційного та стохастичного моделювання тощо.

R має кращі графічні можливості для візуалізації результатів в порівнянні з двома іншими, є безліч вільних пакетів, які надають сучасні графічні можливості. Комерційні пакети, як SAS, наприклад, мають базові графічні можливості, але для їх розширення необхідно придбати додаткові пакети, що надто дорого для використання. Крім того це вимагає вивчення тонкощів, наприклад, пакета SAS Graph для його налаштування.

Звичайно, пропріетарні вартісні пакети SAS та SPSS мають кращу підтримку клієнтів і спільноти в порівнянні з вільними мовами. Спеціальна підтримка і обслуговування клієнтів входить до вартості пакета. R також має велику інтернет-спільноту, але ж це не центр підтримки клієнтів. Ви отримаєте допомогу від спільноти, але не оперативно.

Табл.2. Оцінка функціональності ПЗСА

	Оцінка [10]		Оцінка [11]		Узагальнено	
Критерії / пакети	SAS/SPSS	R	SAS/SPSS	R	SAS/SPSS	R
Освоєння	Складне	Просте	4,5	5	4	4
Операції з даними	Відмінно	Відмінно	4	4	5	5
Моделювання	Відмінно	Відмінно			5	5
Графіка	Низько	Відмінно	3	4,5	3	5
Підтримка	Відмінна	Низька	4	3,5	4	3
Разом					21	22
Оцінка					5	5

3. Загальний рейтинг

Вважаємо, що у вітчизняних умовах при виборі ПЗСА приблизно рівна вага критеріїв поширеності, функціональності, вартості та локалізації. Пакет R повністю безкоштовний отримує оцінку 5, а пакети SAS та SPSS надто вартісні і отримують оцінку 1. Щодо локалізації: пакети SAS та SPSS підтримують російську мову (оцінка 3), а R не локалізований (оцінка 1). Результати дослідження наведено в таблиці 3.

Таблиця 3. Загальна оцінка пакетів та кінцевий рейтинг пакетів

Пакет	Поширеність	Функціональність	Вартість	Локалізація	Рейтинг
SAS	4	5	1	3	13
SPSS	5	5	1	3	14
R	4	5	5	1	15

Висновки. Проведені дослідження показали, що вибір користувачами ПЗСА є компромісом між витратами і набутими вигодами. В наших умовах витрати для користувачів особливо значимі. Ми врахували цей чинник в нашому дослідженні. Проте немає явного переможця серед трьох найбільш розповсюджених пакетів. Вони мають свої переваги і недоліки, і їх сильні сторони забезпечують їм нішу в довгостроковій перспективі. Рейтинг кожного пакета дещо варіюється в залежності від сфери використання, можлива деяка перевага в виборі одного з них. Тенденцією на ринку ПЗСА є рух в бік технологій з відкритим вихідним кодом. Майже всі закордонні дослідники вважають, що для експертів і професіоналів необхідно знати, принаймні два з них, що додало б гнучкості для майбутнього і відкрило нові перспективи.

Для майбутнього варто дослідити пакет з відкритим вихідним кодом KNIME, який є світовим лідером слідом за SAS та SPSS.

References:

1. Dietrich B., Plachy E., Norton M.. Analytics Across the Enterprise: Analytics transforms a 'dinosaur'. // Analytics, INFORMS (the Institute for Operations Research and the Management Sciences), 2014, July, pp. 62–69.
2. Brown M.S. What Analytics Talent Shortage? How To Get And Keep The Talent You Need. Forbes, 2016. http://www.forbes.com/sites/metabrown/2016/06/27/what-analytics-talent-shortage-how-to-get-and-keep-the-talent-you-need/#638f10f221d4.
3. McKieL, Ryan L. An End to the Crisis of Empirical Sociology? Trends and Challenges in Social Research // Oxon, «Sociological Futures», 2016, 200 p.
4. Chaffey D. The best digital marketing statistics sources in 2016. Customer research & analysis. http://www.smartinsights.com/marketplace-analysis/customer-analysis/digital-marketing-statistics-sources.
5. Christopher N., McDonnell F. Elea R for Marketing Research and Analytics. Switzeland, Springer, 2015. – 454 p.
6. Laureano R. M. S., da Silva L. M., Grencho A. R. R. R.. Framework to implement business analytics: Phases and critical success factors. // 11th Iberian Conference on Information Systems and Technologies (CISTI), 2016.
7. Kart L., Herschel G., Linden A., Hare J. Magic Quadrant for Advanced Analytics Platforms. Gartner, Inc., 2016. https://www.gartner.com/doc/reprints? id=1-2YEIILW&ct=160210&st=sb.
8. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. New-York, Springer, 2013. – 418 p.
9. Ju A. Data Analytics Tool Comparison: Excel, SPSS and SAS. Bell, Canada, 2015. https://www.linkedin.com/pulse/data-analytics-tool-comparison-excel-spss-sas-angela-ju.
10. Shethna J. SAS vs R vs Python. EDUCBA, 2016. https://www.educba.com/sas-vs-r-vs-python.
11. Jain K. SAS vs. R (vs. Python) – which tool should I learn? // Analytics Vidhya, 2014. https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn.
12. Soltanpoor R., Sellis T.. Prescriptive Analytics for Big Data. // Databases Theory and Applications (27th Australasian Database Conference, ADC 2016, Sydney, 2016, Proceedings). Springer, 2016, LNCS, vol. 9877, pp. 245–256.
13. Pethuru R. Handbook of Research on Cloud Infrastructures for Big Data Analytics. IGI Global, 2014. - 570 p.
14. Tan Y.,.Shi Y. Data Mining and Big Data: First International Conference, DMBD 2016, Bali, Indonesia, June 25-30, 2016. Proceedings. Springer, 2016 – 569 p.

15. G. Bell, T. Hey, A. Szalay. Beyond the Data Deluge. // Science, 2009 vol. 323, no. 5919, pp. 1297-1298.
16. Muenchen R.A. R for SAS and SPSS Users. Springer-Verlag New York, 2011. – 686 p.
17. Work in Ukraine. HeadHunter Group, 2016. https://hh.ua.
18. Muenchen R.A. The Popularity of Data Analysis Software. // Analyzing the World of Data Science, 2016. http://r4stats.com/articles/popularity.
19. SAS, R, or Python Survey 2016: Which Tool Do Analytics Pros Prefer? Burtch Works: Executive Recrution, 2016. http://www.burtchworks.com/ 2016/07/13/ sas-r-python-survey-2016-tool-analytics-pros-prefer.
20. Walker M. SAS, R and SPSS Comparison. Data Science Association, 2014. http://www.datascienceassn.org/content/sas-r-and-spss-comparison.
21. Diakopoulos N., Cass S. The Top Programming Languages 2016. Ranking IEEE Spectrum, 2016. http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016

Myskiv T.Y.
Senior Lecturer, Samborskii College of Economics and Informatics
Sambor, Ukraine

The choice of software tools for studying statistical disciplines

t.korkuna@gmail.com