УДК [004.422.63:332.012.2]:004.67

Петров Ю. А., Петрова Г. И.

АНАЛИЗ И ОБРАБОТКА БОЛЬШИХ МАССИВОВ ПОЛУСТРУКТУРИРОВАННЫХ ЭКОНОМИЧЕСКИХ ДАННЫХ

Юрий Александрович Петров
кандидат химических наук, доцент
youri1054@gmail.com
ФГАОУ ВО «Российский государственный профессионально-педагогический университет», Россия, Екатеринбург
Галина Ивановна Петрова
кандидат философских наук, доцент
galinapetrova477@gmail.com
ФГАОУ ВО «Российский государственный профессионально-педагогический университет», Россия, Екатеринбург

ANALYSIS AND PROCESSING OF LARGE ARRAYS OF SEMI-STRUCTURED ECONOMIC DATA

Iurii Aleksandrovich Petrov
Russian State Vocation Pedagogical University, Russia, Yekaterinburg
Galina Ivanovna Petrova
Russian State Vocation Pedagogical University, Russia, Yekaterinburg

***Аннотация**. В статье рассматриваются способы обработки больших массивов полу-структурированных социально-экономических данных средствами стандартных офисных программ таких, как Microsoft Excel. Такой подход даёт возможность ранжировать, визуализировать и анализировать данные, а также рассчитывать другие важные показатели, отсутствовавшие в исходных данных.*

***Abstract** The article discusses ways of processing large arrays of semi-structured socio-economic data using standard office programs such as Microsoft Excel.*

This approach makes it possible to rank, visualize and analyze data, as well as calculate other important indicators that were absent in the original data.

Ключевые слова: *большие полу-структурированные данные, анализ, обработка, визуализация, ранжирование, экономические системы.*

Keywords: *large semi-structured data, analysis, processing, visualization, ranking, economic systems.*

Экономические системы относятся к сложным искусственным системам [1], обладающим огромным количеством показателей, свойств и прочих характеристик, часть из которых относятся к данным, другие же относятся к расчётным показателям, полученным на основе этих данных. При этом часто встаёт задача сопоставления, визуализации и анализа этих показателей. Набор экспериментальных данных часто достаточно ограничен и не охватывает всей совокупности интересующих исследователя характеристик. В этом случае нередко используются модели, в которых используются взаимосвязи «свойство — свойство». Идеология и теоретические основы такого подхода, в частности, были разработаны и апробированы на обширном экспериментальном материале в работах [2, 3]. Развитием этих представлений явились их приложения к описанию и прогнозированию некоторых характеристик и свойств социально-экономических систем: взаимосвязи показателей качества жизни [4]; иерархическая матричная модель уровней компетентности [5] и образовательная функция, количественно описывающая траектории формирования и развития компетентностей [6]; демографическое прогнозирование контингента студентов вузов РФ [7]; преобразование данных и приведение их к виду, пригодному для построения моделей прогнозирования в рамках стандартного набора аппроксимирующих функций в MS Excel [8]. Существенным фактором, облегчающим обработку и анализ использованных в этих работах данных, было то, что они уже изначально были сформированы в виде, пригодном для использования в Excel.

В настоящей работе на основе выше названных представлений рассматривается способ обработки большого массива основных экономических показателей 2000 крупнейших публичных компаний мира, таблично представленных в полу-структурированном виде на информационно-аналитическом портале Forbes в разделе Global 2000 [9]. На рисунке 1 приведён начальный фрагмент этой таблицы, скопированной и вставленной в Excel для её дальнейшей обработки. Рейтинг составлен в 2020 году, но данные, приведённые в нём, относятся к 2019 финансовому году.

Rank	Company	Country/Territory	Sales	Profits	Assets	Market Value
1	ICBC	China	$177.2 B	$45.3 B	$4,322.5 B	$242.3 B
2	China Construction Bank	China	$162.1 B	$38.9 B	$3,822 B	$203.8 B
3	JPMorgan Chase	United States	$142.9 B	$30 B	$3,139.4 B	$291.7 B
4	Berkshire Hathaway	United States	$254.6 B	$81.4 B	$817.7 B	$455.4 B
5	Agricultural Bank of China	China	$148.7 B	$30.9 B	$3,697.5 B	$147.2 B
5	Saudi Arabian Oil Company (Saudi Aramco)	Saudi Arabia	$329.8 B	$88.2 B	$398.3 B	$1,684.8 B
7	Ping An Insurance Group	China	$155 B	$18.8 B	$1,218.6 B	$187.2 B
8	Bank of America	United States	$112.1 B	$24.1 B	$2,620 B	$208.6 B
9	Apple	United States	$267.7 B	$57.2 B	$320.4 B	$1,285.5 B
10	Bank of China	China	$135.4 B	$27.2 B	$3,387 B	$112.8 B

Рисунок 1 — Начальный фрагмент рейтинга Global 2000

Полностью весь массив данных занимает 20 страниц на сайте (по 100 компаний на каждой странице), что не слишком удобно для навигации и поиска интересующих данных. Кроме того, все 8000 ячеек с данными представлены не в числовом, а в текстовом формате, что удобно для чтения и зрительного восприятия показателей, но не пригодно для их обработки, вычислений, визуализации и анализа, в том числе и в программе Excel.

Для решения этой задачи, скопированные и вставленные на 1 страницу книги Excel, данные подверглись последовательной обработке с использованием имеющихся в программе инструментов. А именно (с помощью инструмента «найти» > «заменить»): удалены символы $, а также удалены символы, обозначающие единицы измерения — B (миллиарды) или M (миллионы); раз-

делители «точка», заменены на разделители «запятая»; все численные показатели приведены к одинаковым единицам измерения (миллиарды). В результате этих преобразований все данные были переведены в численные и стали пригодными для дальнейших вычислений, группировок и сортировок, а также для их визуализации в виде диаграмм, рисунков и таблиц. Рисунок 2 показывает начальный фрагмент данных после их преобразования.

Rank	Company	Country/Territory	Sales	Profits	Assets	Market Va	Profitability
1	Walmart	United States	524	14,9	236,5	344,4	2,84%
2	Sinopec	China	369,2	3,3	254,8	76,6	0,89%
3	PetroChina	China	364,1	6,6	392,3	65,9	1,81%
4	Saudi Arabian Oil C	Saudi Arabia	329,8	88,2	398,3	1684,8	26,74%
5	Royal Dutch Shell	Netherlands	311,6	9,9	394	126,5	3,18%
5	Amazon	United States	296,3	10,6	221,2	1233,4	3,58%
7	Toyota Motor	Japan	280,5	22,7	495,1	173,3	8,09%
8	Volkswagen Group	Germany	275,2	12	538,9	70,4	4,36%
9	BP	United Kingdom	271,6	-3,3	273,9	79,4	-1,22%
10	Apple	United States	267,7	57,2	320,4	1285,5	21,37%

Рисунок 2 — Начальный фрагмент данных Global 2000 после их преобразования и обработки в Excel

Данные, представленные на рисунке 2, сгруппированы по основному финансово-экономическому показателю — выручке, который чаще всего принимается при рассмотрении вопроса о величине (размере) компании. Кроме того, в этих данных приведён и ещё один важнейший показатель экономической эффективности — рентабельность. Этот показатель отсутствует в исходных данных, но был вычислен нами для всех 2000 компаний в одно действие — делением столбца «прибыль» на столбец «выручка» с использованием инструмента Excel «специальная вставка» > «разделить» и выбора «процентного» отображения результата.

Одной из целей данной работы являлось не только ранжирование крупнейших компаний мира по основным показателям (выручка, прибыль, активы, маркетинговая капитализация и рентабельность), причём, как в глобальном

представлении, так и в национальном разрезе (по отдельным ведущим странам, включая Россию), но также и визуализация этих данных для выявления характера распределения ключевых показателей по всей совокупности компаний. Для удобства сравнения все показатели были сгруппированы по 10 %-группам — по 200 компаний в каждой группе для всех 2000 компаний, либо меньшее, но соответствующее, число компаний для группировок по отдельным странам. Такое представление принято, в частности, для сопоставления среднемесячной зарплаты в отдельных отраслях и в целом по экономике России и отдельных её субъектов.

Рисунок 3 показывает распределение среднего размера выручки в каждой 10%-группе всей совокупности 2000 крупнейших компаний мира.

	1	2	3	4	5	6	7	8	9	10
Sales Aver.	1,25	2,861	4,69	6,756	9,05	11,708	15,216	21,255	35,167	103,801
Average	21,175	21,175	21,175	21,175	21,175	21,175	21,175	21,175	21,175	21,175

Рисунок 3 — Распределение средней выручки по 10%-группам крупнейших публичных компаний мира

Как показано в данных, представленных на рисунке 3, средний размер выручки компаний, входящих в список 2000 крупнейших компаний мира, со-

ставляет 21,175 млрд. $. При этом распределение выручки среди этих компаний далеко не линейное и только примерно ¼ компаний имеет выручку на уровне средней и выше неё, а ¾ компаний имеют выручку ниже средней.

Примерно так же распределены и все остальные показатели: прибыль, активы, рыночная капитализация, рентабельность. Средний показатель варьируется лишь в незначительных пределах : 75+/- 3%. Показатели независимые, но закономерность их распределения практически одинаковая. И чем больше группа охвата данных, тем более выражена эта закономерность и тем меньше флуктуации случайных отклонений.

На рисунке 4 представлены данные по распределению выручки среди крупнейших компаний России. Таких компаний в список Global 2000 вошло 23 и поэтому они не были подразделены на группы, а на рисунке отображены данные по всем компаниям Российского списка.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sales	0,7	2,2	2,3	3,6	4	7,4	7,9	10	10	10	13	14	15	16	16	16	21	23	25	48	116	123	127
Average	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27

Рисунок 4 — Распределение выручки среди крупнейших публичных компаний России

Из рисунка 4 видно, что средняя выручка компаний-флагманов Российской экономики заметно выше среднего показателя по всем 2000 крупнейших компаний мира — 27 млрд.$ в России по сравнению с 21,2 млрд. $ в среднем

по Global 2000. Но при этом только 4 компании (3 нефтегазовых гиганта — Роснефть, Газпром и Лукойл, а также крупнейший из банков — Сбербанк) имеют выручку выше средней, а 19 остальных компаний — ниже. И, несмотря на то, что со статистической точки зрения, выборка мала, мы опять наблюдаем близкую закономерность — чуть более 17% компаний имеют выручку выше средней, а около 82% — ниже. При большем охвате компаний, возможно, и результат мог бы быть немного иным.

Таким образом, рассмотренный здесь подход, позволяет, используя стандартное ПО пакета Microsoft Office, обрабатывать и структурировать большие массивы данных, а также исследовать, выявлять и анализировать закономерности и взаимосвязи в разнообразных социально-экономических системах.

Список литературы

1. *Губарев, А. В.* Семантические, аксиоматические и методологические основы феноменологической теории развития искусственных систем / А. В. Губарев, Ю. А. Петров, Г. И. Петрова. Текст: непосредственный // Наука. Информатизация. Технологии. Образование: материалы XI международной научно-практической конференции, Екатеринбург, 26 февраля – 2 марта 2018 г. / Рос. гос. проф.-пед. ун-т. Екатеринбург, 2018. С. 49–63.

2. *О взаимосвязи* свойств, описываемых методом кластерных компонентов / Ю. А. Петров, А. Ю. Попков, А. Н. Мень, В. М. Камышов, Г. И. Чуфаров. Текст: непосредственный // Доклады Академии наук СССР. 1983. Т. 272, № 4. С. 906.

3. *Петров, Ю. А.* Исследование кристаллохимических и магнитных свойств замещенных железоиттриевых гранатов: специальность 02.00.04 «Физическая химия»: диссертация на соискание ученой степени кандидата химических наук / Петров Юрий Александрович. Свердловск, 1984. 163 с. Текст: непосредственный.

4. *Петров, Ю. А.* Качество жизни: о взаимосвязи некоторых из основных показателей / Ю. А. Петров, Г. И. Петрова. Текст: непосредственный //

Академическая наука – проблемы и достижения = Academic science – problems and achievements: материалы VI международной научно-практической конференции, 25–26 мая 2015 г. North Charleston, USA, 2015. С. 36–40.

5. *Петров, Ю. А.* Уровни компетентности: модель, классификация, иерархия / Ю. А. Петров, Г. И. Петрова. Текст: непосредственный // Образовательные технологии. 2014. № 4. С. 65–70.

6. *Петров, Ю. А.* Образовательная функция в матричной модели уровней компетентности / Ю. А. Петров, Г. И. Петрова. Текст: непосредственный // Новые информационные технологии в образовании: материалы IX международной научно-практической конференции, Екатеринбург, 15–18 марта 2016 г. / Рос. гос. проф.-пед. ун-т [и др.]. Екатеринбург, 2016. С. 305–311.

7. *Петров, Ю. А.* Демографическое прогнозирование контингента студентов в вузах Российской Федерации / Ю. А. Петров, Г. И. Петрова. Текст: непосредственный // Демографический потенциал стран ЕАЭС: сборник статей VIII Уральского демографического форума, Екатеринбург, 08–09 июня 2017 г. / Ин-т экономики Урал. отд-ния Рос. акад. наук. Екатеринбург, 2017. С. 186–190.

8. *Петров, Ю. А.* Линеаризация аппроксимирующих функций при описании свойств экономических систем / Ю. А. Петров, Г. И. Петрова. Текст: непосредственный // Наука. Информатизация. Технологии. Образование: материалы XIII международной научно-практической конференции, Екатеринбург, 24–28 февраля 2020 г. / Рос. гос. проф.-пед. ун-т [и др.]. Екатеринбург, 2020. С. 632–641.

9. Информационно-аналитический портал Forbes. Раздел Крупнейшие публичные компании мира Global 2000. URL: https://www.forbes.com/global2000/#10f2617f335d. Текст: электронный.