ISSN 1998-0663 (print), ISSN 2587-8166 (online)
English version: ISSN 2587–814X (print), ISSN 2587–8158 (online)

Т. 14 № 2 – 2020

БИЗНЕС-ИНФОРМАТИКА

НАУЧНЫЙ ЖУРНАЛ НИУ ВШЭ

Издатель:
Национальный исследовательский университет «Высшая школа экономики»

Подписной индекс в каталоге агентства «Роспечать» –72315

Выпускается ежеквартально

Журнал включен в Перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук

Адрес редакции:
119049, г. Москва,
ул. Шаболовка, д. 28/11, стр. 4
Тел./факс: +7 (495) 772-9590 *26311
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

За точность приведенных сведений и содержание данных, не подлежащих открытой публикации, несут ответственность авторы

При перепечатке ссылка на журнал «Бизнес-информатика» обязательна

Тираж:
русскоязычная версия – 200 экз.,
англоязычная версия – 200 экз.,
онлайн-версии на русском и английском – свободный доступ

Отпечатано в типографии НИУ ВШЭ
г. Москва, Кочновский проезд, 3

СОДЕРЖАНИЕ

О ЖУРНАЛЕ

«Бизнес-информатика» – рецензируемый междисциплинарный научный журнал, выпускаемый с 2007 года Национальным исследовательским университетом «Высшая школа экономики» (НИУ ВШЭ). Администрирование журнала осуществляется школой бизнес-информатики НИУ ВШЭ. Журнал выпускается ежеквартально.

Миссия журнала – развитие бизнес-информатики как новой области информационных технологий и менеджмента. Журнал осуществляет распространение последних разработок технологического и методологического характера, способствует развитию соответствующих компетенций, а также обеспечивает возможности для дискуссий в области применения современных информационно-технологических решений в бизнесе, менеджменте и экономике.

Журнал публикует статьи по следующей тематике:

- ✦ автоматизация процессов управления и производства
- ✦ анализ данных и интеллектуальные системы
- ✦ информационные системы и технологии в бизнесе
- ✦ математические методы и алгоритмы бизнес-информатики
- ✦ программная инженерия
- ✦ интернет-технологии
- ✦ моделирование и анализ бизнес-процессов
- ✦ стандартизация, сертификация, качество, инновации
- ✦ правовые вопросы бизнес-информатики
- ✦ принятие решений и бизнес-интеллект
- ✦ моделирование социальных и экономических систем
- ✦ информационная безопасность.

В соответствии с решением президиума Высшей аттестационной комиссии Российской Федерации журнал включен в Перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук, по следующим группам научных специальностей: 05.13.00 – информатика, вычислительная техника и управление; 05.25.00 – документальная информация; 08.00.00 – экономические науки.

Журнал входит в базы Web of Science Emerging Sources Citation Index (WoS ESCI) и Russian Science Citation Index на платформе Web of Science (RSCI).

Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор), свидетельство ПИ № ФС77-66609 от 08 августа 2016 г.

Международный стандартный серийный номер (ISSN): 1998-0663 (на русском), 2587-814X (на английском).

Главный редактор: Голосов Алексей Олегович, кандидат технических наук, Президент компании «ФОРС – Центр разработки».

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

ISSN 1998-0663 (print), ISSN 2587-8166 (online)
English version: ISSN 2587–814X (print), ISSN 2587–8158 (online)

Vol. 14 No 2 – 2020

BUSINESS INFORMATICS

HSE SCIENTIFIC JOURNAL

Publisher:
National Research University
Higher School of Economics

**Subscription index
in the Rospechat catalog –
72315**

The journal is published quarterly

The journal is included
into the list of peer reviewed
scientific editions established
by the Supreme Certification
Commission of the Russian Federation

CONTENTS

Address:
28/11, build. 4, Shablovka Street
Moscow 119049, Russia

Tel./fax: +7 (495) 772-9590 *26311
http://bijournal.hse.ru
E-mail: bijournal@hse.ru

Circulation:
English version – 200 copies,
Russian version – 200 copies,
online versions in English and Russian –
open access

Printed in HSE Printing House
3, Kochnovsky Proezd, Moscow,
Russia

ABOUT THE JOURNAL

Business Informatics is a peer reviewed interdisciplinary academic journal published since 2007 by National Research University Higher School of Economics (HSE), Moscow, Russian Federation. The journal is administered by School of Business Informatics. The journal is published quarterly.

The mission of the journal is to develop business informatics as a new field within both information technologies and management. It provides dissemination of latest technical and methodological developments, promotes new competences and provides a framework for discussion in the field of application of modern IT solutions in business, management and economics.

The journal publishes papers in the areas of, but not limited to:

- ✦ automation of management and production processes
- ✦ data analysis and intelligence systems
- ✦ information systems and technologies in business
- ✦ mathematical methods and algorithms of business informatics
- ✦ software engineering
- ✦ internet technologies
- ✦ business processes modeling and analysis
- ✦ standardization, certification, quality, innovations
- ✦ legal aspects of business informatics
- ✦ decision making and business intelligence
- ✦ modeling of social and economic systems
- ✦ information security.

The journal is included into the list of peer reviewed scientific editions established by the Supreme Certification Commission of the Russian Federation.

The journal is included into Web of Science Emerging Sources Citation Index (WoS ESCI) and Russian Science Citation Index on the Web of Science platform (RSCI).

International Standard Serial Number (ISSN): 2587-814X (in English), 1998-0663 (in Russian).

EDITORIAL BOARD

DOI: 10.17323/2587-814X.2020.2.7.20

Трансфертное обучение и доменная адаптация на основе моделирования социально-экономических систем

О.Д. Казаков
E-mail: kod8383@mail.ru

О.В. Михеенко
E-mail: miheenkoov@mail.ru

Брянский государственный инженерно-технологический университет
Адрес: 241037, г. Брянск, пр. Станке Димитрова, д. 3

Аннотация

В статье рассматриваются вопросы применения методов трансфертного обучения (transfer learning) и доменной адаптации (domain adaptation) в рекуррентной нейронной сети, построенной по архитектуре долгой краткосрочной памяти (long short-term memory, LSTM), для повышения эффективности управленческих решений и государственной экономической политики. Обзор существующих в данной области подходов позволяет сделать вывод о необходимости решения ряда практических вопросов повышения качества предиктивной аналитики для задач разработки прогнозов развития социально-экономических систем. В частности, в контексте применения алгоритмов машинного обучения одной из проблем представляется ограниченное количество размеченных данных. Авторами реализовано обучение исходной рекуррентной нейронной сети на синтетических данных, полученных в результате имитационного моделирования, с последующим трансфертным обучением и доменной адаптацией. Для реализации этой цели на основе комбинирования нотаций системной динамики с агентным моделированием в системе AnyLogic разработана имитационная модель, позволяющая исследовать влияние совокупности факторов на ключевые параметры эффективности социально-экономической системы. Обучение исходной LSTM осуществлялось с помощью открытой программной библиотеки для машинного обучения TensorFlow. Предложенный подход позволяет расширить возможности комплексного применения методов имитационного моделирования для построения нейронной сети в целях обоснования параметров развития социально-экономической системы и позволяет получить информацию о ее перспективном состоянии.

Ключевые слова: трансфертное обучение; доменная адаптация; имитационное моделирование; системы поддержки принятия решений; социально-экономическое развитие регионов.

Цитирование: Казаков О.Д., Михеенко О.В. Трансфертное обучение и доменная адаптация на основе моделирования социально-экономических систем // Бизнес-информатика. 2020. Т. 14. № 2. С. 7–20. DOI: 10.17323/2587-814X.2020.2.7.20

Введение

Управление развитием социальных-экономических систем в своей основе опирается на документы, содержащие плановые значения индикаторов по той или иной тематике (стратегия, концепция, прогноз и т.п.). На сегодняшний день управление регионом осуществляется посредством мониторинга путем корректировки плановых значений в соответствии с фактически достигнутыми [1]. Это означает, что в основе будущего развития по большей части заложены показатели прошлых периодов, полученные с существенным опозданием, если принимать во внимание реальную ситуацию с публикацией официальных статистических данных. В связи с этим разработка инструментария для обоснования значений прогнозных экономических параметров, позволяющего с высокой степенью надежности достигать плановых контрольных цифр, является важной научной задачей. Эта задача в конечном счете выступает в качестве объективного условия реализации эффективной экономической политики.

Основу всей совокупности методов социально-экономического прогнозирования традиционно составляют статистические методы, применяемые для построения адекватных моделей временных рядов [2, 3]. Среди наиболее распространенных методов анализа временных рядов можно выделить следующие [4]: регрессионные модели прогнозирования (множественная и нелинейная регрессия), модели экспоненциального сглаживания (ES), модель по выборке максимального подобия (MMSP), модель на цепях Маркова (Markov chains), модель на классификационно-регрессионных деревьях (CART), модель на основе генетического алгоритма (GA), модель на опорных векторах (SVM). Широчайшим и наиболее применимым из классов моделей являются авторегрессионные модели прогнозирования (ARIMAX, GARCH, ARDLM).

В последнее время доказывают свою эффективность методы глубокого машинного обучения, метрики качества которых значительно лучше по сравнению с классическими методами. Однако применение подобных моделей требует наличия огромного объема размеченных данных, получить которые в реальных условиях не всегда представляется возможным. В то же время при прогнозировании большинства показателей, характеризующих социально-экономические системы и процессы, используются статистические данные за один десяток лет и, в лучшем случае, в разрезе месяца. Иначе говоря, на входе имеется всего сотня размеченных записей. Проблема тем или иным образом могла бы решиться в случае применения обучения без учителя, что, к сожалению, на данном этапе развития серийных вычислительных систем не может быть реализовано на практике.

Для решения данной проблемы предлагается использовать рекуррентную нейронную сеть, построенную по архитектуре долгой краткосрочной памяти (long short-term memory, LSTM) и обученную на синтетических данных, полученных в результате имитационного моделирования с последующим трансфертным обучением (transfer learning) и доменной адаптацией (domain adaption). Это позволит, имея реальную статистику за несколько десятков лет, с высокой степенью точности прогнозировать значения экономических параметров с учетом современных векторов развития. Системы поддержки принятия решений, построенные на данных алгоритмах, позволяют наиболее точно обосновывать экономические планы и прогнозы развития территорий и обеспечивать достижение стратегических ориентиров развития.

1. Методы

1.1. Трансфертное обучение и доменная адаптация в LSTM

Основная идея трансфертного обучения состоит в решении поставленной проблемы на основе «готовых данных», полученных в результате решения аналогичных задач. Это означает, что сначала можно обучить нейросеть на большом объеме данных, а впоследствии дообучить ее на конкретном целевом наборе. В этой связи выделяют два основных преимущества использования трансфертного обучения [5]:

✦ значительное снижение времени и затрат в контексте использования соответствующей инфраструктуры для обучения, за счет обучения только определенной части конечной модели;

✦ повышение эффективности конечной модели за счет использования моделей, обученных на доступных данных.

Результаты данного исследования тесно связаны со вторым из перечисленных преимуществ, поскольку в предиктивной аналитике социально-экономических систем это является определяющим фактором.

В качестве доступных данных использовались синтетические данные, полученные в результате

имитационного моделирования. Имитационное моделирование представляет собой экспериментальный способ изучения реальности с помощью компьютерной модели [6]. В имитационных моделях реальные экономические процессы описаны так, как если бы они происходили в действительности [7]. Таким образом, имитационные модели могут применяться для исследования реальных социально-экономических систем с условием, что экономические объекты и процессы заменяются совокупностью математических зависимостей, которые определяют в какое состояние перейдет система из изначально заданного [8].

Веса из обученной на синтетических данных модели, полученные в результате имитационного моделирования, переносятся на новую модель. Для этого авторами использовалась открытая программная библиотека для машинного обучения TensorFlow.

Методы трансфертного обучения и доменной адаптации, как правило, зависят от алгоритмов машинного обучения, используемых для решения поставленных задач [9]. Одним из наиболее эффективных инструментов предиктивной аналитики социально-экономических систем являются рекуррентные нейронные сети с долгой краткосрочной памятью (LSTM-сети). В частности, модели, построенные по архитектуре LSTM, являются очень эффективными для прогнозирования временного ряда – одной из самых распространенных задач в управлении социально-экономическими системами [10]. Следует отметить, что эта эффективность не снижается при прогнозировании нескольких шагов.

Базовая архитектура рекуррентной сети, разработанная еще в 1980-е годы, строится из узлов, каждый из которых соединен со всеми другими узлами. Для обучения с учителем с дискретным временем на входные узлы при каждом очередном шаге подаются данные. При этом прочие узлы (выходные и скрытые) завершают свою активацию и выходные сигналы готовятся для передачи нейронам следующего уровня [11]. Таким образом, рекуррентная сеть с долгосрочной памятью позволяет использовать полученную в прошлом информацию для решения текущих задач. В частности, она дает возможность прогнозирования значений временного ряда, поскольку не использует функцию активации внутри своих рекуррентных компонентов, и хранимое значение не размывается во времени (*рисунок 1*) [12, 13].

Модуль LSTM имеет пять основных компонентов, которые позволяют ему моделировать как долгосрочные, так и краткосрочные данные [13]:

$$\begin{cases} f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_t\right) \\ i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \\ o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ \tilde{c}_t = tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \\ c_t = f_t \odot c_{t-1} + i_t \odot c_t \\ h_t = o_t \odot tanh\left(c_t\right) \end{cases}, \quad (1)$$

где c_t – «состояние ячейки», представляющее ее внутреннюю память, которая хранит как кратковременную, так и долговременную информацию;

h_t – «скрытое состояние»: такая информация о состоянии вывода, которая рассчитана по текущему входу, предыдущему скрытому состоянию и текущему входу ячейки, которые будут использоваться для прогнозирования того или иного временного ряда. Скрытое состояние может принять решение извлечь кратковременную или долгосрочную, либо оба типа информации из сохраненной в c_t;

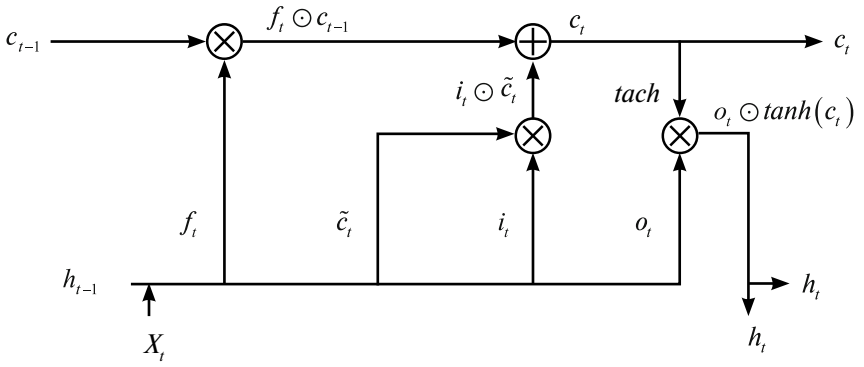


Рис. 1. Архитектура LSTM [12, 13]

```
my_new_model = tf.keras.models.load_model('Kazakov_LSTM.h5')
```

```
for i in zip(my_new_model.layers[0].trainable_weights,
             my_new_model.layers[0].get_weights()):
  print('Parametr %s:\n%s' % (i[0], i[1]))
```

Рис. 2. Листинг «Вывод параметров LSTM»

i_t – «входные ворота»: определяют объем информации поступающей из текущего ввода в c_t;

f_t – «переходные ворота»: определяют объем информации, перетекающей из текущего и предыдущего c_{t-1} вводов в текущий c_t;

o_t – «выходные ворота»: определяют объем информации, попадающей из текущего c_t в скрытое состояние.

Допустим, имеется хорошо работающая модель для предсказательной аналитики временных рядов "Kazakov_LSTM.h5" (процесс ее обучения представлен в следующей разделе статьи). Тогда для просмотра параметров данной модели можно воспользоваться следующими инструкциями (*рисунок 2*).

Таким образом, получим следующий вывод:

```
Parametr lstm_3_W_i:
[[ 0.00075, ...]]
Parametr lstm_3_U_i:
[[ 1.090001, ...]]
Parametr lstm_3_b_i:
[ 0., ...]
Parametr lstm_3_V_c:
[[-1.17770085, ...]]
```
(2)

где *W*-матрицы – матрицы, которые преобразуют входные данные;

U-матрицы – матрицы, которые преобразуют предыдущее скрытое состояние в другое внутреннее значение;

b-векторы – смещение для каждого блока;

V – вектор, определяющий значения, которые следует выводить из нового внутреннего состояния.

Понятие доменной адаптации тесно связано с трансфертным обучением. Суть этой адаптации заключается в обучении модели на данных из домена-источника так, чтобы она показывала сравнимое качество на целевом домене [14]. Домен-источник может представлять собой синтетические данные, которые можно просто сгенерировать в результате прогона соответствующей имитационной модели, а целевой домен – это временной ряд, отражающий динамику тех или иных ключевых показателей социально-экономической системы. Тогда задача доменной адаптации заключается в тренировке модели на синтетических данных, которая будет хорошо работать с реальными объектами.

Этап адаптации домена сводится к замораживанию весов в модели "Kazakov_LSTM.h5" в их предварительно подготовленном состоянии. Веса уровня адаптации домена обучаются на целевом наборе данных. Для этой цели в модель после LSTM добавим полносвязанные (dense) слои.

1.2. Системно-динамическое моделирование индикаторов инновационного развития социально-экономических систем

В целях формирования набора данных в рамках домена-источника построим системно-динамическую модель, позволяющую определить параметры социально-экономической системы, в частности, произвести оценку значений индикаторов инновационного развития регионов. Основным документом, задающим стратегические ориентиры государственной политики в сфере инновационного развития в целях противостояния современным глобальным вызовам и угрозам, является Стратегия инновационного развития Российской Федерации на период до 2020 года [15]. Стратегия определяет долгосрочные приоритеты развития всех субъектов инновационной деятельности, а также устанавливает ряд целевых индикаторов, которые, в соответствии с установкой Правительства страны, должны учитываться при разработке концепций и программ социально-экономического развития России и ее регионов.

Стратегией определены значения целевых индикаторов на 2020 год. При этом 2010 год закреплен как базовый, а 2013 и 2016 годы являются промежуточными контрольными точками. Анализ фактических значений большинства целевых индикаторов за 2016 год выявил общую тенденцию отставания от запланированного уровня.

Поскольку статические службы готовят аналитические данные за отчетный период с существенным временным лагом, серьезной проблемой является то обстоятельство, что государственные власти, ответственные за реализацию Стратегии, пытаются вырабатывать управленческие решения, ориентируясь на неактуальные результаты деятельности. Назвать такой процесс эффективным управлением крайне сложно.

На наш взгляд, наиболее действенным подходом будет обратное движение, когда значение конкретного целевого показателя на определенную дату дифференцируется по субъектам федерации и доводится до региональных властей заблаговременно, в виде рекомендуемых прогнозных значений. В этом случае государственные службы на местах станут непосредственными участниками процесса реализации национальных стратегических инициатив, в том числе, с учетом определенной ответственности за недостижение целевых показателей. Кроме того, появится возможность управлять на основе актуальной карты, отражающей инновационное развитие по долгосрочным задачам Стратегии в разрезе регионов.

Основными компонентами системной модели, позволяющей определить инновационное развитие региона в соответствии с государственной стратегией, являются стратегические задачи по ключевым направлениям. Связь между итоговым показателем инновационного развития и данными компонентами (субиндексами) может быть описана следующим образом:

$$I = \sum_{j=1}^{m} w_j \cdot I_j, \qquad (3)$$

где I_j – значение j-го субиндекса;

m – число субиндексов;

w_j – коэффициент весомости j-го субиндекса.

Субиндексы представляют собой сводные показатели, отражающие формирование первичных индикаторов в рамках решения конкретной стратегической задачи по приоритетным направлениям инновационной деятельности. Число первичных индикаторов по направлениям варьируется от 2 до 12, при этом каждый из них может быть рассмотрен как самостоятельная сложная системно-динамическая модель. Рассмотрим модель формирования частного индикатора первого порядка «Коэффициент изобретательской активности», который определяется в рамках приоритетной стратегической задачи «Инновационный бизнес». Модели остальных индикаторов могут быть сформированы аналогичным образом и не представлены в рамках данной статьи ввиду значительного масштаба проведенного исследования.

Системно-динамическая модель уровня изобретательской активности представлена на *рисунке 3*. Модель основана на определении отношения числа патентных заявок, поданных отечественными изобретателями, к общей численности населения. Количество разработанных и поданных патентных заявок зависит от числа организаций, осуществляющих деятельность в области НИОКР, численности персонала, занимающегося исследованиями и разработками, а также объема внутренних затрат организаций на НИОКР.

Соответствующая математическая модель может быть представлена следующим образом:

$$\begin{aligned}
&\frac{d(\text{Изобретения_разработано})}{dt} = -\,\text{patent}\\
&\frac{d(\text{Полезные_модели_разработано})}{dt} =\\
&\quad = -\,\text{Полезные_модели_подано}\\
&\frac{d(\text{patent})}{dt} = \text{Изобретения_подано} +\\
&\quad + \text{Полезные_модели_подано}\\
&\frac{d(\text{population})}{dt} = \text{Рождение} - \text{Смерть} +\\
&\quad + \text{Миграция} - \text{Эмиграция}
\end{aligned} \qquad (4)$$

$$\begin{aligned}
&\text{Изобретения}\,(\text{Полезные_модели})\,\text{_разработано} =\\
&\quad = \text{corp_research} \times \text{person_research}\\
&\quad \times \text{Внутренние затраты НИОКР}
\end{aligned}$$

$$\begin{aligned}
&\text{Смерть} = \text{Смерть_трудосп_возраст} +\\
&\quad + \text{Смерть_младенческая} + \text{прочее}
\end{aligned}$$

$$\begin{aligned}
&\text{Миграция} = \text{Беженцы} + \text{Временное_убежище} +\\
&+ \text{Вынужденны_переселенцы} + \text{Прибывшие}
\end{aligned}$$

$$\text{coeff\_inv\_activity} = \frac{\text{patent}}{\text{population}} \times 10\,000\,.$$

При первом рассмотрении возникает вопрос целесообразности применения имитационного моделирования для оценки уровня изобретательской активности, поскольку каждая из составляющих его компонент поддается прогнозированию (например, в рамках моделей ARIMA или GARCH, хорошо за-

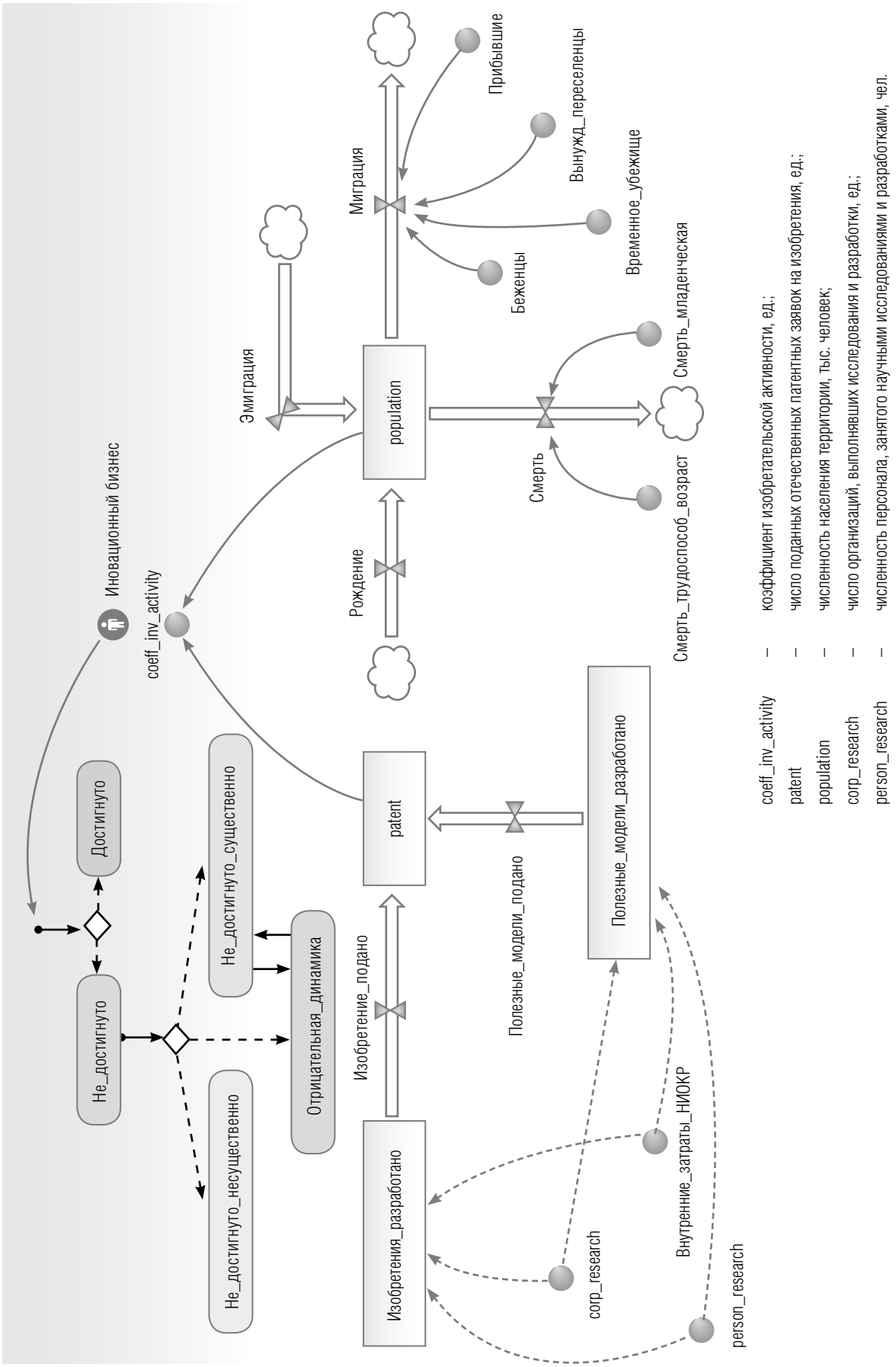


Рис. 3. Системно-динамическая модель уровня изобретательской активности

рекомендовавших себя в сфере прогнозирования демографических и социально-экономических показателей). В действительности зависимость изобретательской активности от многих показателей является стохастической, к тому же на практике по большинству их них собрать достаточный по объему набор значений не представляется возможным. Поэтому имитационное моделирование в данном контексте рассматривается как способ построения модели социально-экономической системы, описывающей сложное поведение объектов и процессов, связанных с управлением инновациями на уровне региона. Данную модель возможно реализовать любое число раз. В этом случае результаты будут обусловлены случайным характером процессов [16]. Используя такие результаты, можно получить устойчивую синтетическую статистику уровня изобретательской активности, которая впоследствии используется для обучения нейронной сети.

2. Эксперимент

2.1. Генерация синтетических данных с помощью системно-динамической модели

Имитационное моделирование было реализовано посредством набора математических инструментальных средств и специального программного обеспечения AnyLogic, позволивших провести целенаправленное моделирование в режиме «имитации» структуры исследуемого индикатора, а также оптимизацию некоторых его параметров [17]. В соответствии с результатами исследования, проведенного группой ученых Казанского технического университета [18], достоверность системы AnyLogic признана удовлетворительной, а в рейтинге аналогичного программного обеспечения данная система входит в тройку лидеров.

Конфигурационные настройки модели, графически описывающей поставленную пользователем проблему в терминах языка AnyLogic, задаются с помощью экспериментов. Дискретно-событийное моделирование реализует возможность аппроксимации реальных процессов дискретными событиями, рассматривающими наиболее важные моменты жизни моделируемой системы [19].

В системе AnyLogic был проведен эксперимент «Варьирование параметров», суть которого состояла в многократном запуске построенной имитационной модели. Для эксперимента была определена доверительная вероятность 0,95 и точность 0,01. Число прогонов модели, рассчитанное по функции Лапласа, составило 9604 [20]. Варьируя разные значения параметров, модель выдавала значение метки в пределах от 1,4725 до 2,1105. Математическое ожидание составило 1,8114, а рассеивание значений относительно математического ожидания – 0,1539, что является допустимым и позволяет сделать вывод об успешной валидации предложенной имитационной модели. Результаты эксперимента представлены *в таблице 1.*

В *таблице 2* представлены синтетические данные, полученные в результате эксперимента «Варьирование параметров» в системе AnyLogic и используемые для обучения первичной нейронной сети.

Исходный набор данных содержит пять функций, изменение во времени которых представлено на *рисунке 4.*

На рисунке видно, что все временные ряды обладают свойством сезонности, но в явном виде этот фактор мы не будем учитывать при дальнейшем обучении исходной сети.

Таблица 1.

Результаты эксперимента «Варьирование параметров» системно-динамической модели уровня изобретательской активности в системе AnyLogic

Наименование	population	patent	corp_research	person_ research	coeff_inv_activity
Изменение параметра	[135600…154235]	[21627…30732]	[3317…4384]	[672493…932115]	[1,4725…2,1105]
Математическое ожидание	144705,3752	26204,7906	3782,2631	778989,9941	1,8114
Дисперсия	6347086,0729	4467795,1931	46155,5052	3873708843	0,0237
Стандартное отклонение	2519,3424	2113,7160	214,8383	62239,1263	0,1539

Таблица 2.

Синтетические данные для обучения исходной нейронной сети

№	Признаки				Метка
	population	patent	corp_research	person_ research	coeff_inv_activity
1	146890	28688	4099	887729	1,950000
2	146841	28362	4098	887553	1,931477
3	146792	28036	4097	887377	1,909913
4	146743	27710	4096	887201	1,888335
5	146694	27384	4095	887025	1,886743
…	…	…	…	…	…
9601	143267	24072	3604	732274	1,732500
9602	146545	29269	4175	738857	2,100000
9603	146804	26795	4032	722291	1,921500
9604	146880	22765	3944	707887	1,627500

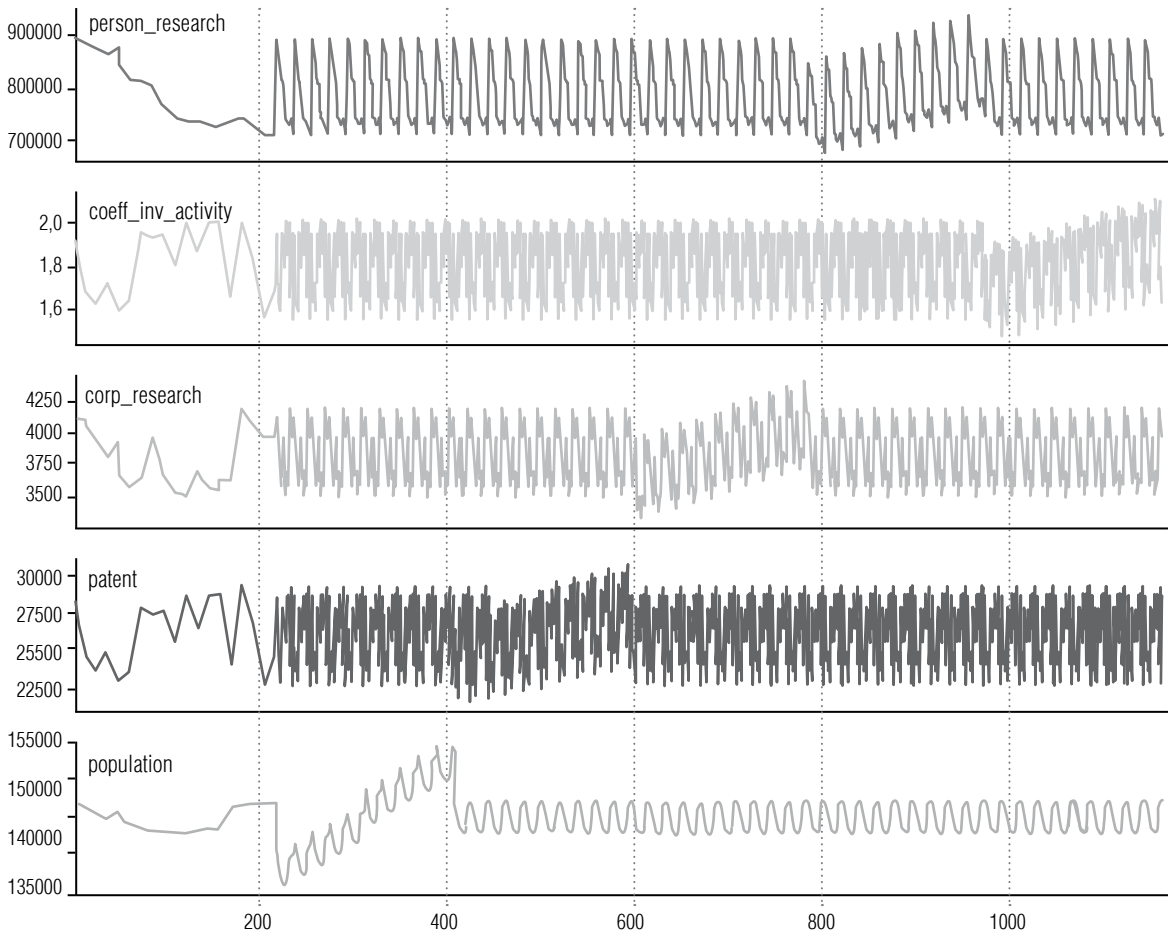


Рис. 4. Изменение во времени исходных функций

2.2. Обучение исходной LSTM и перенос обучения

Как было отмечено выше, обучение исходной LSTM осуществляется на синтетических данных, полученных в результате имитационного моделирования. Для этого используется открытая программная библиотека для машинного обучения TensorFlow, которая предоставляет хороший вспомогательный интерфейс прикладного программирования (RNN API) для реализации предсказательных моделей временных рядов.

Прежде всего, для выполнения процесса обучения загрузим синтетические данные и выполним стандартизацию набора данных, используя функции mean() и std().

Далее задача сводится к прогнозированию многомерного временного ряда на основе некоторой предоставленной истории. Сформируем обучающие и валидационные данные и выполним непосредственное обучение исходной LSTM.

Функция multivariate_data выполняет задачу управления окнами. Она выбирает прошлые наблюдения на основе заданного размера шага (*рисунок 5*). Далее фиксируем веса предварительно обученной нейронной сети (*рисунок 6*). Затем создаем составную нейронную сеть на основе «Kazakov.h5» и компилируем ее (*рисунок 7*).

Для подбора наилучших гиперпараметров нейронной сети использовался оптимизатор Keras Tuner, разработанный командой Google и входящий в открытую библиотеку Keras. В качестве основного типом Keras Tuner был определен RandomSearch. Листинг выбора наилучшей модели с помощью Keras Tuner выглядит следующим образом (*рисунок 8*).

Для создания нейронной сети с перебором основных значений гиперпараметров использовалась следующая функция (*рисунок 9*).

Для двух полносвязанных слоев в качестве функции активации Keras Tuner определил relu – выпрямитель (rectifier) и оптимизатор adam – метод

```
x_train_single, y_train_single = multivariate_data(dataset, dataset[:, 1], 0,
                                                   TRAIN_SPLIT, past_history,
                                                   future_target, STEP,
                                                   single_step=True)
x_val_single, y_val_single = multivariate_data(dataset, dataset[:, 1],
                                               TRAIN_SPLIT, None, past_history,
                                               future_target, STEP,
                                               single_step=True)
```

```
single_step_model = tf.keras.models.Sequential()
single_step_model.add(tf.keras.layers.LSTM(32,
                                           return_sequences=True,
                                           input_shape=x_train_multi.shape[-2:]))
single_step_model.add(tf.keras.layers.LSTM(16, activation='relu'))
single_step_model.add(tf.keras.layers.Dense(1))

single_step_model.compile(optimizer=tf.keras.optimizers.RMSprop(clipvalue=1.0),
                          loss='mae')
```

Рис. 5. Листинг «Обучение исходной LSTM»

```
my_new_model = tf.keras.models.load_model('Kazakov_LSTM.h5')
```

```
my_new_model.trainable = False
```

Рис. 6. Листинг «Назначение весов из предварительно обученной нейронной сети»

```
single_step_model = tf.keras.models.Sequential()
single_step_model.add(my_new_model)
single_step_model.add(tf.keras.layers.Dense(14))
single_step_model.add(tf.keras.layers.Dense(1))
```

Рис. 7. Листинг «Формирование архитектуры составной нейронной сети»

```
[ ]  tuner = RandomSearch(build_model)
     tuner.search(x_train,
                  y_train,
                  epochs=20,
                  verbose = 1)
     models = tuner.get_best_models(num_models=1)
```

Рис. 8. Листинг «Подбор гиперпараметров нейронной сети»

```
def build_model(hp):
    model = Sequential()
    activation_choice = hp.Choice('activation', values=['relu', 'sigmoid', 'tanh'])
    model.add(Dense(units=hp.Int('units_input',
                                 min_value=7,
                                 max_value=30),
                    activation=activation_choice))
    model.add(Dense(1, activation=activation_choice))
    model.compile(
        optimizer=hp.Choice('adam', values=['adam','rmsprop','SGD']),
        loss='mse',
        metrics=['mae'])
    return model
```

Рис. 9. Листинг «Функция подбора гиперпараметров нейронной сети»

стохастического градиентного спуска, основанный на адаптивной оценке моментов первого и второго порядка. В качестве функции потерь представлена среднеквадратичная ошибка (mse), в качестве метрики качества – средняя абсолютная ошибка (mae). На последней эпохе обучения данные параметры приняли значения 0,3995 и 0,1739 соответственно.

Таким образом, были добавлены два полносвязанных слоя для реализации доменной адаптации на основе фактических данных, полученных из официальной статистики и представленных в *таблице 3.*

Составлено на основании данных официального сайт Федеральной службы государственной статистики (https://www.gks.ru/)

Предполагается, что прогнозирование динамики изобретательской активности будет осуществляться на один шаг, поэтому на выходе последнего слоя сети останется один нейрон.

3. Обсуждение результатов

Разработанная имитационная модель динамики изобретательской активности позволяет формировать потенциально неограниченное количество записей для обучения исходной сети. Исследуемые методики трансфертного обучения и доменной адаптации в LSTM позволили использовать предобученную исходную сеть в новой смешанной архитектуре. Таким образом, несмотря на имеющийся критически малый набор фактических данных для

Таблица 3.

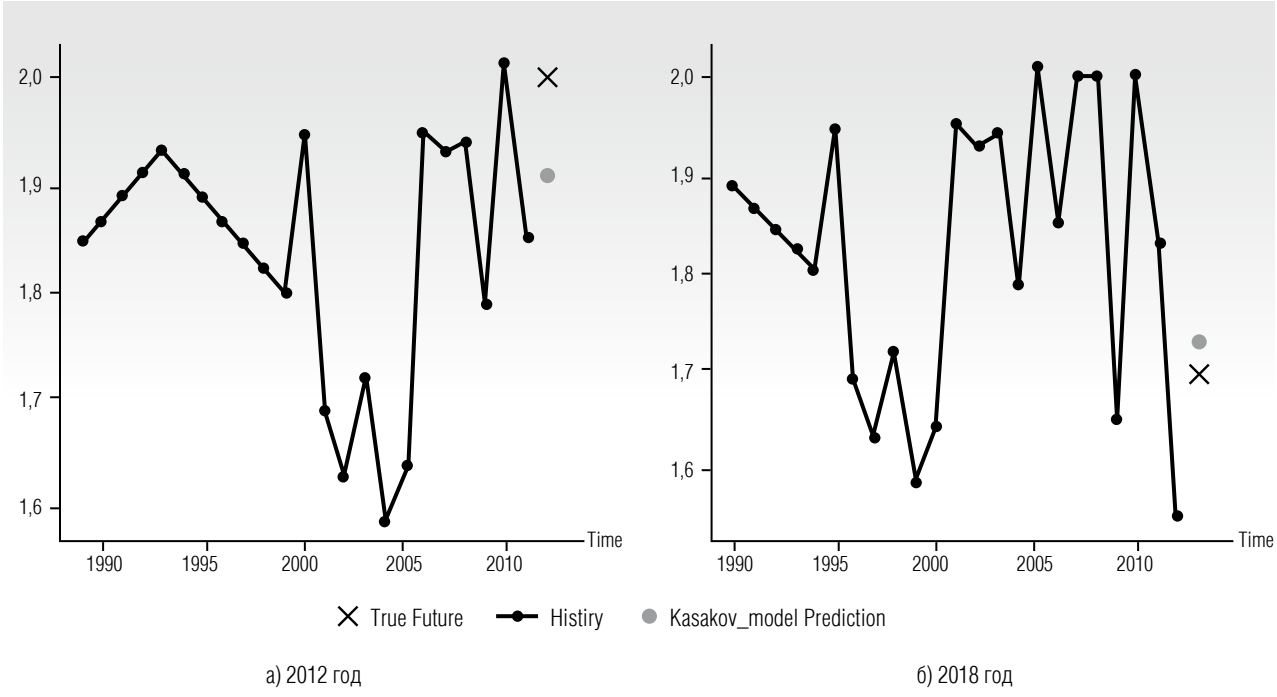
Фактические данные для реализации доменной адаптации

№ п/п	Год	Признаки				Метка
		population	patent	corp_research	person_ research	coeff_inv_activity
0	2001	146304	24777,0	4037	885568	1,69
1	2002	145649	23712,0	3906	870878	1,63
2	2003	144964	24969,0	3797	858470	1,72
…	…	…	…	…	…	…
15	2016	146804	26795,0	4032	722291	1,83
16	2017	146880	22765,0	3944	707887	1,55
17	2018	146781	24952,8	3944	707887	1,70

обучения нейронной сети, получена возможность осуществлять прогноз экономических показателей.

С помощью обученной нейронной сети визуализируем прогнозные значения коэффициента инновационной активности на 2012 год (на основе данных для валидационной выборки) и на 2018 год (на основе данных для тестовой выборки) (*рисунок 10*).

Данные, полученные на 2012 год, показали значение метки 1,91 при фактически зафиксированном в данный период значении показателя 2,00 (*рисунок 10а*). Коэффициент изобретательской активности, определенный сетью на 2018 год, равен 1,73 против фактического 1,70 (*рисунок 10б*). Таким образом, отклонение составило 4,5% в 2012 году и 1,8% в 2018 году соответственно. Полученные результаты можно считать удовлетворительными, что позволяет транслировать данный метод на перспективу.



а) 2012 год

б) 2018 год

Рис. 10. Определение значения целевой переменной с помощью обученной нейронной сети

Заключение

Представленный в исследовании подход, базирующийся на построении системно-динамической модели и рекуррентной нейронной сети может быть адаптирован к другим социально-экономическим системам и процессам в части решения задач предиктивной аналитики. Авторский подход к обучению и использованию LSTM-сетей в социально-экономических системах позволит существенно повысить эффективность управленческих решений. Несомненным преимуществом применения данной методики, на наш взгляд, является возможность раннего определения трендов в процессах даже в условиях ограниченного набора данных.

Предложенный подход может стать универсальным инструментом LSTM предиктивной аналитики, поскольку исследуемые методики трансфертного обучения и доменной адаптации в LSTM позволили использовать исходную сеть, обученную на синтетических данных, и с высокой степенью точности прогнозировать значение целевой переменной. Практическая значимость исследования состоит в расширении возможностей комплексного применения методов имитационного моделирования для построения нейронной сети. Вместе с тем разработанный подход может быть использован органами государственной власти для обоснования параметров развития социально-экономической системы и позволяет получить информацию о ее перспективном состоянии. ■

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-41-320003 «Математическое моделирование социально-экономического развития региона в системах поддержки принятия решений с использованием адаптивных методов машинного обучения и имитационного моделирования в условиях неопределенности».

Литература

1. Региональное и муниципальное управление социально-экономическим развитием в Сибирском федеральном округе / под ред. А.С. Новоселова. Новосибирск: ИЭОПП СО РАН, 2014.
2. Канторович Г.Г. Анализ временных рядов // Экономический журнал. 2002. № 1. С. 87–110.
3. Андерсон Т. Статистический анализ временных рядов. Москва: Мир, 1976.
4. Обзор моделей прогнозирования временных рядов: проба пера / Сообщество IT-специалистов. [Электронный ресурс]: https://habr.com/ru/post/180409/ (дата обращения: 15.03.2020).
5. Pan S.J., Yang Q. A survey on transfer learning // IEEE Transactions on Knowledge and Data Engineering. 2010. Vol. 22. No 10. P. 1345–1359. DOI: 10.1109/TKDE.2009.191.
6. Царегородцев Е.И., Баркалова Т.Г. Имитационное моделирование в прогнозировании социально-экономических систем // Вестник ТИСБИ. 2017. № 3. С. 126–134.
7. Манкаев Н.В. Исследование и моделирование процесса управления социально-экономическими системами // Мягкие измерения и вычисления. 2019. № 1 (14). С. 21–30.
8. Звягин Л.С. Практические приемы моделирования экономических систем // Материалы IV Международной научной конференции «Проблемы современной экономики». Челябинск, 20–23 февраля 2015 г. С. 14–19.
9. Domain adaptation with latent semantic association for named entity recognition / H. Guo [et al.] // Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL – 2009). Boulder, Colorado, USA, 31 May – 5 June 2009. P. 281–289. DOI: 10.3115/1620754.1620795.
10. Cortes C., Mohri M. Domain adaptation and sample bias correction theory and algorithm for regression // Theoretical Computer Science. 2014. No 519. P. 103–126. DOI: 10.1016/j.tcs.2013.09.027.
11. Гафаров Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения. Казань: Казанский университет, 2018.
12. Olah C. Understanding LSTM networks // GITHUB blog. 27 August 2015. [Электронный ресурс]: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (дата обращения: 12.03.2020).
13. Ganegedara T. Stock market predictions with LSTM in Python // GITHUB blog. 3 May 2018. [Электронный ресурс]: https://www.datacamp.com/community/tutorials/lstm-python-stock-market (дата обращения: 15.03.2020).
14. Кондрашова Д.А., Насыров Р.В. Сравнение эффективности методов автоматической классификации текстов // Труды VII Всероссийской научной конференция «Информационные технологии интеллектуальной поддержки принятия решений». Уфа, 28–30 мая 2019 г. С. 146–149.
15. О Стратегии инновационного развития РФ на период до 2020 г. Распоряжение Правительства РФ от 8 декабря 2011 г. № 2227-р. [Электронный ресурс]: https://www.garant.ru/products/ipo/prime/doc/70006124/#review (дата обращения: 27.11.2019).
16. Создание эксперимента Монте-Карло. [Электронный ресурс]: https://studme.org/286158/informatika/sozdanie_eksperimenta_monte_karlo (дата обращения: 27.11.2019).
17. Звягин Л.С. Ключевые аспекты имитационного моделирования сложных систем // Молодой ученый. 2016. №12. С. 19–23.
18. Дровянников В.И., Хаймович И.Н. Имитационное моделирование управления социальным кластером в системе AnyLogic // Фундаментальные исследования. 2015. № 8–2. С. 361–366.

19. Якимов И.М., Кирпичников А.П., Исаева Ю.Г., Аляутдинова Г.Р. Сравнение результатов имитационного моделирования вероятностных объектов в системах: AnyLogic, Arena, Bizagi modeler, GPSS W // Вестник технологического университета. 2015. Т. 18. №. 16. С. 260–264.

20. Gе́ron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. Sebastopol, CA: O'Reilly Media, 2017.

Об авторах

Казаков Олег Дмитриевич

кандидат экономических наук, доцент;

заведующий кафедрой информационных технологий,
Брянский государственный инженерно-технологический университет,
241037, г. Брянск, пр. Станке Димитрова, д.3;

E-mail: kod8383@mail.ru;

ORCID: 0000-0001-9665-8138

Михеенко Ольга Валерьевна

кандидат экономических наук;

доцент кафедры государственного управления, экономической и информационной безопасности,
Брянский государственный инженерно-технологический университет,
241037, г. Брянск, пр. Станке Димитрова, д.3;

E-mail: miheenkoov@mail.ru;

ORCID: 0000-0003-0917-8406

Transfer learning and domain adaptation based on modeling of socio-economic systems

Oleg D. Kazakov
E-mail: kod8383@mail.ru

Olga V. Mikheenko
E-mail: miheenkoov@mail.ru

Bryansk State Technological University of Engineering
Address: 3, Stanke Dimitrov Avenue, Bryansk 241037, Russia

Abstract

This article deals with the application of transfer learning methods and domain adaptation in a recurrent neural network based on the long short-term memory architecture (LSTM) to improve the efficiency of management decisions and state economic policy. Review of existing approaches in this area allows us to draw a conclusion about the need to solve a number of practical issues of improving the quality of predictive analytics for preparing forecasts of the development of socio-economic systems. In particular, in the context of applying machine learning algorithms, one of the problems is the limited number of marked data. The authors have implemented training of the original recurrent neural network on synthetic data obtained as a result of simulation, followed by transfer training and domain adaptation. To achieve this goal, a simulation model was developed by combining notations of system dynamics with agent-based modeling in the AnyLogic system, which allows us to investigate the influence of a combination of factors on the key parameters of the efficiency of the socio-economic system. The original LSTM training was realized with the help of TensorFlow, an open source software library for machine learning. The suggested approach makes it possible to expand the possibilities of complex application of simulation methods for building a neural network in order to justify the parameters of the development of the socio-economic system and allows us to get information about its future state.

Key words: transfer learning; domain adaptation, simulation modeling; decision support systems; socio-economic development of regions.

Citation: Kazakov O.D., Mikheenko O.V. (2020) Transfer learning and domain adaptation based on modeling of socio-economic systems. *Business Informatics*, vol. 14, no 2, pp. 7–20. DOI: 10.17323/2587-814X.2020.2.7.20

References

1. Novoselov A.S., ed. (2014) *Regional and municipal management of socio-economic development in the Siberian Federal district*. Novosibirsk: IEIE SB RAS (in Russian).
2. Kantorovich G.G. (2002) Time series analysis. *Economic Journal*, no 1, pp. 87–110 (in Russian).
3. Anderson T. (1976) *Statistical analysis of time series*. Moscow: Mir (in Russian).
4. Community of IT specialists (2013) *Overview of time series forecasting models*. Available at: https://habr.com/ru/post/180409/ (accessed: 15 March 2020) (in Russian).
5. Pan S.J., Yang Q. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no 10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.
6. Tsaregorodtsev E.I., Barkalova T.G. (2017) Simulation modeling in forecasting of socio-economic systems. *Herald of TISBI*, no 3, pp. 126–134 (in Russian).
7. Mankaev N.V. (2019) Research and modeling of the process of socio-economic systems management. *Soft Measurements and Computing*, no 1 (14), pp. 21–30 (in Russian).
8. Zvyagin L.S. (2015) Practical methods of modeling economic systems. Proceedings of the *IV International Scientific Conference on Problems of the Modern Economy. Chelyabinsk, 20–23 February 2015*, pp. 14–19 (in Russian).
9. Guo H., Zhu H., Guo Z., Zhang X., Wu X., Su Z. (2009) Domain adaptation with latent semantic association for named entity recognition. Proceedings of the *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL – 2009). Boulder, Colorado, USA, 31 May – 5 June 2009*, pp. 281–289. DOI: 10.3115/1620754.1620795.
10. Cortes C., Mohri M. (2014) Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, no 519, pp. 103–126. DOI: 10.1016/j.tcs.2013.09.027.
11. Gafarov F.M., Galimyanov A.F. (2018) *Artificial neural networks and applications*. Kazan: Kazan University (in Russian).
12. Olah C. (2015) *Understanding LSTM networks. GITHUB blog*. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed 12 March 2020).
13. Ganegedara T. (2018) *Stock market predictions with LSTM in Python*. GITHUB blog. Available at: https://www.datacamp.com/community/tutorials/lstm-python-stock-market (accessed 12 March 2020).
14. Kondrashova D.A., Nasyrov R.V. (2019) Comparison of the effectiveness of automatic text classification methods. Proceedings of the *VII All-Russian Scientific Conference on Information Technologies of Intellectual Decision Support. Ufa, 28–30 May 2019*, pp. 146–149 (in Russian).
15. Decree of the Government of the Russian Federation No 2227-R of 8 December 2011. *About the strategy of innovative development of the Russian Federation for the period up to 2020*. Available at: https://www.garant.ru/products/ipo/prime/doc/70006124/#review (accessed: 27 November 2019) (in Russian).
16. *Creating the Monte Carlo experiment*. Available at: https://studme.org/286158/informatika/sozdanie_eksperimenta_monte_karlo (accessed: 27 November 2019) (in Russian).
17. Zvyagin L.S. (2016) Key aspects of complex systems simulation. *Young Scientist*, no 12, pp. 19–23 (in Russian).
18. Drovyannikov V.I., Khaimovich I.N. (2015) Simulation of social cluster management in the AnyLogic system. *Fundamental Study*, no 8–2, pp. 361–366 (in Russian).
19. Yakimov I.M., Kirpichnikov A.P., Isaeva Yu.G., Alyautdinova G.R. (2015) Comparison of simulation results for probabilistic objects in the systems: AnyLogic, Arena, Bizagi modeler, GPSS W. *Bulletin of the Technological University*, vol. 18, no 16, pp. 260–264 (in Russian).
20. Géron A. (2017) *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media.

About the authors

Oleg D. Kazakov

Cand. Sci. (Econ.), Associate Professor;

Head of Department of Information Technology, Bryansk State Technological University of Engineering, 3, Stanke Dimitrov Avenue, Bryansk 241037, Russia;

E-mail: kod8383@mail.ru;

ORCID: 0000-0001-9665-8138

Olga V. Mikheenko

Cand. Sci. (Econ.);

Associate Professor, Department of Public Administration, Economic and Information Security, Bryansk State Technological University of Engineering, 3, Stanke Dimitrov Avenue, Bryansk 241037, Russia;

E-mail: miheenkoov@mail.ru;

ORCID: 0000-0003-0917-8406

DOI: 10.17323/2587-814X.2020.2.21.35

Моделирование и оптимизация планов грузовых железнодорожных перевозок, выполняемых транспортным оператором

Ф.А. Белоусов
E-mail: sky_tt@list.ru

И.В. Неволин
E-mail: i.nevolin@cemi.rssi.ru

Н.К. Хачатрян
E-mail: nerses@cemi.rssi.ru

Центральный экономико-математический институт Российской академии наук
Адрес: 117418, г. Москва, Нахимовский проспект, д. 47

Аннотация

В работе предложен один из подходов для решения задачи, возникающей перед операторами железнодорожного транспорта. Задача состоит в оптимальном с точки зрения максимизации прибыли управлении парком грузовых железнодорожных вагонов. Исходными данными для транспортного оператора являются список заявок, поступающих от заказчиков, и местоположение вагонов к началу планового периода. Заявка, сформированная заказчиком, содержит станцию отправления, станцию назначения, а также наименование и объем груза, который заказчик хотел бы перевезти. К заявке добавляется ставка, которую заказчик платит транспортному оператору за каждый перевезенный вагон груза. Планирование осуществляется на месяц вперед и заключается, с одной стороны, в выборе наиболее выгодных к исполнению заявок, с другой стороны – в построении такой последовательности грузовых и порожних перегонов, которые исполнят выбранные заявки с наибольшей эффективностью. Непосредственная транспортировка грузовых и порожних вагонов осуществляется силами РЖД с заранее известными тарифами и временными нормативами движения по каждому из маршрутов. Тарифы на грузовые перегоны являются дополнительными издержками заказчика, указанного в заявке маршрута (заказчики платят как транспортному оператору за использование вагонов, так и РЖД). При этом транспортировку порожних вагонов оплачивают транспортные операторы. Для решения поставленной задачи предложен один из возможных способов сведения данной задачи к задаче линейного программирования большой размерности. Предложен алгоритм, результатом выполнения которого является задача, записанная в виде задачи линейного программирования. Для наглядности демонстрации подхода рассматривается упрощенная постановка, учитывающая лишь основные факторы моделируемого процесса. Также в работе продемонстрирован пример численного решения поставленной задачи на основе простых модельных данных.

Ключевые слова: железнодорожные грузоперевозки; планирование железнодорожных грузоперевозок; оптимальный план железнодорожных грузоперевозок; оптимальное управление парком вагонов; линейное программирование; теория расписаний; исследование операций; беспилотные транспортные средства.

Цитирование: Белоусов Ф.А., Неволин И.В., Хачатрян Н.К. Моделирование и оптимизация планов грузовых железнодорожных перевозок, выполняемых транспортным оператором // Бизнес-информатика. 2020. Т. 14. № 2. С. 21–35. DOI: 10.17323/2587-814X.2020.2.21.35

Введение

В математической области, которая носит название «исследование операций», и в ее подразделе, – теории расписаний, – существует множество задач, связанных с оптимизацией управления железнодорожным транспортом. Целую серию таких задач, а также их классификацию можно найти в работах [1–5]. Среди задач составления расписаний железнодорожного транспорта отдельно можно выделить класс задач, связанных с составлением расписаний для пассажирского транспорта. Примерами таких моделей являются работы [6–9].

В данной статье рассматривается сфера грузовых железнодорожных перевозок. Одной из актуальных задач, возникающих в этой сфере, является организация процесса грузоперевозок. Ей, в частности, посвящены работы [10–13]. В них описаны и исследованы динамические модели организации грузовых железнодорожных перевозок как между двумя узловыми станциями, так и на замкнутой цепочке станций. В данной работе будет рассмотрена не менее актуальная задача – управление парком грузовых железнодорожных вагонов.

Эта задача возникает перед операторами железнодорожного транспорта (далее – транспортными операторами), управляющими парком грузовых железнодорожных вагонов в коммерческих целях. В зависимости от особенностей регулирования и особенностей рынка, для каждого конкретного региона могут быть построены свои модели, учитывающие ту или иную специфику. В качестве примера можно рассмотреть работу [14], в которой представлена модель, применяемая одним из крупнейших железнодорожных транспортных операторов на территории Латинской Америки. Другим примером является работа [15], в которой рассматривается сразу несколько моделей оптимизации доставки грузов швейцарской железнодорожной грузовой компанией Cargo Express Service of Swiss Federal Railways. В работах [16, 17] рассматриваются модели, спроектированные с учетом особенностей рынка грузовых перевозок в Италии. В работах [18, 19] представлены модели минимизации издержек транспортировки грузов по железнодорожной сети, покрывающей несколько европейских стран.

Также существуют модели, созданные для российского рынка железнодорожных перевозок, пример такой модели представлен в работах [20, 21]. В данной работе постановка задачи схожа с постановкой из [20, 21], а основные отличия состоят в методах решения поставленной задачи. Если в работах [20, 21] решения ищутся с помощью метода генерации колонок [22], а также с помощью модифицированного метода генерации колонок [23], то в данной работе поиск решения осуществляется на основе поиска оптимальных потоков в сети, покрывающей все возможные маршруты, путем сведения к задаче линейного программирования большой размерности.

Иначе говоря, если в методе генерации колонок и обобщенном методе генерации колонок поиск решения осуществляется итерационно (при этом на каждой итерации решается задача линейного программирования на подмножестве всех возможных маршрутов) то в данной работе решается одна задача линейного программирования достаточно большой размерности на множестве всех возможных маршрутов. Перед тем как получить такую задачу линейного программирования большой размерности необходимо построить сеть всех возможных маршрутов (в нашем случае строится более широкая сеть, которая включает в себя сеть всех возможных маршрутов). После этого задача может быть формализована как задача линейного программирования.

Сеть всех возможных маршрутов представляет из себя ориентированный пространственно-временной граф без циклов. Такой граф строится с фиксированным временным шагом, за шаг берется один день. Каждая вершина графа содержит информацию о количестве вагонов на определенной станции s в некоторый определенный день t планового периода. Каждое ребро этого графа характеризует маршрут отправлением из некоторой станции s_i в день t_i с прибытием на другую станцию s_j в день t_j. При этом разница между t_i и t_j ($t_i < t_j$)соответствует тому количеству дней, которое требуется, чтобы перевести вагоны из станции s_i в станцию s_j в соответствии с известными временными нормативами движений грузов по железной дороге. В такой интерпретации задача состоит в поиске потоков в построенном графе, которые суммарно обеспечивают максимальный выигрыш. Под потоком подразумевается цепочка грузовых и порожних маршрутов, а получаемый выигрыш соответствует прибыли за расчетный период. Более подробно построение подобных моделей и сведение их к задачам линейного программирования представлено в книгах [24–26].

Данный подход, заключающийся в решении задачи линейного программирования большой раз-

мерности, требует большего времени для поиска решений по сравнению с подходом, связанным с генерацией колонок. Преимуществом предложенного подхода является поиск оптимального плана перевозок на множестве всевозможных маршрутов, тогда как в методах, связанных с генерацией колонок, решается серия задач линейного программирования на подмножествах множества всех маршрутов. Как следствие, решение, полученное с помощью метода генерации колонок, может отличаться от оптимального. На практике в этом случае упущенная выгода транспортного оператора может измеряться десятками миллионов рублей в месяц.

1. Общая постановка задачи

Рассматривается задача управления парком грузовых вагонов транспортным оператором, целью которого является максимизация прибыли. Составление плана осуществляется на месяц вперед в момент, когда известна вся необходимая для этого информация. Для осуществления планирования необходима информация о начальном расположении вагонов в следующем месяце, а также готовый список заявок на транспортировку грузов в следующем месяце. Под начальным расположением вагонов в плановом месяце подразумевается то, на какой станции и в какой день планового месяца окажется каждый из вагонов, отправленный в предыдущем месяце. Список заявок состоит из запросов со стороны клиентов, в каждом из которых указывается груз, его объем (в вагонах), пункт отправления и пункт назначения. Также в заявке указывается ставка, которую заказчик заплатит за каждый вагон перевезенного груза. Допускается ситуация, при которой начало исполнения заявки будет иметь место в плановом месяце, но ее завершение окажется в месяце, следующим за плановым. В этом случае вся прибыль, связанная с исполнением такой заявки, будет учтена в плановом месяце. Предполагается, что заказчику неважно, в какой именно день планового месяца будет исполнен его заказ: если оператор берется за выполнение данного заказа, то он будет исполнен в наиболее удобный для транспортного оператора день (или несколько дней, если заявка будет исполняться несколькими рейсами). Транспортный оператор не обязан выполнять все поступающие заявки: как правило, он физически не может этого сделать за отведенный месяц. Поэтому оператор может либо выполнить заявку полностью, либо исполнить ее частично, либо отклонить ее. Таким образом, при создании плана на предстоящий месяц задача транспортного оператора заключается, во-первых, в выборе тех заявок, которые наиболее выгодны к исполнению, во-вторых, в выборе таких цепочек грузовых и порожних маршрутов, которые с наибольшей эффективностью обеспечат выполнение выбранных заявок.

Непосредственная транспортировка вагонов осуществляется силами Российских железных дорог (ОАО «РЖД»), которые устанавливают свои тарифы на перегоны как порожних, так и груженых рейсов. Также заранее известны нормативы по времени таких перегонов по каждому из возможных маршрутов. В рамках модели будем считать, что тарифы не зависят от количества вагонов, перевозимых одним рейсом. Если заявка заказчика исполнена, то он, помимо платежа оператору в соответствии с указанной ставкой за использование его вагонов, отдельно платит РЖД за транспортировку этих вагонов. Перемещение порожних вагонов силами РЖД является статьей расходов транспортного оператора. Поскольку тарифы РЖД на транспортировку грузовых вагонов являются издержками заказчиков (транспортные операторы к ним никакого отношения не имеют), в модели эти тарифы отсутствуют. В данной работе рассматривается упрощенная постановка, в рамках которой транспортный оператор управляет одним типом грузовых вагонов, а горизонт планирования составляет один месяц.

Дополнительно будем предполагать, что значительная часть непосредственной транспортировки силами РЖД осуществляется беспилотными транспортными средствами – беспилотными локомотивами. Их использование позволит снизить влияние человеческого фактора, который зачастую приводит к тем или иным аварийным ситуациям, что в свою очередь влечет за собой сбои в расписаниях. Данное предположение об использовании беспилотных транспортных средств дает возможность рассматривать детерминированную модель, без учета стохастических компонент. В противном случае в рамках модели необходим учет случайных факторов, что неизбежно привело бы к значительному ее усложнению.

2. Математическая постановка задачи

В данном разделе приведена математическая постановка в некотором промежуточном формате, который на следующем этапе преобразуется в формат задачи линейного программирования в яв-

ном виде. Преимущество такого промежуточного формата состоит в том, что он гораздо более нагляден и удобен для понимания сути предлагаемого подхода. Схожий формат математической постановки данной задачи можно найти в работах [3, 4], однако в этих работах запись математической постановки имеет достаточно громоздкую форму и с трудом может быть применима на практике. В данной работе для облегчения изложения представлен наиболее простой вариант постановки задачи.

Введем ряд обозначений:

N – количество станций, участвующих в планировании;

T – горизонт планирования, который измеряется в днях. Для упрощения в данной работе за горизонт планирования принят один месяц (т.е. $T = 30$ или 31). Однако на практике правильнее рассматривать более длительный горизонт планирования, например, два или более месяцев, но для простоты изложения в данной работе принимается одновременно короткий и правдоподобный промежуток;

t – дискретный параметр, характеризующий время. Он измеряется в днях и принимает значения $t = 1, 2, ... T$;

$C = \left\{C_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют тарифы, установленные РЖД за порожний перегон одного вагона от станции i до станции j;

$\Theta 1 = \left\{\Theta 1_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют время (в днях) движения груженого вагона от станции i до станции j в соответствии с нормативами РЖД (время округляется до большего целого числа);

$\Theta 2 = \left\{\Theta 2_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют время (в днях) движения порожнего вагона от станции i до станции j в соответствии с нормативами РЖД (время округляется до большего целого числа). Отметим, что диагональные элементы данной матрицы принимаются равными единице. Это означает, что если вагон остается на станции до следующего дня, то это эквивалентно тому, что он отправляется в рейс-петлю длительностью в один день, где станции отправления и назначения совпадают;

$P = \left\{P_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют ставку, указанную заказчиком в заявке на транспортировку одного вагона груза от станции i до станции j;

$\bar{Q} = \left\{\bar{Q}_{ij}\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют количество грузовых вагонов, указанное в соответствующей заявке на транспортировку груза от станции i до станции j. Все элементы матрицы принимают неотрицательные целочисленные значения;

$\bar{S}^0(t) = \left\{\bar{S}_i^0(t)\right\}_{i=1}^{N}$ – вектор размерности N, характеризующий начальное расположение вагонов в день t; i-й элемент данного вектора равен количеству вагонов, отправка которых была осуществлена в предыдущем месяце, прибывших на станцию i в момент времени $t \in \{1, ..., T\}$. Все значения данного вектора принимают неотрицательные целочисленные значения.

План перевозок характеризуется следующими матрицами:

$K1(t) = \left\{K1_{ij}(t)\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют количество отправляемых груженых вагонов от станции i до станции j на момент времени $t \in \{1, ..., T\}$. Все элементы матрицы принимают неотрицательные целочисленные значения;

$K2(t) = \left\{K2_{ij}(t)\right\}_{i,j=1}^{N}$ – $(N \times N)$-матрица, элементы которой характеризуют количество порожних вагонов, отправляемых от станции i до станции j на момент времени $t \in \{1, ..., T\}$. Все элементы матрицы принимают неотрицательные целочисленные значения.

Обозначим через $K1$ и $K1$ набор соответствующих матриц для всех моментов времени $t \in \{1, ..., T\}$, иначе говоря, $K1 = \left\{K1(t)\right\}_{t=1}^{T}$, $K2 = \left\{K2(t)\right\}_{t=1}^{T}$.

Также интерес представляет итоговое распределение вагонов по станциям и по времени в плановый месяц в соответствии с предложенным планом $K1$, $K2$. Для этого в рассмотрение вводится набор векторов $\left\{\bar{S}_i(t, K1, K2)\right\}_{i=1}^{N}$.

$\bar{S}(t, K1, K2) = \left\{\bar{S}_i(t, K1, K2)\right\}_{i=1}^{N}$ – вектор длинной N, каждый элемент которого характеризует количество вагонов на станции i в момент времени $t \in \{1, ..., T\}$, которое реализуется в соответствии с предложенным планом $K1$ и $K1$ и начальным распределением вагонов $\bar{S}^0(t)$. Нетрудно заметить, что для всех $t \in \{1, ..., T\}$ и любой станции значение элемента $i \in \{1, ..., N\}$ определяется формулой:

$$\bar{S}_i(t, K1, K2) = \bar{S}_i^0(t) + \sum_{\tau=1}^{t}\left(\sum_{j=1}^{N} K1_{ji}(\tau) I\left(t - \tau - \Theta 1_{ji}\right) + \sum_{j=1}^{N} K2_{ji}(\tau) I\left(t - \tau - \Theta 2_{ji}\right)\right), \quad (1)$$

где функция $I(\cdot)$ определяется по правилу

$$I(x)=\begin{cases}1, \text{ если } x=0;\\ 0, \quad \text{иначе.}\end{cases}$$

Иначе говоря, количество вагонов, которое наблюдается на станции i в момент времени t, равно количеству прибывших из предыдущего месяца вагонов в соответствии со значением $\overline{S}_i^0(t)$, а также тому количеству вагонов, которое было направлено на станцию i порожними либо груженными рейсами в дни τ, предшествующие текущему дню t ($\tau \in \{1, ..., t-1\}$), и которые прибывают на станцию i именно в этот день t.

Теперь можно сформулировать математическую постановку данной задачи. Критерием максимизации является прибыль

$$\sum_{t=1}^{T}\left(\sum_{i,j=1}^{N}P_{ij}K1_{ij}(t)-\sum_{i,j=1}^{N}C_{ij}K2_{ij}(t)\right)\to\max_{\{K1_{ij}(t),K2_{ij}(t)\}}. \quad (2)$$

При этом должны выполняться следующие ограничения:

$$\overline{S}_i(t,K1,K2)=\sum_{j=1}^{N}\left(K1_{ij}(t)+K2_{ij}(t)\right),\quad i=\overline{1,N},\ t=\overline{1,T}; \quad (3)$$

$$\sum_{t=1}^{T}K1_{ij}(t)\le\overline{Q}_{ij},\quad i=\overline{1,N},\ j=\overline{1,N}; \quad (4)$$

$$K1_{ij}(t)\in\mathbb{N}\cup\{0\},\ K2_{ij}(t)\in\mathbb{N}\cup\{0\},\quad i=\overline{1,N},\ j=\overline{1,N},\ t=\overline{1,T}. \quad (5)$$

Целевая функция (2) представляет собой прибыль от всех грузовых перегонов за вычетом издержек, связанных с перегоном порожних вагонов. Оптимизация осуществляется за счет управления груженными рейсами $K1_{ij}(t)$ и порожними рейсами $K2_{ij}(t)$. Ограничение (3) является балансовым ограничением и означает, что количество вагонов, отправляемых со станции i в момент времени t, в точности равно количеству вагонов, которые в этот день туда прибыли. При этом количество прибывающих вагонов $\overline{S}_i(t, K1, K2)$ определяется в соответствии с формулой (1). Ограничение (4) говорит о том, что количество грузовых вагонов, отправленных со станции i на станцию j во все дни расчетного периода, не должно превышать того количества, которое указано в соответствующей заявке.

Задача (2)–(5) может быть решена напрямую методами целочисленного программирования (Mixed-Integer Programming). Однако в силу большой размерности задача поиска целочисленного решения может оказаться невыполнимой в пределах разумного времени. Поэтому в таких случаях вместо исходной задачи рассматривают ее линейную релаксацию. Это означает, что вместо условия (5) рассматривается следующее, более слабое условие:

$$K1_{ij}(t)\ge 0,\ K1_{ij}(t)\ge 0,\ i=\overline{1,N},\ t=\overline{1,T} \quad (6)$$

Таким образом, вместо задачи (2)–(5) далее рассматривается задача (2), (3), (4), (6). Результатом решения такой задачи может оказаться дробное решение, которое, очевидно, не может быть применено на практике. В этом случае, для получения целочисленного решения к полученному дробному решению необходимо применить методы округления.

На следующем этапе переформулируем эту задачу в виде задачи линейного программирования.

3. Сведение исходной задачи к задаче линейного программирования

Для того чтобы решить сформулированную задачу с помощью программных средств, приведенные выше обозначения необходимо представить в ином формате. Для этого матрицы P, C, $\overline{Q}$, а также систему матриц $K1(t)$ и $K2(t)$, $t\in\{1, ..., T\}$ необходимо преобразовать в длинные вектора. Системы векторов $\overline{S}^0$ и $\overline{S}(t, K1, K2)$, $t \in \{1, ..., T\}$ также преобразуются в два единых вектора. С одной стороны, матрицы P и C по специальному правилу трансформируются в один вектор PC, с другой стороны, по похожему правилу матрицы $K1(t)$ и $K2(t)$, $t \in \{1, ..., T\}$, также объединяются в согласованный с вектором PC вектор K такой же размерности. Сделано это так, чтобы прибыль транспортного оператора, формула расчета которой используется в (2), полностью совпадала со скалярным произведением двух новых векторов PC и K.

Вектор K

Поскольку количество элементов во всех матрицах $K1(t)$ и $K2(t)$, $t \in \{1, ..., T\}$ составляет $2TN^2$, то и размерности векторов PC и K также совпадают с этим значением. При формировании вектора K вопрос стоит в том, в каком именно порядке расположить все элементы матриц $K1(t)$ и $K2(t)$, $t\in\{1, ..., T\}$. После того как этот порядок определен, соответствующим образом и в соответствующем порядке составляется вектор PC.

Пусть первые TN^2 элементов вектора K соответствуют всем элементам матриц $K1(t)$, $t \in \{1, ..., T\}$. При этом первые N^2 элементов взяты из матрицы $K1(t)$: первые N элементов совпадают с первой строкой этой матрицы (первый элемент строки ставится на первую позицию вектора, второй элемент строки – на вторую позицию и т.д.), следующие N элементов соответствуют второй строке и так далее. Следующие N^2 элементов вектора K записываются с помощью аналогичного упорядочивания элементов матрицы $K1(2)$. По такому же принципу вектор K заполняется элементами остальных матриц $K1(t)$, $t = 3, ..., T$. Вторая половина вектора K, также состоящая из TN^2 элементов, заполняется аналогичным путем только элементами матриц $K2(t)$, $t = 3, ..., T$.

Вектор *PC*

Для того, чтобы вектор PC был согласован с вектором K, первые TN^2 его элементов берутся из матрицы P, а остальные TN^2 заполняются на основе матрицы C. Первые TN^2 элементов вектора PC соответствуют первой строке матрицы P (первый элемент строки ставится на первую позицию вектора, второй элемент строки – на вторую позицию и т.д.), следующие TN^2 элементов соответствуют второй строке матрицы P и так далее. После того, как первые N^2 элементов вектора PC определены по указанной процедуре, эта часть вектора копируется и ставится на место следующих N^2 элементов. Повторение этой операции $T - 1$ раз завершает формирование первой половины вектора PC, состоящей из TN^2 элементов. Вторая половина данного вектора заполняется аналогично элементами матрицы C. Единственное отличие в правиле заполнения состоит в том, что перед всеми элементами второй половины вектора PC дополнительно ставится знак минус. Таким образом, первые TN^2 элементов вектора PC неотрицательные, следующие TN^2 элементов данного вектора неположительные.

Векторы *Q*, S^0 и *S*

Из матрицы $\bar{Q}$ по аналогичному правилу составляется вектор Q (первые N элементов вектора соответствуют первой строке матрицы, следующие N элементов соответствуют второй строке и т.д.). Размерность нового вектора составляет N^2.

Системы векторов $\bar{S}^0(t)$ и $\bar{S}(t, K1, K2)$, $t \in \{1, ..., T\}$ преобразуются в два вектора путем последовательной конкатенации векторов, соответствующих каждому моменту времени (первые N элементов каждого из новых векторов соответствуют вектору $\bar{S}^0(1)$ и $\bar{S}(1, K1, K2)$, соответственно, следующие N элементов соответствуют вектору $\bar{S}^0(2)$ и $\bar{S}(2, K1, K2)$, соответственно, и т.д.). В итоге получится два новых вектора размера TN, первый обозначим через S^0, второй – через S.

Имея векторы PC и K, в новом формате могут быть представлены критерий оптимизации (2) и условие (6). Однако для формулирования условий (3) и (4) необходимо ввести специальные матрицы, при умножении которых на вектор K можно было бы получить нужные ограничения.

Матрица A_Q

Построим матрицу A_Q, которая позволит переписать условие (4). Нетрудно заметить, что размерность этой матрицы равна ($N^2 \times 2TN^2$). Это связано с тем, что данная матрица справа умножается на вектор K, и результат такого умножения сравнивается с вектором Q размерности N^2. Исходя из того, что элементы матрицы $K2$ (которые соответствуют последним TN^2 элементам вектора K) в данном ограничении не участвуют, все элементы матрицы A_Q, расположенные в последних TN^2 колонках, равны нулю.

Теперь необходимо определить первые TN^2 столбцов данной матрицы. Условие (4) означает, что для каждого маршрута со станции i на станцию j суммарное количество отправляемых груженых вагонов во все дни планового периода не может превышать того объема, который указан в заявке – $\bar{Q}_{ij}$. Исходя из этого, легко заметить, что в первом столбце первой строки стоит единица (она соответствует элементу $K1_{11}(1)$). Следующая единица встретится в первой строке на позиции $N^2 + 1$ (она соответствует элементу $K1_{11}(2)$) и так далее: единицы будут встречаться в первой строке матрицы A_Q T раз с периодом N^2 элементов. Далее, во второй строке матрицы A_Q единица стоит на второй позиции (она соответствует элементу $K1_{12}(1)$), следующая единица во второй строке – на позиции $N^2 + 2$ (она соответствует элементу $K1_{12}(2)$) и так далее: также с периодичностью N^2 элементов будут стоять единицы в количестве T штук. Для остальных строк матрицы A_Q алгоритм заполнения аналогичен.

Таким образом, если матрицу A_Q разбить на T блоков, каждый размером N^2 на N^2 элементов, отбросив нулевую часть (последние TN^2 колонок), то каждый из этих блоков будет представлять собой

единичную матрицу. После определения матрицы A_Q в новом формате условие (4) принимает вид неравенства $A_Q \cdot K \le Q$, где под знаком «≤» подразумевается поэлементное сравнение компонент обоих векторов.

Матрицы A_{in} и A_{out}

Осталось переписать условие (3). Для его записи в матричной форме необходимо определить две матрицы, обозначенные через A_{in} и A_{out}. С помощью матрицы A_{in} рассчитывается количество вагонов, въезжающих на каждую из станций в каждый момент времени $t \in \{1, ..., T\}$. С помощью матрицы A_{out} рассчитывается количество вагонов, отбывающих с каждой станций в каждый момент времени $t \in \{1, ..., T\}$. Размерность каждой из матриц составляет $(TN \times 2TN^2)$. Обусловлено это с тем, что данные матрицы умножаются справа на вектор K. Результат произведения вектора K на любую из этих матриц связан с распределением вагонов по станциям и времени, поэтому количество строк в матрицах и размерность получаемого в результате произведения вектора равна TN. При этом первые N элементов полученного вектора соответствуют распределению вагонов по всем станциям в первый период, следующие N элементов соответствуют распределению вагонов по всем станциям во второй период и так далее до периода T. Матрица A_{in} позволяет представить формулу (1) в виде $S = A_{in} \cdot K$.

Сформируем матрицу A_{out}. Как уже было сказано, эта матрица предназначена для подсчета количества вагонов, отправленных с произвольной станции i в произвольный момент времени t. Куда именно отправлены эти вагоны, в данном случае интереса не представляет. Учитывая упорядоченность элементов вектора K, можно утверждать, что первые N элементов первой строки матрицы A_{out} отвечают за груженые вагоны, исходящие в первый период времени с первой станции. Соответственно эти элементы равны единице. Далее, начиная с элемента $N + 1$, до элемента TN^2 включительно расположены нулевые элементы. Начиная с $TN^2 + 1$, до элемента $TN^2 + N$ снова расположены единичные элементы, отвечающие за исходящие порожние вагоны. Оставшиеся элементы первой строки матрицы A_{out} равны нулю. Вторая строка матрицы A_{out} имеет похожую структуру с тем отличием, что все сдвинуто на N элементов вправо. Иначе говоря, первые N элементов второй строки равны нулю, следующие Nэлементов равны единице, затем, начиная с $TN^2 + N + 1$, до $TN^2 + 2N$ расположены единичные элементы, а остальные элементы второй строки равны 0. Аналогично в каждой следующей строке координаты единичных элементов сдвигаются на N элементов. Повторение этой операции для всех строк матрицы A_{out} завершает ее формирование.

Осталось определить матрицу A_{in}. Для каждого из груженых и порожних маршрутов со станции i на станцию j известна длительность выполнения такого перегона: она равна $\Theta 1_{ij}$ для грузовых и $\Theta 2_{ij}$ для порожних перегонов. Посмотрим, как эта информация отражается в матрице A_{in}. Для этого рассмотрим произвольный перегон груженых вагонов в количестве $K1_{ij}(t)$ со станции i на станцию j в день t планового периода. В векторе K ему соответствует элемент, расположенный на позиции $TN^2 + (i-1)N + j$. Время движения по данному маршруту составляет $\Theta 1_{ij}$. Это означает, что выехав в день t, отправленные вагоны окажутся на станции j в день $t + \Theta 1_{ij}$. В терминах уравнения (1) можно утверждать, что в результате данного груженого рейса произойдет увеличение значения элемента $\overline{S}_j\left(t + \Theta 1_{ij}, K1, K2\right)$ на $K1_{ij}(t)$ единиц. Следовательно, если $t + \Theta 1_{ij}$ не превышает T, то результатом умножения A_{in} на K является увеличение координаты $(t + \Theta 1_{ij} - 1)N + j$ вектора S на $K1_{ij}(t)$ единиц. Для того чтобы указанное перемножение дало такой результат, необходимо, чтобы элемент матрицы A_{in} с координатами $\left(\left(t + \Theta 1_{ij} - 1\right)N + j, tN^2 + \left(i-1\right)N + j\right)$ равнялся единице. Произвольному порожнему перегону в количестве $K2_{ij}(t)$ вагонов со станции i на станцию j в день t планового периода, при условии, что $t + \Theta 2_{ij}$ не превышает T, соответствует единичный элемент матрицы A_{in} с координатами $\left(\left(t + \Theta 2_{ij} - 1\right)N + j, tN^2 + \left(i-1\right)N + j\right)$. Видно, что в случае порожних перегонов ко второй компоненте координаты единичного элемента матрицы A_{in} дополнительно прибавляется величина TN^2. Связанно это с тем, что порожним перегонам в векторе K соответствует вторая половина этого вектора, которая начинается с координаты $TN^2 + 1$ и заканчивается последним элементом с координатой $2TN^2$. Таким образом, алгоритм конструирования матрицы A_{in} состоит в том, чтобы взять нулевую матрицу размером TN на $2TN^2$ элементов и расставить в ней единичные элементы, перебрав все элементы вектора K (или, что тоже самое, перебрав все элементы матриц $K1(t)$ и $K2(t)$, $t \in \{1, ..., T\}$).

После того как матрицы A_{in} и A_{out} определены, ограничение (3) может быть переписано в виде

$A_{in} \cdot K + S_0 = A_{out} \cdot K$ или, что тоже самое, в виде $(A_{out} - A_{in}) \cdot K = S_0$.

Задача линейного программирования

После того, как все вектора и матрицы в новом формате определены, можно переписать задачу (2), (3), (4), (6) в новом формате:

$$PC^T \cdot K \to \max_K, \quad (7)$$

при ограничениях

$$(A_{out} - A_{in}) \cdot K = S_0; \quad (8)$$

$$A_Q \cdot K \le Q; \quad (9)$$

$$K \ge 0. \quad (10)$$

Задача (7)–(10) абсолютно идентична задаче (2), (3), (4), (6), но ее преимущество заключается в том, что она записана в формате классической задачи линейного программирования, что позволяет решать ее, применяя соответствующие методы и программные инструменты.

4. Решение транспортной задачи на гипотетическом примере

В качестве примера решения рассматриваемой транспортной задачи продемонстрирован простой модельный пример с небольшим числом станций и коротким горизонтом планирования.

Пусть число станций равно 4 ($N = 4$), горизонт планирования составляет 3 дня ($T = 3$). Список поступивших заявок состоит из пяти позиций. Приведем эти заявки в виде *таблицы 1.*

Таблица 1.

Список заявок на перевозку груза в модельном примере

№	Пункт отправления	Пункт назначения	Объем заявки (в вагонах)	Ставка (в условных единицах)
1	1	3	3	2,9
2	2	1	5	1,1
3	2	3	4	2,3
4	3	2	7	1,9
5	3	4	6	2,1

На основе списка заявок необходимо составить две матрицы – матрицу тарифов P, элементы которой записаны в условных единицах, и матрицу объемов заявок $\bar{Q}$:

$$P = \begin{pmatrix} 0 & 0 & 2,9 & 0 \\ 1,1 & 0 & 2,3 & 0 \\ 0 & 1,9 & 0 & 2,1 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \ \bar{Q} = \begin{pmatrix} 0 & 0 & 3 & 0 \\ 5 & 0 & 4 & 0 \\ 0 & 7 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Далее приведем время движения как груженых, так и порожних маршрутов в виде значений матриц $\Theta 1$ и $\Theta 2$:

$$\Theta 1 = \begin{pmatrix} 0 & 2 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 \\ 1 & 2 & 1 & 0 \end{pmatrix}; \ \Theta 2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 \end{pmatrix}.$$

Напомним, что диагональные элементы матрицы берутся равными единице. Связано это с тем, что если вагоны необходимо оставить на станции до следующего дня, то это равносильно тому, что они как бы отправляются из данной станции на саму себя в рейс длительностью один день.

Значения тарифов РЖД на порожние перегоны, так же как и ставки, выраженные в условных единицах, характеризуются значениями элементов матрицы C:

$$C = \begin{pmatrix} 0 & 1,9 & 1,3 & 1,9 \\ 1,2 & 0 & 1,8 & 0,9 \\ 1,1 & 1,2 & 0 & 1,6 \\ 1,3 & 1,5 & 1,2 & 0 \end{pmatrix}.$$

В рамках данной задачи предполагается, что вагоны могут оставаться на станциях до следующего дня бесплатно, поэтому диагональные элементы матрицы C нулевые.

Начальное распределение вагонов характеризуется следующими векторами:

$$\bar{S}^0(1) = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 3 \end{pmatrix}; \ \bar{S}^0(2) = \begin{pmatrix} 5 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

В период $t = 3$ вагоны не прибывают, что эквивалентно нулевому вектору $\bar{S}^0(3)$.

Перепишем эту задачу в новых обозначениях. Вектор PC имеет размерность $2TN^2 = 2 \cdot 3 \cdot 16 = 96$. Представление такого вектора в явном виде не уместится на странице, поэтому рассмотрим промежуточные векторы p и c длинной $N^2 = 16$, составлен-

ные на основе матриц P и C:

$$p = \begin{pmatrix} 0 \\ 0 \\ 2,9 \\ 0 \\ 1,1 \\ 0 \\ 2,3 \\ 0 \\ 0 \\ 1,9 \\ 0 \\ 2,1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad c = \begin{pmatrix} 0 \\ 1,9 \\ 1,3 \\ 1,9 \\ 1,2 \\ 0 \\ 1,8 \\ 0,9 \\ 1,1 \\ 1,2 \\ 0 \\ 1,4 \\ 1,3 \\ 1,5 \\ 1,2 \\ 0 \end{pmatrix}.$$

Вектор PC получается последовательной конкатенацией этих векторов, а именно:

$$PC = \left(p^T,\ p^T,\ p^T,\ c^T,\ c^T,\ c^T \right)^T.$$

Вектор Q, получаемый из матрицы $\overline{Q}$ и имеющий размерность $N^2 = 16$, принимает следующий вид:

$$Q = (0\ \ 0\ \ 3\ \ 0\ \ 5\ \ 0\ \ 4\ \ 0\ \ 0\ \ 7\ \ 0\ \ 6\ \ 0\ \ 0\ \ 0\ \ 0)^T.$$

Вектор S^0, размерность которого составляет $TN = 3 \cdot 4 = 12$, сконструирован на основе векторов $\overline{S}^0(1)$ и $\overline{S}^0(2)$ и имеет следующий вид:

$$S^0 = \left(0\ \ 2\ \ 1\ \ 3\ \ 5\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\right)^T.$$

Осталось получить матрицы A_{in}, A_{out} и A_Q. Начнем с матрицы A_Q, ее размерность составляет $(N^2 \times 2TN^2) = (16 \times 96)$. Представим ее в формате разреженной матрицы, т.е. укажем координаты ненулевых элементов, равные единице. Список координат единичных элементов матрицы A_Q следующий (здесь и далее нумерация строк и столбцов начинается с единицы):

(1, 1), (2, 2), (3,3), (4, 4), (5, 5), (6, 6), (7, 7), (8, 8), (9, 9), (10, 10), (11, 11), (12, 12), (13, 13), (14, 14), (15, 15), (16, 16), (1, 17), (2, 18), (3, 19), (4, 20), (5, 21), (6, 22), (7, 23), (8, 24), (9, 25), (10, 26), (11, 27), (12, 28), (13, 29), (14, 30), (15, 31), (16, 32), (1, 33), (2, 34), (3, 35), (4, 36), (5, 37), (6, 38), (7, 39), (8, 40), (9, 41), (10, 42), (11, 43), (12, 44), (13, 45), (14, 46), (15, 47), (16, 48).

В таком же формате приводятся матрицы A_{in} и A_{out}, размерность которых составляет $(TN \times 2TN^2 = (12 \times 96)$. Список координат единичных элементов матрицы A_{in}:

(5, 1), (7, 3), (12,4), (5, 5), (6, 6), (8, 8), (5, 9), (10, 10), (7, 11), (5, 13), (10, 14), (7, 15), (8, 16), (9, 17), (11, 19), (9, 21), (10, 22), (12, 24), (9, 25), (11, 27), (9, 25), (11, 27), (9, 29), (11, 31), (12, 32), (5, 49), (7, 51), (5, 53), (6, 54), (8, 56), (5, 57), (10, 58), (7, 59), (5, 61), (10, 62), (7, 63), (8, 64), (9, 65), (10, 70), (12, 72), (9, 73), (11, 75), (9, 77), (11, 79), (12, 80).

Список координат единичных элементов матрицы A_{out}:

(1, 1), (1, 2), (1, 3), (1, 4), (2, 5), (2, 6), (2, 7), (2, 8), (3, 9), (3, 10), (3, 11), (3, 12), (4, 13), (4, 14), (4, 15), (4, 16), (5, 17), (5, 18), (5, 19), (5, 20), (6, 21), (6, 22), (6, 23), (6, 24), (7, 25), (7, 26), (7, 27), (7, 28), (8, 29), (8, 30), (8, 31), (8, 32), (9, 33), (9, 34), (9, 35), (9, 36), (10, 37), (10, 38), (10, 39), (10, 40), (11, 41), (11, 42), (11, 43), (11, 44), (12, 45), (12, 46), (12, 47), (12, 48), (1, 49), (1, 50), (1, 51), (1, 52), (2, 53), (2, 54), (2, 55), (2, 56), (3, 57), (3, 58), (3, 59), (3, 60), (4, 61), (4, 62), (4, 63), (4, 64), (5, 65), (5, 66), (5, 67), (5, 68), (6, 69), (6, 70), (6, 71), (6, 72), (7, 73), (7, 74), (7, 75), (7, 76), (8, 77), (8, 78), (8, 79), (8, 80), (9, 81), (9, 82), (9, 83), (9, 84), (10, 85), (10, 86), (10, 87), (10, 88), (11, 89), (11, 90), (11, 91), (11, 92), (12, 93), (12, 94), (12, 95), (12, 96).

После того как все векторы и матрицы определены, задача (7)–(10) может быть решена. Результатом решения данной задачи является вектор K, размерность которого составляет $2TN^2 = 96$. Приведем одно из найденных решений K, выписав значения только ненулевых элементов:

$K_7 = 2$; $K_{10} = 1$; $K^{19} = 2$; $K_{37} = 2$; $K_{42} = 4$; $K_{44} = 2$; $K_{62} = 3$; $K_{65} = 2$; $K_{67} = 1$; $K_{79} = 1$; $K_{81} = 2$.

Переведем данное решение в формат матриц $K1(t)$ и $K2(t)$:

$$K1(1) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad K1(2) = \begin{pmatrix} 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad (11)$$

$$K1(3) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix};$$

$$K2(1)=\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&1&2&0\end{pmatrix};\ K2(2)=\begin{pmatrix}0&0&2&0\\0&0&0&0\\0&0&0&0\\0&0&1&0\end{pmatrix};$$
$$K2(3)=\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}. \quad (12)$$

Полученное решение (11), (12), в силу малого масштаба задачи, также может быть представлено в виде схемы, изображенной на *рисунке 1.*

Прибыль от полученного решения можно рассчитать двумя эквивалентными способами: либо на основе формулы, используемой в (2), либо как произведение векторов $PC^T \cdot K$. Для данной задачи ее значение, выраженное в условных единицах, получается равным 32,3.

Проанализируем полученное решение. Видно, что одна из заявок на транспортировку пяти вагонов со станции 2 на станцию 1 осталась неисполненной. Заявка на транспортировку семи вагонов из станции 3 на станцию 2 должна выполняется частично в объеме пяти вагонов. Остальные заявки в соответствии с планом выполняются полностью. Заявки со станции 1 на станцию 3, а также со станции 3 на станцию 4 должны быть выполнены в соответствии с найденным планом полностью единичными рейсами. Заявки со станции 2 на станцию 3 и со станции 3 на станцию 2 в соответствии с планом выполняются поэтапно – в два рейса для первой заявки и в три рейса для второй. Также найденный план предусматривает несколько порожних рейсов. Так, в первый день из станции 4 на станцию 2 направляется один порожний вагон, и из станции 4 на станцию 3 – два порожних вагона. Длительность первого порожнего перегона составляет двое суток (вагон приедет на станцию 2 к третьему периоду), длительность второго порожнего рейса – одни сутки. Аналогично, во второй день осуществляется два порожних перегона: одного вагона из станции 4 на станцию 3 и двух вагонов из станции 1 на станцию 3, длительность обоих перегонов составляет одни сутки. В заключительном третьем периоде отправлений порожних рейсов не запланировано.

Заключение

В работе рассматривается транспортная задача, которая возникает перед транспортными операторами при управлении парком грузовых железнодорожных вагонов. Предложенный подход позволяет найти оптимальный план на всем множестве возможных маршрутов, тогда как способы, связанные с методом генерации колонок, осуществляют поиск оптимальных планов на подмножествах всех возможных маршрутов. Существенным недостатком предложенного подхода является то, что время вычислений при его применении заметно выше по

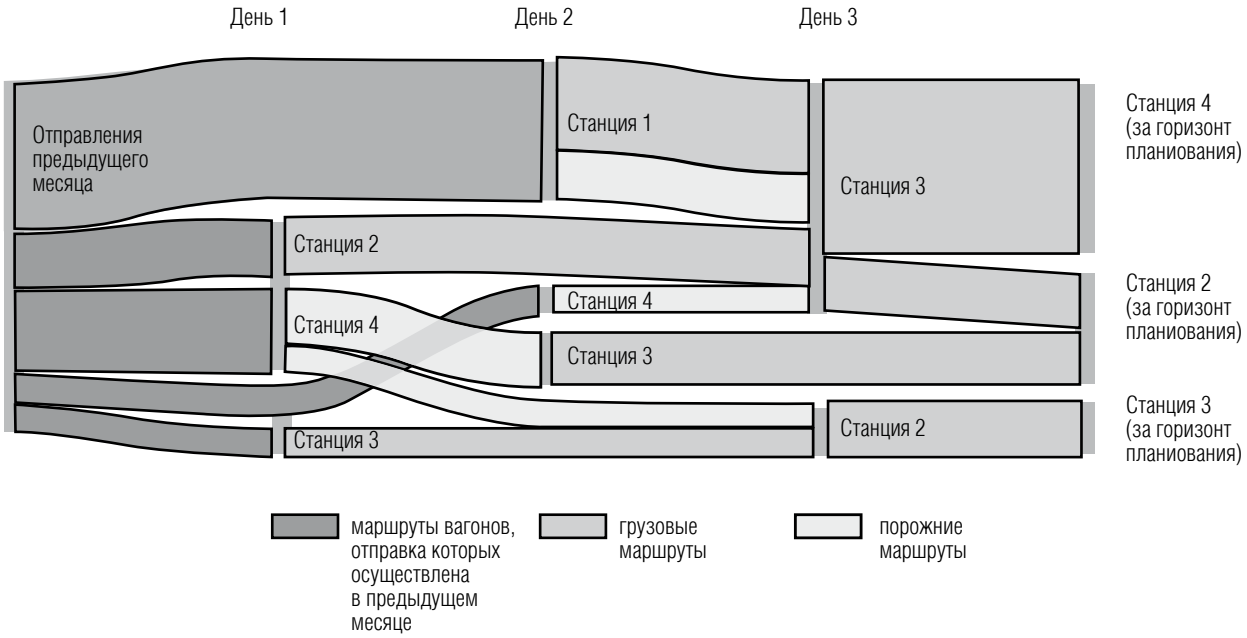


Рис. 1. Схематичное представление решения (11), (12)

сравнению с временем, которое затрачивается при использовании методов, связанных с методом генерации колонок. Поэтому важно предложить способы модификации данного подхода, которые приведут к ускорению вычислительных процессов.

Большая размерность задачи во многом обусловлена тем, что в представленной постановке (2)–(4), (6) при решении берутся в расчет все возможные маршруты, в том числе и те, которых фактически не существует. В частности, это касается грузовых маршрутов: в рассматриваемой постановке перебираются все грузовые маршруты, хотя только малая их доля является актуальной (актуальны те, которые указаны в заявках).

При решении задачи на основе реальных данных количество станций может быть равным примерно 1000, размерность вектора K в этом случае составит порядка $2TN^2 = 1{,}2 \cdot 10^8$, при горизонте планирования $T = 60$. Понятно, что на практике задачу линейного программирования такой размерности невозможно решить за разумное время.

Вектор K состоит из двух частей: первая составлена на основе матриц $K1(t)$, вторая – на основе матриц $K2(t)$. Если для порожних маршрутов заранее неизвестно, какие из них будут востребованы, а какие нет, то в случае с грузовыми маршрутами известно, что востребованными будут только те, которые указаны в заявках. Поэтому остальные маршруты из рассмотрения могут быть удалены. Это означает, что при осуществлении расчетов более рационально учитывать только те элементы матриц $K1(t)$, $t \in \{1, ..., T\}$, которые соответствуют грузовым маршрутам, указанным в поданных заявках. В векторе K также должны быть учтены только эти грузовые маршруты: рассматривать остальные грузовые маршруты не имеет смысла. Поэтому размерность вектора K может заметно сократиться. Очевидно, что исключать такие маршруты необходимо не только из вектора K: аналогично должны быть модифицированы алгоритмы формирования вектора PC, а также алгоритмы формирования матриц A_Q, A_{in} и A_{out}. В данной работе для упрощения изложения указанные алгоритмы формирования векторов и матриц не рассмотрены.

Для оценки того, насколько сильно удается сократить размерность задачи, обратимся к данным реальной задачи, в которой $N = 1126$, а количество заявок, равное количеству востребованных грузовых маршрутов, составляет 1616. В таких условиях первая часть модифицированного вектора K, отвечающая за грузовые перегоны, при $T = 60$ имеет размерность $T \cdot 1616 = 96960$ вместо $2TN^2 = 76072560$. Видно, что почти в 800 раз сокращается размерность первой части вектора K. Если алгоритм формирования второй части вектора K, отвечающей за порожние перегоны, оставить неизменным, то размерность вектора K в модифицированном алгоритме составит $T \cdot 1616 + 2TN^2 = 76169520$ вместо $2TN^2 = 152145120$, т.е. сократится почти вдвое.

Отдельно отметим, что для решения реальных задач горизонт планирования, равный одному месяцу, оказывается слишком коротким. Это связано с тем, что при таком горизонте планирования результатом оптимизации может оказаться план, при котором на последние дни расчетного месяца будет запланировано большое количество заявок с максимальной ставкой, а также максимальной дальностью и длительностью перегонов. Реализация такого плана приведет к максимально возможной прибыли за расчетный месяц. Однако это чревато тем, что в месяце, следующим за расчетным, вагоны могут быть крайне неудачно распределены как в пространстве, так и по времени прибытия на станции назначения. Соответственно, это может привести к заметному снижению показателей прибыли в следующем расчетном месяце. Исходя из этого, горизонт планирования в оптимизационной модели должен быть увеличен. При этом если увеличивать горизонт планирования до нескольких месяцев, то для каждого из этих месяцев должно осуществляться планирование относительно своего отдельного списка заявок. Вообще говоря, списки заявок для различных месяцев в точности совпадать не должны, но поскольку достоверно известен только текущий список заявок и для следующих месяцев заявки неизвестны, в рамках модели можно предполагать, что списки заявок, ожидаемые в будущем, совпадают с текущим списком. Мы считаем разумным в качестве горизонта планирования брать два месяца, такое расширение горизонта планирования приведет к двукратному увеличению размерности задачи. Дальнейшая работа по развитию модели должна быть проделана в направлении учета большего количества факторов, а также в направлении оптимизации вычислительных процессов путем снижения размерности задачи. Все вышеперечисленное предполагается учесть в последующих исследованиях. ■

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-06003.

Литература

1. Лазарев А.А., Мусатова Е.Г. Гафаров Е.Р., Кварацхелия А.Г. Теория расписаний. Задачи железнодорожного планирования. М.: ИПУ РАН, 2012.
2. Лазарев А.А., Мусатова Е.Г. Кварацхелия А.Г., Гафаров Е.Р. Теория расписаний. Задачи управления транспортными системами. М.: МГУ, 2012.
3. Lusba R., Larsen J., Ehrgott M., Ryan D. Railway track allocation: models and methods // OR Spectrum. 2011. Vol. 33. No 4. P. 843–883. DOI: 10.1007/s00291-009-0189-0.
4. Cordean J.-F., Toth P., Vigo V. A survey of optimization models for train routing and scheduling // Transportation Science. 1988. Vol. 32. No 4. P. 380–404. DOI: 10.1287/trsc.32.4.380.
5. Ravindra K., Möhring R.H., Zaroliagis C.D. (Eds.) Robust and online large-scale optimization: Models and techniques for transportation systems. Berlin, Heidelberg: Springer 2009. DOI: 10.1007/978-3-642-05465-5.
6. Brucker P., Hurink J., Rolfes T. Routing of railway carriages: A case study. Technical Report 1498. Netherlands: University of Twente, 1999.
7. Brucker P., Hurink J., Rolfes T. Routing of railway carriages // Journal of Global Optimization. 2003. No 27. P. 313–332. DOI: https://doi.org/10.1023/A:1024843208074.
8. Cordean J.F., Soumis F., Desrosiers J. Simultaneous assignment of locomotives and cars to passenger trains // Operations Research. 2001. Vol. 49. No 4. P. 531–548. DOI: 10.1287/opre.49.4.531.11226.
9. Eidenbenz S., Pagourtzis A., Widmayer P. Flexible train rostering. Algorithms and Computation // ISAAC 2003. Lecture Notes in Computer Science. Vol. 2906. P. 615–624. DOI: https://doi.org/10.1007/978-3-540-24587-2_63.
10. Бекларян Л.А., Хачатрян Н.К. Об одном классе динамических моделей грузоперевозок // Журнал вычислительной математики и математической физики. 2013. Т. 53. № 10. С. 1649–1667. DOI: 10.7868/S0044466913100037.
11. Khachatryan N.K., Akopov A.S. Model for organizing cargo transportation with an initial station of departure and a final station of cargo distribution // Business Informatics. 2017. No 1. P. 25–35. DOI: 10.17323/1998-0663.2017.1.25.35.
12. Khachatryan N.K., Akopov A.S., Belousov F.A. About quasi-solutions of traveling wave type in models for organizing cargo transportation // Business Informatics. 2018. No 1. P. 61–70. DOI: 10.17323/1998-0663.2018.1.61.70.
13. Бекларян Л.А., Хачатрян Н.К. Динамические модели организации грузопотока на железнодорожном транспорте // Экономика и математические методы. 2019. Т. 55. № 3. С. 62–73. DOI: 10.31857/S042473880005780-7.
14. Fukasawa R., Aragão M.P., Porto O., Uchoa E. Solving the freight car flow problem to optimality // Electronic Notes in Theoretical Computer Science. 2002. Vol. 66. No 6. P. 42–52. DOI: 10.1016/S1571-0661(04)80528-0.
15. Optimizing the cargo express service of Swiss Federal Railways / A. Ceselli [et al.] // Transportation Science. 2008. Vol. 42. No 4. P. 450–465. DOI: 10.1287/trsc.1080.0246.
16. Lulli G., Pietropaoli U., Ricciardi N. Service network design for freight railway transportation: the Italian case // Journal of the Operational Research Society. 2011. Vol. 62. No 12. P. 2107–2119. DOI: 10.1057/jors.2010.190.
17. Campetella M., Lulli G., Pietropaoli U., Ricciardi N. Fright service design for the Italian railways company // 6th Workshop on Algorithmic Approach for Transportation Modelling, Optimization, and Systems (ATMOS 2006), Zurich, Switzerland, 14 September 2006. P. 1–13. DOI: 10.4230/OASIcs.ATMOS.2006.685.
18. Andersen J., Christiansen M. Designing new European rail freight services // Journal of the Operational Research Society. 2009. No 60. P. 348–360. DOI: 10.1057/palgrave.jors.2602559.
19. Jeong S.-J., Lee C.-G., Bookbinder J. The European freight railway system as a hub-and-spoke network // Transportation Research, Part A: Policy and Practice. 2007. Vol. 41. No 6. P. 523–536. DOI: 10.1016/j.tra.2006.11.005.
20. Sadykov R., Lazarev A., Shiryaev V., Stratonnikov A. Solving a freight railcar flow problem arising in Russia // 13th Workshop on Algorithmic Approach for Transportation Modelling, Optimization, and Systems (ATMOS'13). Sophia Antipolis, France, 5 September 2013, P. 55–67. DOI: 10.4230/OASIcs.ATMOS.2013.55.
21. Лазарев А.А., Садыков Р.Р. Задача управления парком грузовых железнодорожных вагонов // XII Всероссийское совещание по проблемам управления (ВСПУ 2014). Москва, ИПУ РАН, Россия, 16–19 июня 2014. С. 5083–5093.
22. Desaulniers J., Desrosiers J., Solomon M. Column generation. New York: Springer, 2005.
23. Sadykov R., Vanderbeck F. Column generation for extended formulations // EURO Journal on Computational Optimization. 2013. Vol. 1. No 1–2. P. 81–115. DOI: 10.1007/s13675-013-0009-9.
24. Ahuja R., Magnanti T., Orlin J. Network flows: Theory, algorithms, and applications. Prentice Hall, 1993.
25. Williamson D. Network flow algorithms. Cambridge: Cambridge University Press, 2019. DOI: 10.1017/9781316888568.
26. Evans J.R., Minieka E. Optimization algorithms for networks and graphs. New York: Marcel Dekker, 1992.

Об авторах

Белоусов Федор Анатольевич

кандидат экономических наук;

научный сотрудник лаборатории динамических моделей экономики и оптимизации,
Центральный экономико-математический институт Российской академии наук,
117418, г. Москва, Нахимовский проспект, д. 47;

E-mail: sky_tt@list.ru

ORCID: 0000-0002-3040-3148

Неволин Иван Викторович

кандидат экономических наук;

ведущий научный сотрудник лаборатории экспериментальной экономики, Центральный экономико-математический институт Российской академии наук, 117418, г. Москва, Нахимовский проспект, д. 47;

E-mail: i.nevolin@cemi.rssi.ru

Хачатрян Нерсес Карленович

кандидат физико-математических наук;

ведущий научный сотрудник лаборатории динамических моделей экономики и оптимизации, Центральный экономико-математический институт Российской академии наук, 117418, г. Москва, Нахимовский проспект, д. 47;

E-mail: nerses@cemi.rssi.ru

ORCID: 0000-0003-2495-5736

Modeling and optimization of plans for railway freight transport performed by a transport operator

Fedor A. Belousov
E-mail: sky_tt@list.ru

Ivan V. Nevolin
E-mail: i.nevolin@cemi.rssi.ru

Nerses K. Khachatryan
E-mail: nerses@cemi.rssi.ru

Central Economics and Mathematics Institute, Russian Academy of Sciences
Address: 47, Nakhimovsky Prospect, Moscow 117418, Russia

Abstract

This paper offers an approach for solving a problem that arises for railway transport operators. The task is to manage the fleet of freight railcars optimally in terms of profit maximization. The source data for the transport operator is a list of requests received from customers, as well as the location of railcars at the beginning of the planning period. The request formed by each customer consists of departure station, destination station, name and volume of cargo that the customer would like to transport. The request also contains the rate that the customer has to pay to the transport operator for each loaded wagon transported. Planning is carried out for a month in advance and consists, on the one hand, in selecting the most profitable requests for execution, on the other hand – in building a sequence of cargo and empty runs that will fulfill the selected requests with the greatest efficiency. Direct transportation of loaded and empty railway cars is carried out by Russian Railways with pre-known tariffs and time standards for each of the routes. At the same time, tariffs for driving loaded wagons are additional costs for the customer of the route specified in the request (customers pay both the transport operator for the use of wagons and Russian Railways); transportation of empty wagons is paid by transport operators. To solve this problem, one of the possible ways to reduce it to a large-

dimensional linear programming problem is proposed. An algorithm is proposed, the result of which is a problem written in the format of a linear programming problem. To demonstrate the approach clearly, a simplified problem statement is considered that takes into account only the main factors of the modeled process. The paper also shows an example of a numerical solution of the problem based on simple model data.

Key words: railway cargo transportation; planning of railway cargo transportation; optimal plan of railway cargo transportation; optimal management of railway car fleet; linear programming; schedule theory; operations research; unmanned vehicles.

Citation: Belousov F.A., Nevolin I.V., Khachatryan N.K. (2020) Modeling and optimization of plans for railway freight transport performed by a transport operator. *Business Informatics,* vol. 14, no 2, pp. 21–35.
DOI: 10.17323/2587-814X.2020.2.21.35

References

1. Lazarev A.A., Musatova E.G., Gafarov E.R., Kvaratskheliya A.G. (2012) *Schedule Theory. Problems of railway planning.* Moscow: IPU RAS (in Russian)
2. Lazarev A.A., Musatova E.G., Kvaratskheliya A.G., Gafarov E.R. (2012) *Schedule Theory. Problems of transport system management.* Moscow: MSU (in Russian).
3. Lusba R., Larsen J., Ehrgott M., Ryan D. (2011) Railway track allocation: models and methods. *OR Spectrum*, vol. 33, no 4, pp. 843–883. DOI: 10.1007/s00291-009-0189-0.
4. Cordean J.-F., Toth P., Vigo V. (1988) A survey of optimization models for train routing and scheduling. *Transportation Science*, vol. 32, no 4, pp. 380–404. DOI: 10.1287/trsc.32.4.380.
5. Ravindra K., Möhring R.H., Zaroliagis C.D., eds. (2009) *Robust and online large-scale optimization: Models and techniques for transportation systems.* Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-05465-5.
6. Brucker P., Hurink J., Rolfes T. (1999) *Routing of railway carriages: A case study. Technical Report 1498.* Netherlands: University of Twente.
7. Brucker P., Hurink J., Rolfes T. (2003) Routing of railway carriages. *Journal of Global Optimization*, no 27, pp. 313–332. DOI: https://doi.org/10.1023/A:1024843208074.
8. Cordean J.F., Soumis F., Desrosiers J. (2001) Simultaneous assignment of locomotives and cars to passenger trains. *Operations Research*, vol. 49, no 4, pp. 531–548. DOI: 10.1287/opre.49.4.531.11226.
9. Eidenbenz S., Pagourtzis A., Widmayer P. (2003) Flexible train rostering. Algorithms and Computation. ISAAC 2003. *Lecture Notes in Computer Science*, vol. 2906, pp. 615–624. DOI: https://doi.org/10.1007/978-3-540-24587-2_63.
10. Beklaryan L.A., Khachatryan N.K. (2013) On one class of dynamic transportation models. *Computational Mathematics and Mathematical Physics*, vol. 53, no 10, pp. 1649–1667 (in Russian). DOI: 10.7868/S0044466913100037.
11. Khachatryan N.K., Akopov A.S. (2017) Model for organizing cargo transportation with an initial station of departure and a final station of cargo distribution. *Business Informatics*, no 1, pp. 25–35. DOI: 10.17323/1998-0663.2017.1.25.35.
12. Khachatryan N.K., Akopov A.S., Belousov F.A. (2018) About quasi-solutions of traveling wave type in models for organizing cargo transportation. *Business Informatics*, no 1, pp. 61–70. DOI: 10.17323/1998-0663.2018.1.61.70.
13. Beklaryan L.A., Khachatryan N.K. (2019) Dynamic models of cargo flow organization on Railway Transport. *Economics and Mathematical Methods*, vol. 55, no 3, pp. 62–73 (in Russian). DOI: 10.31857/S042473880005780-7.
14. Fukasawa R., Aragão M.P., Porto O., Uchoa E. (2002) Solving the freight car flow problem to optimality. *Electronic Notes in Theoretical Computer Science*, vol. 66, no 6, pp. 42–52. DOI: 10.1016/S1571-0661(04)80528-0.
15. Ceselli A., Gatto M., Lübbecke M., Nunkesser M., Schilling H. (2008) Optimizing the cargo express service of Swiss Federal Railways. *Transportation Science*, vol. 42, no 4, pp. 450–465. DOI: 10.1287/trsc.1080.0246.
16. Lulli G., Pietropaoli U., Ricciardi N. (2011) Service network design for freight railway transportation: the Italian case. *Journal of the Operational Research Society*, vol. 62, no 12, pp. 2107–2119. DOI: 10.1057/jors.2010.190.
17. Campetella M., Lulli G., Pietropaoli U., Ricciardi N. Fright service design for the Italian railways company. Proceedings of the *6th Workshop on Algorithmic Approach for Transportation Modelling, Optimization, and Systems (ATMOS 2006), Zurich, Switzerland, 14 September 2006*, pp. 1–13. DOI: 10.4230/OASIcs.ATMOS.2006.685.
18. Andersen J., Christiansen M. (2009) Designing new European rail freight services. *Journal of the Operational Research Society*, no 60, pp. 348–360. DOI: 10.1057/palgrave.jors.2602559.
19. Jeong S.-J., Lee C.-G., Bookbinder J. (2007) The European freight railway system as a hub-and-spoke network. *Transportation Research, Part A: Policy and Practice*, vol. 41, no 6, pp. 523–536. DOI: 10.1016/j.tra.2006.11.005.

20. Sadykov R., Lazarev A., Shiryaev V., Stratonnikov A. (2013) Solving a freight railcar flow problem arising in Russia. Proceedings of the *13th Workshop on Algorithmic Approach for Transportation Modelling, Optimization, and Systems (ATMOS'13), Sophia Antipolis, France, 5 September 2013*, pp. 55–67. DOI: 10.4230/OASIcs.ATMOS.2013.55.

21. Lazarev A.A., Sadykov R.R. (2014) Management problem of railway cars fleet. Proceedings of the *XII All-Russian Meeting on Management Issues (VSPU 2014), Moscow, IPU RAS, Russia, 16–19 June 2014*, pp. 5083–5093 (in Russian).

22. Desaulniers J., Desrosiers J., Solomon M. (2005) *Column generation*. New York: Springer.

23. Sadykov R., Vanderbeck F. (2013) Column generation for extended formulations. *EURO Journal on Computational Optimization*, vol. 1, no 1–2, pp. 81–115. DOI: 10.1007/s13675-013-0009-9.

24. Ahuja R., Magnanti T., Orlin J. (1993) *Network flows: Theory, algorithms, and applications*. Prentice Hall.

25. Williamson D. (2019) *Network flow algorithms*. Cambridge: Cambridge University Press. DOI: 10.1017/9781316888568.

26. Evans J.R., Minieka E. (1992) *Optimization algorithms for networks and graphs*. New York: Marcel Dekker.

About the authors

Fedor A. Belousov
Cand. Sci. (Econ.);
Researcher, Laboratory of Dynamic Models of Economy and Optimization,
Central Economics and Mathematics Institute, Russian Academy of Sciences,
47, Nakhimovsky Prospect, Moscow 117418, Russia;
E-mail: sky_tt@list.ru
ORCID: 0000-0002-3040-3148

Ivan V. Nevolin
Cand. Sci. (Econ.);
Leading Researcher, Laboratory of Experimental Economics,
Central Economics and Mathematics Institute, Russian Academy of Sciences,
47, Nakhimovsky Prospect, Moscow 117418, Russia;
E-mail: i.nevolin@cemi.rssi.ru

Nerses K. Khachatryan
Cand. Sci. (Phys.-Math.);
Senior Researcher, Laboratory of Dynamic Models of Economy and Optimization,
Central Economics and Mathematics Institute, Russian Academy of Sciences,
47, Nakhimovsky Prospect, Moscow 117418, Russia;
E-mail: nerses@cemi.rssi.ru
ORCID: 0000-0003-2495-5736

DOI: 10.17323/2587-814X.2020.2.36.47

Преимущества когнитивного менеджмента для управления предприятием в современных условиях

Р.А. Караев
E-mail: karayevr@rambler.ru

Н.Ю. Садыхова
E-mail: natella5@rambler.ru

Институт систем управления, Национальная академия наук Азербайджанской Республики
Адрес: Азербайджанская Республика, AZ1141, г. Баку, ул. Б. Вагабзаде, д. 9

Аннотация

В статье приводится краткая характеристика когнитивного менеджмента, открывающего уникальные возможности для эффективного управления предприятиями в современных сложных и нестабильных условиях. Обсуждаются проблемы коммерческой реализации этой перспективной парадигмы. Отмечается, что главной, критической из этих проблем является отсутствие развитой инженерии когнитивного менеджмента. Предлагается концептуальная схема решения этой проблемы, основанная на конвергенции идей и методов «когнитивной школы» и эмпирического опыта, накопленного в инженерии знаний. Приводятся результаты использования концептуальной схемы в четырех исследовательских проектах с различной отраслевой ориентацией, различными внутренними условиями и различной динамикой внешней среды. Обсуждаются инжиниринговые перспективы предлагаемой схемы в части коммерциализации когнитивной школы, идентифицированной Г. Минцбергом, Б. Альстрендом и Д. Лемпелом еще тридцать лет назад.

Ключевые слова: когнитивный менеджмент; концептуальная схема; технология анализа и выбора; когнитивный подход.

Цитирование: Караев Р.А., Садыхова Н.Ю. Преимущества когнитивного менеджмента для управления предприятием в современных условиях // Бизнес-информатика. 2020. Т. 14. № 2. С. 36–47.
DOI: 10.17323/2587-814X.2020.2.36.47

Введение

Начало исследований по стратегическому менеджменту принято относить к середине 1960-х годов. Иногда называют и 1951 год [1]. Но значительно раньше появились работы по военному стратегическому строительству: например, знаменитый трактат о военном искусстве Сунь-Цзы относят к V в. до н. э. Круг исследований по стратегическому менеджменту начал стремительно расширяться в начале 1980-х годов, когда командиры бизнеса столкнулись с растущей сложностью делового мира и осознали необходимость стратегического анализа перспектив своего делового успеха.

В последние годы бизнес сталкивается с новыми вызовами, требующими существенного пересмотра наработанных концепций и традиционных инструментов стратегического менеджмента [2]. Одной из наиболее серьезных проблем, с которыми сталкиваются менеджеры в последние годы, является понимание сложных причинных цепочек, определяющих влияние внешних и внутренних условий предприятия на цели и свойства разрабатываемой стратегии. Сегодня эта проблема усугубляется растущей сложностью и нестабильностью экономической среды, приводящей к многочисленным неопределенностям и рискам.

В новых условиях применение традиционных инструментов стратегического менеджмента повсеместно сталкивается с серьезными ограничениями, диктующими необходимость создания новых инструментов, адекватных исследовательскому характеру современной управленческой практики [3]. Широкие возможности для создания такого рода инструментов открывают идеи и методы когнитивного подхода [4, 5]. Инструменты поддержки, основанные на когнитивном подходе, могут быть отнесены к инструментам нового поколения, формирующегося в рамках исследовательской парадигмы когнитивной науки [6]. Эта парадигма сегодня ориентирована на решение широкого круга сложных проблем современного мира в самых различных областях: экономика, социология, политика, экология, промышленность, бизнес, чрезвычайные ситуации и др., включая проблемы стратегического менеджмента.

К настоящему времени вопросу когнитивного менеджмента посвящено достаточно много работ. Однако большинство из них выполнены теоретиками менеджмента, математиками или психологами, и они очень далеки от запросов управленческой практики. Прошло уже более тридцати лет с момента зарождения когнитивной школы менеджмента [7]. В многочисленных публикациях широко обсуждаются заманчивые и многообещающие перспективы этой школы, однако до настоящего времени все еще нет развитой инженерии когнитивного менеджмента, являющейся (как и в случае всех других “knowledge-based” технологий) критически важным звеном когнитивной аналитики [8].

Показательно, что из проанализированных нами публикаций по когнитивному менеджменту (академические журналы, монографии, труды конференций) мы не встретили ни одну, написанную специалистом, имеющим опыт практической разработки коммерческих моделей. В большинстве из этих публикаций дается анализ проблем когнитивного процесса, и обсуждаются теоретические проблемы для будущих исследований.

Вместе с тем, многолетняя практика убедительно свидетельствует о том, что важной предпосылкой для создания инженерии всех без исключения “knowledge-based” технологий [8] являются критический анализ и обобщение результатов прикладных разработок, формирующих эмпирическую «аксиоматику» проблемной области, не менее важную, чем многочисленные теоретические конструкции.

Еще тридцать лет назад авторы книги [7] отмечали, что «работы сторонников когнитивной школы образуют не столько единую научную школу, сколько собрание не связанных между собой исследований, что не мешает, однако, объединить их в одно направление. Если их авторы успешно справятся с поставленной задачей, вполне возможно, что вскоре изменятся и преподавание, и практика стратегического управления» [7].

На наш взгляд, решению этой сложной «застоявшейся» задачи может способствовать создание развитой инженерии когнитивного менеджмента, включающей две базовые компоненты: единую концептуальную схему, объединяющую все многообразие существующих стратегический концепций, и систематизированную библиотеку когнитивных инструментов, позволяющих реализовать эти концепции в конкретных стратегических проектах.

По-видимому, наиболее эффективным способом решения этой задачи может быть конвергенция идей и методов когнитивной школы стратегического менеджмента и эмпирического опыта, накопленного в инженерии знаний [9]. В статье предлагается одна из возможных версий концептуальной схемы, построенная таким способом. Приводятся полученные в рамках этой схемы результаты разработки и тестирования когнитивных моделей для четырех исследовательских проектов, отличающихся отраслевой ориентацией, внутренними условиями и различной динамикой макроэкономического окружения. В этих проектах не ставилась цель создания коммерческих инструментов поддержки. Основная цель, которую преследовали авторы проектов, состояла в том, чтобы проверить возможность использования единой концептуальной схемы при создании прикладных моделей для предприятий с самой различной конфигурацией. Разработки носили исследовательский характер и выполнялись с 2004 по 2018 годы четырьмя самостоятельными группами.

Ниже приводятся:

(1) общая характеристика когнитивного подхода к проблеме моделирования стратегического менеджмента, отражающая особенности управления предприятиями в условиях нестабильного экономического окружения;

(2) предлагаемая концептуальная схема когнитивного менеджмента, интегрирующая современный когнитивный опыт в рамках единого онтологического проекта;

(3) краткая характеристика когнитивных моделей, разработанных при реализации указанных проектов;

(4) полученные нами некоторые, на наш взгляд, наиболее значимые результаты разработки и тестирования моделей, которые, хотя и учитывают особенности конкретных проектов, но, тем не менее, могут представить интерес для широкого круга специалистов, интересующихся инжиниринговой проблематикой.

Выводы, приведенные в заключительной части, отражают не только мнение разработчиков моделей, но и мнения стейкхолдеров проектов, а также опыт разработки когнитивных моделей в других областях.

1. Общая характеристика когнитивного подхода

Современные представления когнитивного подхода к вопросу управления сложными проблемными ситуациями изложены в многочисленных публикациях, например, [4, 5, 10]. В рамках этих представлений когнитивное моделирование – это моделирование на основе когнитивных карт, выполняемое с целью исследования проблемной ситуации и поиска оптимальной (в том или ином смысле) стратегии управления этой ситуацией. Основными компонентами когнитивных моделей являются когнитивная карта и методы ее анализа.

1.1. Когнитивная карта

Когнитивная карта – это эксплицитное представление «ментальных моделей» [11] субъектов управления о концептуальной структуре, законах или закономерностях проблемной ситуации.

Основными компонентами когнитивной карты являются:

(1) множество базисных факторов, характеризующих проблемную ситуацию (объект управления и его внешнюю среду);

(2) причинно–следственные (каузальные) отношения между базисными факторами, представляемые с помощью тех или иных формализмов.

В настоящее время в подавляющем большинстве случаев для конструирования когнитивных карт используется формализм орграфов [12]. Вершинами орграфа являются факторы проблемной ситуации, а дугами – каузальные отношения между ними.

На множестве базисных факторов задаются:

- подмножество целевых факторов, отражающих желаемое состояние объекта управления;
- подмножества неуправляемых и управляемых факторов объекта управления;
- подмножество факторов внешней среды, которые могут оказать воздействие на проблемную ситуацию.

Для каждого из факторов устанавливаются его текущее значение или тенденция его изменения. Для каузальных отношений устанавливаются характер влияния факторов-причин на факторы-следствия.

Управление ситуацией заключается в таком изменении управляемых факторов объекта управления, которое приводило бы к желательным изменениям целевых факторов.

Следует отметить, что сегодня наиболее популярными в когнитивных исследованиях являются карты на основе орграфов, однако спектр формализмов, которые могут быть использованы при конструировании когнитивных карт существенно шире. Наряду с формализмом орграфов при разработке когнитивных карт могут быть использованы и такие формализмы, как фреймы М. Минского, генетические сети, реляционные матрицы Американской ассоциации качества, сценарные сети и др.

1.2. Методы анализа когнитивных карт

Методы анализа когнитивных карт разрабатываются для проведения модельных экспериментов над когнитивной картой с целью решения широкого круга управленческих задач. Такими задачами могут быть, например, выбор стратегических целей предприятия, стратегическая диагностика внешней и внутренней среды предприятия, разработка «оптимальной» (том или ином смысле) стратегии предприятия в условиях неопределенной и динамичной экономической среды, аудит стратегии, оценка ее устойчивости и эффективности, корректировка стратегии в меняющихся условиях и др.

2. Концептуальная схема когнитивного менеджмента

Задача поиска оптимальной стратегии в сегодняшних бизнес-средах исключительно сложна. Она в полной мере может быть отнесена к классу сложных задач, решение которых находится вне компетенции традиционной теории стратегического управления.

«Феномен сложности» обусловлен пятью важнейшими особенностями современного стратегического менеджмента, с которыми управленческая практика сегодня сталкивается повсеместно. Такими особенностями являются: уникальность каждого из стратегических проектов, многофакторность, многоаспектность (мультидисциплинарность), динамичность и неопределенность задачи стратегического выбора, высокая роль ментальности разработчиков стратегии, а также лиц, принимающих стратегические решения.

Известны попытки создания универсальных инструментов когнитивной поддержки для лиц, решающих стратегические задачи [13–16]. Однако практика показывает, что в столь сложных и многообразных условиях, в которых работают современные предприятия, создание универсальных инструментов бесперспективно и себя не оправдывает. Необходимы не универсальные инструменты, а некая единая методология, позволяющая конструировать когнитивные модели для конкретных предприятий, на конкретный период времени с учетом стратегического видения владельцев и менеджеров этого предприятия [17].

Первоочередным шагом при формировании такого рода методологии является разработка концептуальной схемы когнитивного менеджмента, единой для всех ее приложений. При этом принципиальным является то обстоятельство, что разработка концептуальной схемы в столь сложной проблемной среде не может быть сделана на основе традиционных формально-аксиоматических подходов. Более уместным здесь, по-видимому, является эмпирический подход, широко используемый в «knowledge based» технологиях [18].

В рамках этого подхода концептуальная схема когнитивного менеджмента может быть представлена следующим образом:

$$P(CM)\colon S^o(C) \Rightarrow S^c(C)\,|_{U(P)},$$

где $P(CM)$ – полное знание проблемной области когнитивного менеджмента;

$S^o(C)$ – текущее состояние анализируемой бизнес-ситуации, задаваемое на когнитивной карте;

$S^c(C)$ – целевое состояние анализируемой бизнес-ситуации, задаваемое на когнитивной карте;

$U(P)$ – стратегия управления, устанавливающая последовательность стратегических шагов, обеспечивающих перевод бизнес ситуации из S^o в S^c.

Ясно, что полное знание $P(CM)$ должно отражать накопленный теоретический и практический опыт проблемной области.

Изучение и критический анализ обширного материала, посвященного когнитивному моделированию управленческих задач [19–26], а также личный опыт авторов дают основание полагать, что полное знание $P(CM)$ может быть представлено в виде онтологического проекта, включающего следующие разделы:

1. Прикладные задачи, которые могут быть решены с помощью инженерии когнитивного менеджмента. При этом задачи могут быть представлены

в виде отдельных вопросов, на которые могут быть получены ответы и, таким образом, они является предметом инженерии.

2. Набор постулатов или аксиом, которые показывают, какие допущения были сделаны в ходе разработки инженерии. Аксиомы или постулаты описывают условия и границы применимости инженерии.

3. Перечень концепций стратегического менеджмента, сложившихся в настоящее время в управленческой науке и практике.

4. Систематизированная библиотека прикладных инструментов, позволяющих реализовать эти концепции в конкретных стратегических проектах.

4.1 Инструменты поддержки для выбора концепции стратегического управления предприятием.

4.2. Инструменты поддержки для выбора языка представления (формализма) когнитивных карт.

4.3. Инструменты поддержки для структурно-функциональной идентификации и параметризации когнитивных карт.

4.5. Множество методов анализа когнитивных карт для решения прикладных задач.

Представленная структура P(CM) по сути ела является «протофреймом» инженерии и решает две важные задачи.

Во-первых, она расширяет традиционное представление о схемах решения прикладных задач стратегического менеджмента путем включения этапов постулирования когнитивного анализа, выбора концепции стратегического менеджмента, выбора адекватного языка представления когнитивных карт, проведения структурно-функциональной идентификации и параметризации когнитивных карт. Без решения этих вопросов когнитивный бизнес анализ в большинстве случаев теряет всякий практический смысл.

Во-вторых, она систематизирует направления усилий по развитию и накоплению прикладных возможностей когнитивного анализа при решении важных для управленческой практики вопросов. К их числу, в частности, относятся:

✦ выявление противоречий между целями, устанавливаемыми субъектами управления;

✦ анализ эффективности управляемых факторов когнитивной карты и их важности по степени влияния на установленные цели;

✦ конструирование различных вариантов стратегий управления («стратегия саморазвития» и различные варианты «стратегий управляемого развития»);

✦ моделирование динамики альтернативных стратегий управления при различных сценариях развития внешней среды и выбор оптимальной (в том или ином смысле) стратегии;

✦ исследование устойчивости выбранной стратегии в критических ситуациях, обусловленных возможными угрозами внешней среды;

✦ мониторинг стратегии в процессе ее реализации;

✦ ретроспективный анализ адекватности когнитивной карты и ее корректировка.

Продукционная часть онтологического проекта сегодня включает массу публикаций, связанных с решением главной задачи стратегического менеджмента – выбору стратегии развития предприятия. Предварительная систематизация этих работ может быть сделана с помощью дескрипторов, характеризующих их инжиниринговую эффективность:

1. горизонт анализа, на который ориентирован инструмент (краткосрочный, среднесрочный, долгосрочный);

2. характер внешней среды предприятий, для которого разработан инструмент (статическая, динамическая);

3. этап разработки стратегии, для которого предназначен инструмент (стратегическая идентификация внешней и внутренней среды предприятия, концептуализация стратегического видения владельцев и топ-менеджмента, формализация стратегического видения в форме когнитивной карты, разработка (выбор) методов анализа карты, тестирование когнитивной модели [27]);

4. уровень разработки инструмента (теоретическое предложение, демонстрационный прототип, исследовательский прототип, действующий прототип, промышленная версия [9]).

Данную систематизацию мы используем для характеристики когнитивных моделей, разработанных при реализации четырех упомянутых проектов.

3. Проекты когнитивного анализа

Проект 1. Оценка стратегических перспектив оффшорной нефтяной компании в регионе Каспийского моря, условия функционирования ко-

торой характеризуются типичными для настоящего времени тенденциями: колебаниями мировых цен на нефть, обусловленными геополитическими факторами, глобальным финансовым кризисом, ростом государственного регулирования, ужесточением экологических нормативов. Работа выполнялась совместно с нефтепромысловым факультетом (Petroleum Engineering Department) Техасского A&M университета (College Station, Техас, США) в рамках договора о творческом сотрудничестве с компанией BP Azerbaijan.

Проект 2. Выявление ключевых компетенций для задачи стратегического развития телекоммуникационной компании – регионального дилера Microsoft, Cisco Systems, HP, Intel. Компания занимается разработкой, установкой и обслуживанием промышленных, административных и медицинских информационных систем, а также систем iP- и CTi-телефонии для операторов связи и корпоративных клиентов.

Проект 3. Анализ эффективности менеджмента на птицеводческом предприятии одного из холдингов города Баку. Задача была обусловлена критической ситуацией, сложившейся в связи с увеличением цен на сырье и корма, уменьшением цен на готовую продукцию, ростом конкурентного давления, трудностями получения коммерческих кредитов, недостаточной квалификацией управленческого персонала, а также снижением доли прибыли, идущей на рефинансирование предприятия (в частности, из–за коррупционного давления контролирующих органов).

Проект 4. Выбор стратегии регионального ИКТ-менеджмента. Проект выполнялся при финансовой поддержке Министерства связи и информационных технологий Азербайджана. Актуальность задачи была обусловлена тем, что в связи с переходом Республики к информационной (цифровой) экономике критически важным становился вопрос влияния информационно-коммуникационных технологий (ИКТ) на экономический рост различных субъектов экономики. Наиболее остро данная проблема стояла на региональном уровне. Эта проблема довольно сложна. До настоящего времени она является предметом многочисленных дискуссий и уже более тридцати лет известна в экономической практике как парадокс Р. Соллоу («парадокс производительности») [28].

Проблема носит многофакторный, неопределенный и динамичный характер и не поддается решению с помощью традиционных методов эконометрики. Интересующиеся этим вопросом могут заглянуть в интернет и убедиться, как много безуспешных попыток было предпринято в международной практике для решения этого вопроса. В частности, авторами таких попыток использовались такие методы, как корреляционный и регрессионный анализ, методы оценки на основе производственной функции Кобба –Дугласа, исследования на основе методологии анализа экономического эффекта (economic impact analysis). Также можно упомянуть проект iSociety, поддержанный компаниями Microsoft и PricewaterhouseCoopers, по влиянию ИКТ на эффективность повседневного бизнеса в Великобритании, а также исследование McKinsey, в рамках которого анализировались каналы влияния ИКТ на производительность на основе изучения соответствующих бизнес-процессов.

В нашем случае для решения поставленной задачи был использован когнитивный подход [29]. В качестве концептуальной основы на этапе выбора стратегической концепции были использованы результаты крупномасштабного исследования Economist Intelligence Unit (EIU, 2004), представляющего собой сочетание методологии системного анализа и «business review».

Модельные эксперименты, выполненные в процессе реализации данного проекта, показали, что когнитивное моделирование открывает принципиально новые возможности не только для оценки экономического влияния ИКТ, но и для управления этим влиянием, т.е. для осуществления эффективной стратегии ИКТ-менеджмента.

4. Краткая характеристика разработанных когнитивных моделей

Основными характеристиками моделей, разработанных в вышеперечисленных проектах, являлись:

- ✦ принятая концепция стратегического менеджмента;
- ✦ формализм когнитивной карты, отражающий логику принятой концепции;
- ✦ методы построения и анализа когнитивной карты;
- ✦ шкалы оценок значений базисных факторов и каузальных связей карты.

Характеристики моделей, которые были разработаны в указанных проектах, приведены в *таблице 1.*

Таблица 1.

Характеристики когнитивных моделей, разработанных при реализации проектов

№ проекта (уровень разработки проекта)	Принятая концепция бизнес-стратегии	Структурный тип КК	Методы построения и анализа КК	Шкалы оценок факторов и тенденций
Проект 1. (ДП)	«Ресурсно–ориентированная» (R. Grant, 1987)	Орграф	**Методы построения:** SWOT–анализ. PESTLE–анализ. Имплиативные репертуарные решетки С. Хинкла. Матрица взаимодействий (L. Jones, 1982) **Методы анализа:** Модели линейной динамики Ф. Робертса. Метод сценарного анализа.	Биполярная лингвистическая 5–балльная
Проект 2. (ИП)	«Школа компетенций» (C. Prahalad, G. Hamel, 1990)	Реляционные матрицы (матричная XL–диаграмма ASQ) [8]	**Методы построения:** Диаграммы Исикавы. VRIO–анализ (J. Barney, 1987). Модель бизнес–процессов eTOM.4.0 (Intern. Telecommunication Union/Standardization Sector). «Модель мнений потребителя» (IBM, США). «Модель ресурсов» (American Society For Quality, США). **Методы анализа:** Нечеткий каузальный матричный анализ [8]. Метод анализа иерархий Т.Саати.	Лингвистическая 10–балльная
Проект 3. (ИП)	«Стратегия миграции» [1]	Орграф	**Методы построения:** SWOT–анализ. PESTLE–анализ. «Модель конкурентного анализа» (М. Портер, 2003). Методы психосемантики и многомерного неметрического шкалирования (М. Девидсон, 2003). **Методы анализа:** Качественные модели динамики на основе правил [5] и логики времени (Поспелов, 1986).	Биполярная лингвистическая 5–балльная
Проект 4 (ИП)	Концепция Economist Intelligence Unit (EIU, 2004)	Фреймы Минского	**Методы построения:** Системный анализ + Business review PESTLE–анализ **Методы анализа:** Метод сценарного моделирования	Лингвистическая 10–балльная

Условные обозначения: ДП – демонстрационный прототип; ИП – исследовательский прототип;
МП–методы построения КК; МА – методы анализа КК.

Примеры построения когнитивных карт можно найти в работах [29, 30].

5. Обсуждение

Опыт разработки и тестирования когнитивных моделей в указанных проектах показал следующее:

1. В практических задачах деловой сферы когнитивное моделирование сталкивается с теми же трудностями, что и другие интеллектуальные (knowledge-based) технологии. Это проблема качественного экспертного знания, используемого при построении когнитивных моделей, всевозможные «ловушки» и «парадоксы экспертизы» [9], возникающие при работе с экспертами, трудности выбора адекватного уровня детализации когнитивных карт, настоятельная необходимость в специалисте–посреднике (инженере по знаниям, метаинтерпретаторе–когнитологе) и др.

2. Когнитивное моделирование приобретает практическую значимость лишь в рамках методологии стратегической диагностики, включающей этапы макроэкономического и маркетингового анализа в контексте динамики внешней среды. Игнорирование этих этапов редуцирует когнитивные модели в математические объекты, далекие от реалий бизнес-практики.

3. Одним из ключевых вопросов когнитивного менеджмента является вопрос выбора разумного уровня детализации («грануляции») когнитивных карт. Вопрос связан с тем, что использование когнитивных карт с низким уровнем детализации часто приводит к потере деталей, в которых, как известно, «прячется дьявол». Критический обзор литературных источников и обсуждения с коллегами привели нас к выводу, что вопрос требует дальнейшей проработки и может рассматриваться как одно из направлений дальнейшего совершенствования когнитивного анализа, в частности, разработке эффективного механизма «мультимасштабирования», аналогичного схеме предложенной корпорацией РЭНД Корпорейшн [31].

4. Серьезные трудности возникают при сценарном анализе разрабатываемых стратегий. Немонотонная динамика внешней среды и необходимость многовариантного анализа стратегий в широком «коридоре сценариев» требуют разработки инженерной техники моделирования, отличной от той, которую предлагает сегодня высшая математика [19, 22].

5. Сотрудничество с работниками предприятий показало, что у них не было однозначного мнения относительно эффективности когнитивного моделирования. Вместе с тем, 12 из 17 респондентов проявили активный интерес и посчитали, что когнитивное моделирование – это перспективная и полезная технология. При этом наши наблюдения выявили очень важную латентную особенность когнитивного анализа: он стимулирует познавательную и творческую активность разработчиков стратегии в наиболее сложной и критически важной фазе «стратегического мышления»– фазе афферентного синтеза стратегических решений.

6. Работа в проектах убедительно подтвердила тот факт, что адекватность когнитивных моделей, в первую очередь, зависит от «качества» знаний, закладываемых в ее основу, носителями которых является сами разработчики стратегии, их компетенции и профессиональный опыт. Когнитивное моделирование лишь усиливает и расширяет их аналитический потенциал.

7. Сегодняшняя высокая изменчивость бизнес-сред предъявляет жесткие требования к срокам конструирования и тестирования когнитивных моделей. Это обстоятельство составляет принципиальное отличие когнитивного менеджмента от многих других когнитивных приложений. В связи с этим критически важным становится вопрос создания эффективных инструментов поддержки (языки высокого уровня, системные оболочки, сценарные сети) для «быстрой» разработки и тестирования когнитивных моделей – вопрос, который сегодня полностью игнорируется теоретиками когнитивной школы.

Заключение

Отметим некоторые соображения общего характера, касающихся достоинств и недостатков когнитивного моделирования, которые, как показывает анализ немногочисленных прикладных исследований, актуальны и для других приложений когнитивного подхода.

1. Как и во всех «knowledge-based» технологиях, необходимым условием успеха являются, во-первых, наличие качественных знаний у привлекаемых экспертов (топ-менеджеры, бизнес-консультанты) и, во-вторых, участие в когнитивном процессе инженера по знаниям, играющего ведущую роль на начальных этапах когнитивного процесса.

2. Принципиальными достоинствами когнитивного подхода являются:

- возможность исследования динамики бизнес-ситуации в условиях сложной быстроменяющейся PEST-среды, когда имеющихся данных недостаточно для построения полной имитационной модели;
- возможность исследования бизнес-ситуаций с учетом многофакторной «институциональной оболочки» предприятия [17];
- возможность исследования бизнес-ситуаций при наличии мультипликативных и обратных связей между факторами среды, а также при наличии различных пороговых эффектов.

Ни один из многочисленных традиционных инструментов поддержки стратегического менеджмента такими возможностями не обладает.

3. Когнитивный анализ открывает новые перспективы для теории принятия решений. Возможность целенаправленной генерации оптимальных стратегий, отсутствующая в известных коммерческих DSS-пакетах, определяет принципиально новый подход к принятию решений: не «выбор лучшего решения из множества имеющихся альтернативных» (парадигма РЭНД Корпорейшн), а целенаправленная генерация «лучшего решения».

4. Тестирование когнитивных моделей выявило и ряд трудностей:

- из-за дискретной структуры моделей обеспечивается лишь грубая аппроксимация непрерывных процессов. Приходится внимательно относиться к последовательности факторов в каузальных сетях и учитывать, совпадают ли во времени воздействия от одних факторов на другие, или они смещены друг относительно друга;
- приходится быть осторожным при параметризации когнитивных карт, особенно при оценке силы каузальных связей, которые могут меняться в процессе сценарных трансформаций;
- параметризация когнитивных моделей в случае сложных (многофакторных) когнитивных карт сталкивается с «проблемой целостности». Широко распространенное мнение о том, что значения каждого из факторов и каждой из каузальных связей могут быть определены в индивидуальном порядке, является, на наш взгляд, глубоко ошибочным. По нашим наблюдениям, эти оценки хоть и выставляются в индивидуальном порядке, но являются сильно коррелированными с гештальтом проблемной ситуации, который формируется на интуитивном уровне у разработчиков стратегии [32]. Это обстоятельство сильно ограничивает рабочие форматы когнитивных карт и делает малопродуктивной работу с популярными крупноформатными картами, в которых число базисных факторов и каузальных связей исчисляется десятками или даже сотнями.

5. Сложная экономическая среда, в которой сегодня работают предприятия, существенно ограничивает возможности традиционного экономико-математического моделирования. В задачах стратегического выбора актуальным становится моделирование на основе знаний (knowledge-based modeling). Современные стратегические исследования фактически превращаются в сложное инженерное искусство [18], формирующее новое перспективное направление управленческой бизнес аналитики. Критический анализ показывает, что когнитивный менеджмент в настоящее время является единственной безальтернативной парадигмой, способной обеспечить успешную реализацию этого направления [33]. Но вместе с тем он показывает, что в инженерии когнитивного менеджмента сегодня остается множество вопросов, без решения которых когнитивная школа менеджмента, как и тридцать лет назад, будет оставаться скорее заманчивым и многообещающим потенциалом, чем практическим инструментом. ■

Литература

1. Ньюман У. Управленческое действие: Техника организации и менеджмента. Колумбийский университет, 1951.
2. Томпсон-мл. А.А., Стрикленд III А.Дж. Стратегический менеджмент: концепции и ситуации для анализа. 12-е изд. М.: Вильямс, 2009.
3. Drucker R. Management challenges for the 21st century. London, New York: Routledge. Taylor & Francis Group, 2012.
4. Narayanan V.K., Zane L.K., Kemmerer B. The cognitive perspective in strategy: An integrative review // Journal of Management. 2011. Vol. 37. No 1, pp. 305–323. DOI: 10.1177/0149206310383986.
5. Hodginson G. Cognitive process in strategic management: Some emerging trends and future direction // Handbook of industrial, work & organizational psychology. London: SAGE Publication. 2011. Vol. 2. P. 401–441.
6. Ross D. Economic theory and cognitive science: Microexplanation. MIT Press, 2005.
7. Mintzberg H., Lampel J., Ahlstrand B. Strategy safari: A guide tour through the wilds of strategic management. New York: Free Press, 2005.

8. The Oxford handbook of cognitive engineering. Oxford library of psychology / Eds. J.D. Lee, A. Kirlik, M.J. Dainoff. New York: OUP USA. 2013.
9. Waterman D.A. Guide on expert systems. Addison-Wesley, 1986.
10. Hodgkinson G.P., Healey M.P. Cognition in organizations // Annual Review of Psychology. 2007. No 59. P. 387–417. DOI: 10.1146/annurev.psych.59.103006.093612.
11. Johnson-Laird P.N. Mental models in cognitive science // Cognitive Science.1980. No 4. P. 71–115. DOI: 10.1207/s15516709cog0401_4.
12. Roberts F. Discrete mathematical models with application to social, biological and environmental problems. Englewood Cliffs, New Jersey: Rutgers University, Prentice-Hall, 1976.
13. Schwenk C.R. The cognitive perspective on strategic decision making // Journal of Management Studies. 1988. Vol. 25. No 1. P. 41–55. DOI: 10.1111/j.1467-6486.1988.tb00021.x.
14. Haleblian J., Rajagopalan N. A cognitive model of CEO dismissal: Understanding the influence of board perceptions, attributions and efficacy beliefs // Journal of Management Studies. 2006. No 43. P. 1009–1026. 10.1111/j.1467-6486.2006.00627.x.
15. Porac J.F., Thomas H. Managing cognition and strategy: Issues, trends and future directions // A. Pettigrew, H. Thomas, R. Whittington (Eds.) Handbook of strategy and management. London: SAGE Publisher. 2002. P. 165–181.
16. Eden C., Ackermann F. SODA – The principles // Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict / J. Rosenhead, J. Mingers (Eds.). Chichester: Wiley, 2001. P. 21–41.
17. Балацкий Е. Диалектика познания и новая парадигма экономической науки // Мировая экономика и международные отношения. 2006. № 7.
18. Поспелов Д.А. Теория и практика ситуационного управления. М.: Наука. 2001.
19. IEEE Proceedings of the International Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) (2011–2019) series [Электронный ресурс]: https://edas.info/web/cogsima2019/home.html (дата обращения 20.02.2020).
20. Best value business & consulting toolkits. Created by management consultants previously from: Deloitte, McKinsey&Company, BCG Strategy Consultants. [Электронный ресурс]: https://www.slidebooks.com/products/strategy-toolkit?gclid=EAIaIQobChMIhfSMj8qz4gIVgobVCh1xTwW2EAMYASAAEgJ8XfD_BwE&variant=17648810693 (дата обращения 20.02.2020).
21. Gallopin G.C. Modeling incompletely specified complex systems // Third International Symposium on Trends in Mathematical Modelling, S.C. Bariloche, December 1976, UNESCO-Fundacion Bariloche, 1977.
22. IEEE Proceedings of the International Conference on Cognitive Modeling (ICCM) series (1996–2019). [Электронный ресурс]: https://iccm-conference.github.io/previous.html (дата обращения 20.02.2020).
23. Авдеева З.К., Коврига С.В. Формирование стратегии развития социально-экономических объектов на основе когнитивных карт. Saarbrucken: LAP LAMBERT Academic Publishing, 2011.
24. Images of strategy / S. Cummins [et al.]. Oxford: Blackwell Publishing, 2003.
25. Родин Д.В. Концептуальные подходы к формированию и реализации организационных стратегий. // Системное управление. 2008. Выпуск 1 (2). [Электронный ресурс]: http://sisupr.mrsu.ru/wp-content/uploads/2014/11/21-rodin.pdf (дата обращения 20.02.2020).
26. Issues in cognitive modeling / Eds. A.M. Aithenhead, J.M. Slack. Lea: Hove, 1994.
27. Constructing an expert system / B. Buchanan [et al.] // F. Hayes-Rath, D. Waterman, D. Lenet (eds.) Building expert system. Addison-Wesley, 1983.
28. Solow R. We'd better watch out. Book review // New York Times. 12 July 1987.
29. Karayev R.A. Choice of a strategy of regional ICT-management. Cognitive paradigm // International Journal of Managing Information Technology. 2013. Vol. 5. No 3. P. 17–30. DOI: 10.5121/ijmit.2013.5303.
30. Karayev R.A. Cognitive approach and its application to the modeling of strategic management of enterprises // Knowledge engineering: Principles, methods and applications (Ed. A. Perez Gama). New York: Nova Science, 2015. P. 79–101.
31. Davis P.K., Kahan J.P. Theory and methods for supporting high-level decision making. Santa Monica, CA: RAND Corporation, TR-422-AF, 2007.
32. Bays P.M., Husain M. Dynamic shifts of limited working memory resources in human vision // Science. 2008. Vol. 321. No 5890. P. 851–854. DOI: 10.1126/science.1158023.
33. Полтерович В.М. Институциональные ловушки и экономические реформы // Экономика и математические методы. 1999. Т.35. № 2. С. 3–20.

Об авторах

Караев Роберт Асадулла оглы

доктор технических наук; профессор Международной Экоэнергетической Академии;

руководитель лаборатории моделирования экологических систем, Институт систем управления,
Национальная академия наук Азербайджанской Республики,
Азербайджанская Республика, AZ1141, г. Баку, ул. Б. Вагабзаде, д. 9;

E-mail: karayevr@rambler.ru

Садыхова Нателла Юсиф кызы

научный сотрудник, Институт систем управления,
Национальная академия наук Азербайджанской Республики,
Азербайджанская Республика, AZ1141, г. Баку, ул. Б. Вагабзаде, д. 9;

E-mail: natella5@rambler.ru

The advantages of cognitive approach for enterprise management in modern conditions

Robert A. Karayev
E-mail: karayevr@rambler.ru

Natella Yu. Sadikhova
E-mail: natella5@rambler.ru

Institute of Control Systems, National Academy of Sciences of Azerbaijan
Address: 9, B. Vahabzade Street, Baku AZ 1141, Azerbaijan

Abstract

The paper provides a brief description of cognitive management, which opens up unique opportunities for the effective management of enterprises in modern complex and unstable conditions. The problems of commercializing this promising paradigm are discussed. It is pointed out that the main, critical one of these problems is due to the lack of developed engineering of cognitive management. A conceptual framework for solving this problem is proposed, based on the convergence of the ideas and methods of the "cognitive school" and the empirical experience gained in knowledge engineering. The results of using the conceptual framework in four research projects of different industry orientations, with different internal conditions and different dynamics of the external environment are presented. The engineering prospects of the proposed framework are discussed in terms of the commercialization of the cognitive school identified by H. Mintzberg, B. Ahlstrand and D. Lampel 30 years ago.

Key words: cognitive management; conceptual framework; analysis and choice technology; cognitive approach.

Citation: Karayev R.A., Sadikhova N.Yu. (2020) The advantages of cognitive approach for enterprise management in modern conditions. *Business Informatics*, vol. 14, no 2, pp. 36–47. DOI: 10.17323/2587-814X.2020.2.36.47

References

1. Newman W. (1951) *Management action: Organization and management technique.* Columbian University (in Russian).
2. Thompson Jr. A.A., Striklend III A.G. (2009) *Strategic management: concepts and situations for the analysis.* 12 ed. Moscow: Williams (in Russian).
3. Drucker R. (2012) *Management challenges for the 21st century*. London, New York: Routledge. Taylor & Francis Group.
4. Narayanan V.K., Zane L.K., Kemmerer B. (2011) The cognitive perspective in strategy: An integrative review. *Journal of Management*, vol. 37, no 1, pp. 305–323. DOI: 10.1177/0149206310383986.
5. Hodginson G. (2011) Cognitive process in strategic management: Some emerging trends and future direction. *Handbook of industrial, work & organizational psychology*. London: SAGE Publication, vol. 2, pp. 401–441.
6. Ross D. (2005) *Economic theory and cognitive science: Microexplanation*. MIT Press.
7. Mintzberg H., Lampel J., Ahlstrand B. (2005) *Strategy safari: A guide tour through the wilds of strategic management*. New York: Free Press.
8. Lee J.D., Kirlik A., Dainoff M.J., eds. (2013) *The Oxford handbook of cognitive engineering. Oxford library of psychology*. New York: OUP USA.
9. Waterman D.A. (1986) *Guide on expert systems*. Addison-Wesley.
10. Hodgkinson G.P., Healey M.P. (2007) Cognition in organizations. *Annual Review of Psychology*, no 59, pp. 387–417. DOI: 10.1146/annurev.psych.59.103006.093612.
11. Johnson-Laird P.N. (1980) Mental models in cognitive science. *Cognitive Science*, no 4, pp. 71–115. DOI: 10.1207/s15516709cog0401_4.
12. Roberts F. (1976) *Discrete mathematical models with application to social, biological and environmental problems.* Englewood Cliffs, New Jersey: Rutgers University, Prentice-Hall.
13. Schwenk C.R. (1988) The cognitive perspective on strategic decision making. *Journal of Management Studies*, vol. 25, no 1, pp. 41–55. DOI: 10.1111/j.1467-6486.1988.tb00021.x.

14. Haleblian J., Rajagopalan N. (2006) A cognitive model of CEO dismissal: Understanding the influence of board perceptions, attributions and efficacy beliefs. *Journal of Management Studies*, no 43, pp. 1009–1026. 10.1111/j.1467-6486.2006.00627.x.

15. Porac J.F., Thomas H. (2002) Managing cognition and strategy: Issues, trends and future directions. *Handbook of strategy and management* (Eds. A. Pettigrew, H. Thomas, R. Whittington). London: SAGE Publisher, pp. 165–181.

16. Eden C., Ackermann F. (2001) SODA – The principles. *Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict* (Eds. J. Rosenhead, J. Mingers). Chichester: Wiley, pp. 21–41.

17. Balatsky E. (2006) The dialectic of cognition and the new paradigm of economic science. *World Economy and International Relations*, no 7 (in Russian).

18. Pospelov D.A. (2001) *Theory and practice of situational management*. Moscow: Nauka (in Russian).

19. *IEEE Proceedings of the International Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) (2011–2019) series.* Available at: https://edas.info/web/cogsima2019/home.html (accessed 20 February 2020).

20. *Best value business & consulting toolkits. Created by management consultants previously from*: Deloitte, McKinsey&Company, BCG Strategy Consultants. Available at: https://www.slidebooks.com/products/strategy-toolkit?gclid=EAIaIQobChMIhfSMj8qz4gIVgobVCh1xTwW2EAMYASAAEgJ8XfD_BwE&variant=17648810693 (accessed 20 February 2020).

21. Gallopin G.C. (1977) Modeling incompletely specified complex systems. *Third International Symposium on Trends in Mathematical Modelling, S.C. Bariloche, December 1976*, UNESCO-Fundacion Bariloche.

22. *IEEE Proceedings of the International Conference on Cognitive Modeling (ICCM) series* (1996–2019). Available at: https://iccm-conference.github.io/previous.html (accessed 20 February 2020).

23 Avdeeva Z.K., Kovriga S.V. (2011) *Formation of a strategy for the development of socio-economic objects based on cognitive maps.* Saarbrucken: LAP LAMBERT Academic Publishing.

24. Cummins S., Wilson D., Angwin D, Bilton C., Brocklesby J., Doyle P., Galliers B., Legge K., McGee J., Newell S., Pettigrew A., Smith C., Wensley R. (2003) *Images of strategy*. Oxford: Blackwell Publishing.

25. Rodin D.V. (2008) Conceptual approaches to the formation and implementation of organizational strategies. *System Management*, issue 1(2). Available at: http://sisupr.mrsu.ru/wp-content/uploads/2014/11/21-rodin.pdf (accessed 20 February 2020) (in Russian).

26. Aithenhead A.M., Slack J.M., Eds. (1994) *Issues in cognitive modeling.* Lea: Hove.

27. Buchanan B., Bechtal R., Bennett J., Clancey W., Kulikowski C., Mitchell T., Waterman D.A. (1983) Constructing an expert system. *Building expert system* (F. Hayes-Rath, D. Waterman, D. Lenet, eds.). Addison-Wesley, 1983.

28. Solow R. (1987) We'd better watch out. Book review. *New York Times*, 12 July 1987.

29. Karayev R.A. (2013) Choice of a strategy of regional ICT-management. Cognitive paradigm. *International Journal of Managing Information Technology*, vol. 5, no 3, pp. 17–30. DOI: 10.5121/ijmit.2013.5303.

30. Karayev R.A. (2015) Cognitive approach and its application to the modeling of strategic management of enterprises. *Knowledge engineering: Principles, methods and applications* (Ed. A. Perez Gama). New York: Nova Science, pp. 79–101.

31. Davis P.K., Kahan J.P. (2007) *Theory and methods for supporting high-level decision making.* Santa Monica, CA: RAND Corporation, TR-422-AF.

32. Bays P.M., Husain M. (2008) Dynamic shifts of limited working memory resources in human vision. *Science*, vol. 321, no 5890, pp. 851–854. DOI: 10.1126/science.1158023.

33. Polterovich V.M. (1999) Institutional traps and economic reforms. *Economics and Mathematical Methods*, vol. 35, no 2, pp. 3–20 (in Russian).

About the authors

Robert A. Karayev

Dr. Sci. (Tech.);

Professor, Head of Ecosystems Modeling Laboratory, Institute of Control Systems, Azerbaijan National Academy of Sciences,
9, B. Vahabzade Street, Baku AZ1141, Azerbaijan;

E-mail: karayevr@rambler.ru

Natella Yu. Sadikhova

Researcher

Institute of Control Systems, Azerbaijan National Academy of Sciences,
9, B. Vahabzade Str., Baku AZ 1141, Azerbaijan

E-mail: natella5@rambler.ru

DOI: 10.17323/2587-814X.2020.2.48.63

Многомерное логарифмически нормальное распределение в оценке недвижимого имущества

М.Б. Ласкин

E-mail: laskinmb@yahoo.com

Санкт-Петербургский институт информатики и автоматизации, Российская академия наук
Адрес: 199178, г. Санкт-Петербург, 14 линия, д. 39

Аннотация

Целью исследования является разработка методики оценки рыночной стоимости по совокупности ценообразующих факторов, подчиняющихся совместному логарифмически нормальному распределению. Под совместным логарифмически нормальным распределением понимается распределение случайного вектора, логарифмы компонент которого распределены совместно нормально. Предложен метод оценки рыночной стоимости по условному распределению цен при заданных значениях ценообразующих факторов, а также методы анализа цены предложения с точки зрения ее обоснованности, по условному распределению вектора ценообразующих факторов при заданной цене предложения. Рассмотрены особенности коэффициента застройки в зависимости от площади земельного участка. Приведены дополнительные аргументы в пользу оценки рыночной стоимости как моды условных законов распределения цен. Рассмотрен пример многомерного логарифмически нормального распределения цены и таких ценообразующих факторов, как площадь улучшений (устоявшийся в оценочном сообществе термин, означающий общую площадь возведенных на земельном участке зданий и сооружений) и площадь земельного участка на реальных данных, т.е. для случая трехмерного случайного вектора. Обоснована формула для определения точки абсолютного максимума плотности многомерного логарифмически нормального случайного вектора. Полученные результаты могут быть использованы при создании информационных систем поддержки принятия решений при оценке объектов недвижимого имущества.

Ключевые слова: рыночная стоимость; логарифмически нормальный закон распределения цен; многомерное логарифмически нормальное распределение; оценка недвижимого имущества.

Цитирование: Ласкин М.Б. Многомерное логарифмически нормальное распределение в оценке недвижимого имущества // Бизнес-информатика. 2020. Т. 14. № 2. С. 48–63. DOI: 10.17323/2587-814X.2020.2.48.63

Введение

Одним из распространенных методов оценки рыночной стоимости объектов недвижимого имущества является построение регрессионных зависимостей цены от значений ценообразующих факторов. При этом факторы могут быть как качественными (тип дома, обременение, этажность, вид из окна, состояние квартиры/помещения и т.д.), так и количественными (площадь объекта или земельного участка, расстояние до центра города, до метро, других объектов инфраструктуры и т.д.).

Существуют различные мнения относительно разбиения ценообразующих факторов на классы. В контексте настоящей статьи имеется в виду разбиение на качественные и количественные факторы с точки зрения возможности представления значений фактора вещественным числом (если такая возможность существует, то фактор является количественным). Совмещение в одной регрессионной модели количественных и качественных факторов представляет для аналитиков определенную трудность. Однако, эта проблема выходит за рамки настоящей статьи: здесь мы ограничимся только количественными (вещественными) факторами. Оказывается, что такие факторы, рассматривая их как случайные величины на множестве объектов сравнения, как правило, удается приблизить логарифмически нормальным распределением. Также имеются основания предполагать, что цены объектов сравнения также подчиняются логарифмически нормальному закону распределения.

Теоретическое обоснование формирования логнормальной генеральной совокупности для цен, образованных последовательными сравнениями, было приведено в работе [1]. На факт подчинения арендных ставок в недвижимости логарифмически нормальному распределению указывали еще Айчинсон и Браун в 1963 году [2]. На логарифмическое распределение цен в недвижимости указывали и современные исследователи [3]. Такой подход пока не является традиционным с точки зрения существующей практики оценки недвижимого имущества, поскольку требует применения специальных прикладных статистических пакетов, которые не используются практикующими оценщиками, привыкшими работать с малым количеством объектов сравнения. В то же время быстро меняющаяся информационная среда побуждает исследователей искать новые, нетрадиционные подходы к оценке недвижимого имущества. В качестве примера можно привести работы [4–9], посвященные методу гедонистического ценообразования, т.е. выявления статистической связи между средней или медианной стоимостью жилья, внутренними и внешними ценообразующими факторами.

Статистическая зависимость, как правило, оценивается через модели линейной, логарифмической или частично-логарифмической зависимости. В целом эта же идеология положена в основу отчета о кадастровой стоимости [10], выполненного Санкт-Петербургским ГУ «Кадастровая оценка» в 2018 году. В ряде работ используются нерегрессионные модели оценки объектов жилой недвижимости: например, в работах [11, 12] для прогнозирования стоимости жилых объектов используются нейронные сети, в работах [13, 14] – методы машинного обучения (случайный лес, метод опорных векторов), а в работе [15] сравниваются результаты применения таких методов, как деревья решений, наивный байесовский классификатор и AdaBoost. Такие методы требуют использования больших выборок данных. Другим подходом является использование индексов цен. Например, в работе [16] рассматривается индекс цен на жилье Case-Shiller. В работах [17–19] исследуется индекс повторной продажи, который прогнозирует изменение стоимости перепроданного объекта, исходя из разницы во времени и изменения его атрибутов между первоначальной продажей и последующей перепродажей. Авторы работ [20–23] рассматривают гибридный метод, сочетающий гедонистический подход и метод повторной продажи.

Основным подходом при исследовании ценпузырей является использование вариаций методов авторегрессии, примененных к усредненным ценам, например, в работах [24–28]. Таким образом, привлечение многомерных логарифмически нормальных распределений также находится в русле современных тенденций поиска нетрадиционных методов в оценке недвижимого имущества.

1. Оценка рыночной стоимости по условным распределениям при фиксированных значениях ценообразующих (количественных/вещественных) факторов

Пусть V – цена предложения (или сделки), а $X_1, ..., X_n$ – количественные (вещественные) ценообразующие факторы. Пусть $W = \ln(V)$, $Y_i = \ln(X_i)$, $i = 1, n$ (тогда $v = e^W$, $X_i = e^{Y_i}$).

Рассмотрим многомерный нормальный случайный вектор $(W, Y_1, ..., Y_n)$ с вектором средних

(μ_W, μ_{Y_1}, ..., μ_{Y_n}). Ковариационную матрицу запишем в блочном виде:

$$CV = \begin{pmatrix} \sigma_W^2 & cov\left(W, \vec{Y}\right) \\ cov\left(W, \vec{Y}\right)^T & COV \end{pmatrix},$$

где COV – ковариационная матрица случайного вектора $\vec{Y} = (Y_1, ..., Y_n)$;

$cov\left(W, \vec{Y}\right)$ – вектор $\left(\rho_{WY_1}\sigma_w\sigma_{Y_1}, \ldots, \quad \rho_{WY_n}\sigma_w\sigma_{Y_n}\right)$;

$\sigma_W^2, \sigma_{Y_1}^2 \ldots, \sigma_{Y_n}^2$ – дисперсии случайных величин W, Y_1, ..., Y_n;

$\rho_{WY_1}, \ldots, \quad \rho_{WY_n}$ – соответствующие коэффициенты корреляции.

Тогда, условное математическое ожидание W, при условии $Y_1 = y_1, ..., Y_n = y_n$ равно

$$E\left(W|Y_1 = y_1, \ldots, Y_n = y_n\right) = \\ = \mu_W + \left(COV^{-1} \times cov\left(W, \vec{Y}\right)^T, \left(\vec{Y} - \overrightarrow{\mu_Y}\right)\right),$$

где $\overrightarrow{\mu_Y} = \left(\mu_{Y_1}, \ldots, \mu_{Y_n}\right)$. Условная дисперсия W, при условии, что $Y_1 = y_1, ..., Y_n = y_n$ равна

$$D\left(W|Y_1 = y_1, \ldots, Y_n = y_n\right) = \\ = \sigma_W^2 - \left(COV^{-1} \times cov\left(W, \vec{Y}\right)^T, cov\left(W, \vec{Y}\right)\right).$$

При фиксированных значениях ценообразующих факторов $X_1 = x_1, ..., X_n = x_n$ наиболее вероятное значение цены предложения (или сделки, в зависимости от того, какие цены были в исходных данных) V рассчитывается по формуле условной моды

$$Mode(V|X_1 = x_1 = e^{y_1}, \ldots, X_n = x_n = e^{y_n}) = \\ \exp\left(\mu_W + (COV^{-1} \times cov(W, \vec{Y})\right)^T, \left(\vec{Y} - \overrightarrow{\mu_Y}\right) - \\ -\sigma_W^2 + \left(COV^{-1} \times cov(W, \vec{Y})^T, cov(W, \vec{Y})\right). \quad (1)$$

В соответствии с формулировкой ФЗ-135 [29], под рыночной стоимостью понимается наиболее вероятная цена, по которой объект оценки может быть отчужден на открытом рынке в условиях совершенной конкуренции. На практике оценщики, как правило, используют средние арифметические, реже средние геометрические значения. Такие оценки могут быть построены как оценки условного математического ожидания и условной медианы:

$$E(V|X_1 = x_1 = \exp(y_1), \ldots, X_n = x_n = \exp(y_n)) = \\ = \exp(\mu_W + (COV^{-1} \times cov\left(W, \vec{Y}\right)^T, \left(\vec{Y} - \overrightarrow{\mu_Y}\right)) + \\ \frac{1}{2}\sigma_W^2 - \frac{1}{2}\left(COV^{-1} \times cov\left(W, \vec{Y}\right)^T, \ cov\left(W, \vec{Y}\right)\right)) \ . \quad (2)$$

$$Median\left(V|X_1 = x_1 = \exp(y_1), \ldots, X_n = x_n = \exp(y_n)\right) = \\ = \exp(\mu_W + COV^{-1} \times cov(W, \vec{Y})^T, (\vec{Y} - \overrightarrow{\mu_Y}). \quad (3)$$

Таким образом, если относительно некоторого ансамбля количественных ценообразующих факторов и цен объектов сравнения удается принять в качестве рабочей гипотезу о совместном логарифмически нормальном распределении (совместном нормальном распределении логарифмов) компонент случайного вектора, то в качестве оценки рыночной стоимости могут быть приняты оценки по формуле (1). Могут быть приняты и оценки по формулам (2) и (3); но при этом следует помнить, что такие оценки не отвечают определению рыночной стоимости, закрепленной в ФЗ-135.

Рассмотрим пример, в котором использованы данные, подобранные известными российскими оценщиками и опубликованные на ресурсе [30]. Рассматривается 40 объектов недвижимости производственно-складского назначения с локацией в одном регионе (г. Санкт-Петербург), выставленные на продажу в одном периоде времени. Поскольку авторами примера обоснован отказ от ряда корректировок, в нашем примере мы также будем считать данные сопоставимыми и сравнимыми без дополнительных корректировок. Объекты производственно-складского назначения рассматриваются как единый комплекс, состоящий из земельного участка и улучшений (зданий). Данные представлены в *таблице 1.*

Объекты сравнения рассматриваются как действующие объекты производственно-складского назначения, выставленные на продажу в текущем использовании. Мы построим общий метод оценки рыночной стоимости (без учета скидки на торг), если для него заданы площади улучшений и земельного участка (полученная оценка, естественно, относится к тому же периоду времени, типу объектов и региону).

В данном случае имеют место случайные величины V – приведенная к 1 кв. м улучшений цена предложения, SB – площадь улучшений, SP – площадь земельного участка. Они образуют трехмерный случайный вектор (V, SB, SP). Пусть $W = \ln(V)$, $Y = \ln(SB)$, $Z = \ln(SP)$, (тогда $v = e^W$, $SB = e^Y$, $SP = e^Z$). Для трехмерного нормального случайного вектора (W, Y, Z) вектор средних равен (μ_W, μ_Y, ..., μ_Z). Ковариационная матрица выглядит следующим образом:

Таблица 1.

Исходные данные

Площадь зданий (кв. м)	Площадь земельных участков (кв. м)	Цена за весь объект (руб.)	Отношение цены к площади улучшений (руб./кв. м)
400	2 500	20 500 000	51 250
750	5 000	18 000 000	24 000
1 081	3 378	26 000 000	24 052
1 130	6 638	27 500 000	24 336
1 320	4 167	31 500 000	23 864
1 440	10 000	160 000 000	111 111
1 790	3 462	93 000 000	51 955
1 900	13 000	85 000 000	44 737
2 125	5 623	85 000 000	40 000
2 642	5 183	75 000 000	28 388
2 700	6 800	59 000 000	21 852
1 820	2 737	32 000 000	17 582
2 250	9 252	84 000 000	37 333
2 973	5 388	90 000 000	30 272
3 513	10 000	80 000 000	22 773
3 600	5 000	95 000 000	26 389
4 000	13 558	140 000 000	35 000
4 124	12 866	91 000 000	22 066
4 167	5 000	125 000 000	29 998
4 257	6 861	128 500 000	30 186
5 292	11 143	56 000 000	10 582
5 300	16 000	220 000 000	41 509
6 011	11 319	135 000 000	22 459
6 013	20 781	90 000 000	14 968
6 060	21 790	179 000 000	29 538
6 123	2 390	152 490 000	24 904
6 479	7 337	119 000 000	18 367
6 756	4 220	90 000 000	13 321
10 000	12 000	420 000 000	42 000
10 300	17 000	312 000 000	30 291
10 672	12 194	350 000 000	32 796
10 990	30 000	480 000 000	43 676
12 000	30 000	300 000 000	25 000
13 000	55 000	200 000 000	15 385
14 428	33 000	385 000 000	26 684
15 000	37 000	840 000 000	56 000
18 924	20 600	800 000 000	42 274
22 312	40 162	338 541 000	15 173
34 082	478 000	2 500 000 000	73 353
35 000	160 000	2 400 000 000	68 571

$$CV = \begin{pmatrix} \sigma_W^2 & \rho_{WY}\sigma_W\sigma_Y & \rho_{WZ}\sigma_W\sigma_Z \\ \rho_{YW}\sigma_W\sigma_Y & \sigma_Y^2 & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{ZW}\sigma_W\sigma_Z & \rho_{ZY}\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix},$$

или в блочном виде:

$$CV = \begin{pmatrix} \sigma_W^2 & cov\left(W,\vec{Y}\right) \\ cov\left(W,\vec{Y}\right)^T & COV \end{pmatrix},$$

где $COV = \begin{pmatrix} \sigma_Y^2 & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{ZY}\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix}$; $\vec{Y} = (Y, Z)$;

$cov\left(W,\vec{Y}\right) = (\rho_{WY}\sigma_W\sigma_Y,\ \rho_{WZ}\sigma_W\sigma_Z)$;

$\sigma_W^2, \sigma_Y^2, \sigma_Z^2$ – дисперсии случайных величин W, Y, Z;

$\rho_{WY} = \rho_{YW}, \rho_{WZ} = \rho_{ZW}, \rho_{YZ} = \rho_{ZY}$ – соответствующие коэффициенты корреляции.

Условное математическое ожидание W, при условии, что $Y = y$, $Z = z$:

$$E\left(W|Y = y, Z = z\right) = \\ = \mu_W + \left(COV^{-1} \times cov\left(X,\vec{Y}\right)^T,\ (y - \mu_Y, z - \mu_Z)\right).$$

Условная дисперсия W, при условии, что $Y = y$, $Z = z$:

$$D\left(W|Y = y, Z = z\right) = \\ = \sigma_W^2 - \left(COV^{-1} \times cov\left(X,\vec{Y}\right)^T,\ cov\left(X,\vec{Y}\right)\right).$$

Пусть заданы значения площади улучшений $SB = sb$ и площади земельного участка $SP = sp$. В соответствии с введенными выше обозначениями $Y = \ln(SB)$, $Z = \ln(SP)$, $y = \ln(sb)$, $z = \ln(sp)$. Наиболее вероятное значение цены предложения V при известных значениях площади улучшений и площади земельного участка рассчитывается по формуле [31]:

$$Mode(V|SB = sb, SP = sp) = \\ \exp(\mu_W + (COV^{-1} \times cov\left(X,\vec{Y}\right)^T, (y - \mu_Y, z - \mu_Z) - \\ -\sigma_W^2 + \left(COV^{-1} \times cov\left(W,\vec{Y}\right)^T,\ cov\left(W,\vec{Y}\right)\right))). \quad (4)$$

Условное математическое ожидание рассчитывается по формуле:

$$E(V|SB = sb,\ SP = E) = \\ \mu_W + (COV^{-1} \times cov\left(X,\vec{Y}\right)^T, (y - \mu_Y, z - \mu_Z)) + \\ + \frac{1}{2}\sigma_W^2 - \frac{1}{2}\left(COV^{-1} \times cov\left(W,\vec{Y}\right)^T, cov\left(W,\vec{Y}\right)\right). \quad (5)$$

Условная медиана рассчитывается по формуле:

$$Median(V|SB = sb,\ SP = sp) = \exp(\mu_W + \left(COV^{-1} \times cov\left(W, \vec{Y}\right)^T, \left(y - \mu_Y, z - \mu_Z\right)\right)). \quad (6)$$

Прежде чем применить формулы (4)–(6) к данным *таблицы 1* проверим, есть ли основания предполагать логнормальность распределений компонент случайного вектора (*V*, *SB*, *SP*) (совместную нормальность логарифмов их компонент). Для проверки маргинальных распределений использовался параметрический тест Колмогорова–Смирнова. Получены следующие значения *p*-value:

V – цена 1 кв. м улучшений: при параметрах meanlog = 10,3 и sdlog = 0,43 *p*-value составляет 0,7016;

SB – площадь зданий: при параметрах meanlog = 8,45 и sdlog = 1,02 *p*-value равно 0,9761;

SP – площадь земельного участка: при параметрах meanlog = 9,3 и sdlog = 1,01 *p*-value составляет 0,8963.

Проверим три изучаемые случайные величины на совместную нормальность логарифмов. Для этого воспользуемся известным условием совместной нормальности: для того, чтобы многомерный случайный вектор имел многомерное нормальное распределение необходимо и достаточно, чтобы любая линейная комбинация его компонент была распределена нормально. В среде статистического пакета R была реализована следующая процедура:

✦ переменные прологарифмированы;

✦ полученные логарифмические значения переменных центрированы и нормированы, каждая на свое стандартное отклонение;

✦ с помощью стандартной функции R runif(3,0,1) сгенерированы три весовых коэффициента, коэффициенты нормированы на их сумму;

✦ из центрированных и нормированных логарифмов составлена линейная комбинация со случайными положительными коэффициентами, в сумме равными единице;

✦ полученная линейная комбинация тестируется при помощи теста Колмогорова–Смирнова на нормальность, результат теста в виде значения *p*-value записывается в массив;

✦ процедура со случайной линейной комбинацией повторяется заданное количество раз, каждый раз записывается *p*-value. Затем общий массив *p*-value сравнивается с критическим уровнем (0,05).

На приведенном *рисунке 1* представлена гистограмма, показывающая, какие значения *p*-value были получены при повторении теста 100 000 раз.

Минимальное значение *p*-value равно 0,2867691, что выше критического уровня 0,05. Приведенная процедура 100 000-кратного повторения тестов для случайных линейных комбинаций компонент вектора (*W*, *Y*, *Z*) представляется достаточным основанием для того, чтобы сохранить в качестве рабочей гипотезу о совместной нормальности логарифмов изучаемых переменных.

Для логарифмов переменных «Отношение цены к площади улучшений», «Площадь зданий», «Площадь земельного участка», указанных в *таблице 1*, получены следующие значения вектора средних и ковариационная матрица (*таблица 2*).

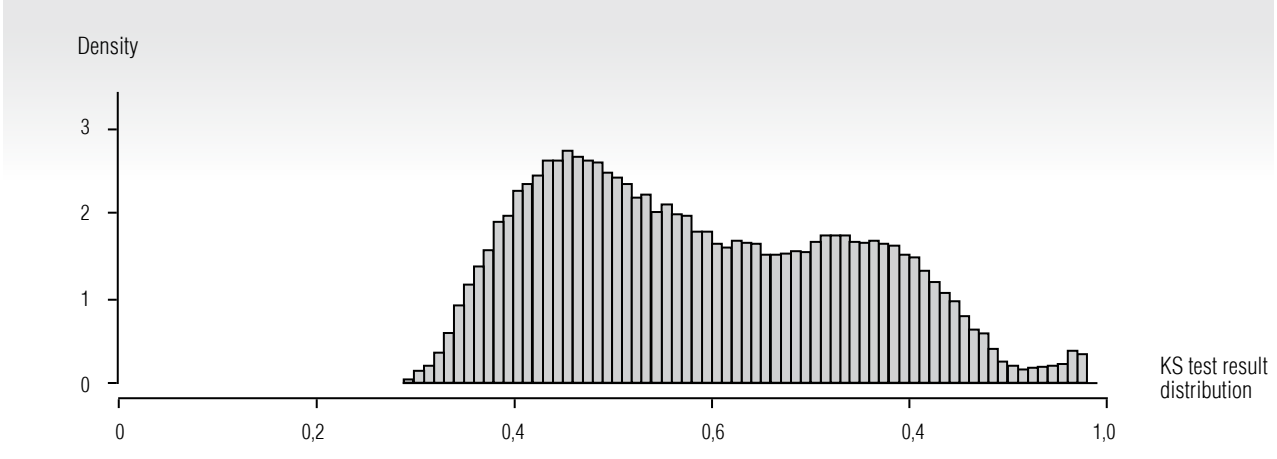


Рис. 1. Результаты тестирования случайных линейных комбинаций центрированных и нормированных компонент вектора (*W*, *Y*, *Z*) = (ln(*V*), ln(*SB*), ln(*SP*)) на совместную нормальность тестом Колмогорова–Смирнова

Таблица 2.

Средние логарифмов переменных

Отношение цены к площади улучшений (V)	Площадь зданий (SB)	Площадь земельного участка (SP)
Вектор средних		
10,2993	8,4469	9,3506
Ковариационная матрица		
0,2381	0,0108	0,1467
0,0108	1,0635	0,8978
0,1467	0,8978	1,2140

Таким образом

$\mu_W = 10{,}2993$; $\mu_Y = 8{,}4469$; $\mu_Z = 9{,}3506$, $\sigma_W^2 = 0{,}2381$,

$\rho_{WY}\sigma_W\sigma_Y = \rho_{YW}\sigma_W\sigma_Y = 0{,}0108$;

$\rho_{WZ}\sigma_W\sigma_Z = \rho_{ZW}\sigma_W\sigma_Z = 0{,}1467$; $\sigma_Y^2 = 1{,}0635$,

$\rho_{YZ}\sigma_Y\sigma_Z = \rho_{ZY}\sigma_Y\sigma_Z = 0{,}8978$; $\sigma_Z^2 = 1{,}2140$.

В среде статистического пакета R реализован программный код, позволяющий рассчитать оценку рыночной стоимости по заданным значениям параметров «Площадь зданий», «Площадь земельного участка» (формула (4)). Аналогичные расчеты могут быть выполнены для оценок по медианным значениям или по математическим ожиданиям (формулы (5) и (6)). Результаты представлены в *таблице 3*.

Из представленной таблицы видно следующее:

✦ оценка по моде всегда ниже оценки по медиане, а оценка по медиане всегда ниже оценки по математическому ожиданию (мнение автора: рыночную стоимость следовало бы определять как оценку по моде, в соответствии с формулировкой ФЗ-135, учитывая наблюдаемые на рынке несимметричные распределения цен, площадей и расстояний);

✦ при постоянной площади земельного участка рыночная стоимость в пересчете на 1 кв. м улучшений уменьшается при увеличении площади улучшений;

✦ при постоянной площади улучшений рыночная стоимость в пересчете на 1 кв. м улучшений увеличивается при увеличении площади земельного участка;

✦ по формуле (4) может быть рассчитана рыночная стоимость объекта недвижимости того же класса, что и объекты сравнения, для любых значений площадей улучшений и земельного участка (на ту же дату, для той же локации). Поскольку в оценочной среде нет общего единства мнений относительно используемых для оценки рыночной стоимости числовых характеристик (мода, медиана, математическое ожидание), могут быть применены формулы (5) и (6), строго говоря, не отвечающие определению рыночной стоимости в соответствии с ФЗ-135.

2. Соотношение площади улучшений и площади земельного участка при известной цене предложения

В статье [30] рассматривается, в том числе, вопрос о ценовых тенденциях на рынке недвижимости производственно-складского назначения и зависимость рыночной стоимости от «коэффициента плотности застройки» земельного участка, под которым понимается отношение площади улучшений к площади всего земельного участка. Рассматриваемая в настоящей статье модель совместного логарифмически нормального распределения компонент случайного вектора (V, SB, SP) также позволяет посмотреть на проблему формирования ценовых тенденций. Отличие заключается в том, что все компоненты случайного вектора (V, SB, SP) распределены на положительной полуоси, т.е. для каждого заданного значения $V = v$ можно указать наиболее вероятные значения компонент SB (площади улучшений) и SP (площади земельного участка), соответствующие цене предложения. В отличие от предыдущего случая (оценка рыночной стоимости по заданным значениям SB и SP), область возможных отклонений от наиболее вероятного (медианного, среднего) значений находится не на числовой оси, а на плоскости и представляет из себя (как будет показано ниже) вложенные множества, получающиеся из эллипсов рассеяния логарифмированных значений SB и SP при обратном экспоненциальном преобразовании плоскости.

Пусть цена предложения $V = v$ известна. Требуется оценить соотношение площади улучшений и земельного участка для класса объектов с такой начальной ценой предложения, т.е. выделить объекты с нижней, средней и верхней ценовыми тенденциями [30]. Обозначения прежние: V – цена предложения, SB – площадь улучшений, SP – площадь земельного участка, $W = \ln(V)$, $Y = \ln(SB)$, $Z = \ln(SP)$ (т.е. $V = e^W$, $SB = e^Y$, $SP = e^Z$).

По прежнему, рассматриваем трехмерный нормальный случайный вектор (W, Y, Z) с вектором средних (μ_W, μ_Y, μ_Z) и ковариационной матрицей

Таблица 3.

Оценки рыночной стоимости, приведенной к 1 кв. м улучшений для различных значений площадей улучшений и земельных участков

Оценка РС по моде		Площадь земельного участка в метрах									
		2 000	7 000	12 000	17 000	22 000	27 000	32 000	37 000	42 000	47 000
Площадь улучшений в метрах	400	26 247	38 298	45 058	50 049	54 096	57 543	60 568	63 279	65 745	68 014
	2 400	16 938	24 714	29 076	32 297	34 909	37 133	39 085	40 835	42 426	43 890
	4 400	14 605	21 310	25 072	27 849	30 101	32 019	33 702	35 211	36 583	37 845
	6 400	13 327	19 445	22 877	25 411	27 466	29 216	30 752	32 129	33 381	34 533
	8 400	12 469	18 194	21 406	23 777	25 700	27 337	28 775	30 062	31 234	32 312
	10 400	11 835	17 269	20 317	22 567	24 392	25 947	27 311	28 533	29 645	30 668
	12 400	11 337	16 542	19 462	21 618	23 366	24 855	26 161	27 332	28 397	29 377
	14 400	10 930	15 948	18 763	20 842	22 527	23 962	25 222	26 351	27 378	28 323
	16 400	10 588	15 449	18 176	20 189	21 822	23 212	24 433	25 527	26 521	27 436
	18 400	10 294	15 021	17 672	19 629	21 217	22 569	23 755	24 818	25 786	26 675

Оценка РС по медиане		Площадь земельного участка в метрах									
		2 000	7 000	12 000	17 000	22 000	27 000	32 000	37 000	42 000	47 000
Площадь улучшений в метрах	400	31 947	46 615	54 843	60 918	65 844	70 039	73 722	77 021	80 023	82 784
	2 400	20 616	30 081	35 391	39 311	42 490	45 197	47 573	49 703	51 640	53 421
	4 400	17 777	25 938	30 517	33 897	36 638	38 972	41 021	42 858	44 528	46 064
	6 400	16 221	23 668	27 846	30 930	33 431	35 561	37 431	39 106	40 630	42 032
	8 400	15 177	22 146	26 055	28 941	31 281	33 274	35 023	36 591	38 017	39 329
	10 400	14 405	21 019	24 729	27 468	29 690	31 581	33 242	34 730	36 083	37 328
	12 400	13 799	20 134	23 688	26 312	28 440	30 252	31 843	33 268	34 564	35 757
	14 400	13 304	19 412	22 838	25 368	27 419	29 166	30 700	32 074	33 324	34 473
	16 400	12 887	18 804	22 123	24 574	26 561	28 253	29 739	31 070	32 281	33 395
	18 400	12 530	18 283	21 510	23 892	25 824	27 470	28 914	30 208	31 385	32 468

Оценка РС по математическому ожиданию		Площадь земельного участка в метрах									
		2 000	7 000	12 000	17 000	22 000	27 000	32 000	37 000	42 000	47 000
Площадь улучшений в метрах	400	35 246	51 428	60 506	67 208	72 643	77 271	81 334	84 974	88 285	91 332
	2 400	22 744	33 187	39 045	43 370	46 877	49 864	52 485	54 835	56 971	58 937
	4 400	19 612	28 616	33 668	37 397	40 421	42 996	45 257	47 283	49 125	50 820
	6 400	17 895	26 112	30 721	34 124	36 883	39 233	41 296	43 144	44 825	46 372
	8 400	16 744	24 432	28 745	31 929	34 511	36 710	38 640	40 369	41 942	43 389
	10 400	15 893	23 189	27 283	30 305	32 755	34 842	36 674	38 315	39 809	41 182
	12 400	15 224	22 213	26 134	29 029	31 377	33 376	35 130	36 703	38 133	39 449
	14 400	14 677	21 416	25 196	27 987	30 250	32 178	33 869	35 385	36 764	38 033
	16 400	14 218	20 746	24 408	27 111	29 304	31 171	32 810	34 278	35 614	36 843
	18 400	13 824	20 170	23 731	26 359	28 491	30 306	31 899	33 327	34 626	35 821

$$CV = \begin{pmatrix} \sigma_W^2 & \rho_{WY}\sigma_W\sigma_Y & \rho_{WZ}\sigma_W\sigma_Z \\ \rho_{YW}\sigma_W\sigma_E & \sigma_Y^2 & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{ZW}\sigma_W\sigma_Z & \rho_{ZY}\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix},$$

или в блочном виде:

$$CV = \begin{pmatrix} \sigma_W^2 & cov\left(W,\vec{Y}\right) \\ cov\left(W,\vec{Y}\right)^T & COV \end{pmatrix},$$

где $COV = \begin{pmatrix} \sigma_Y^2 & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{ZY}\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix}$; $\vec{Y} = (Y,Z)$;

$cov\left(W,\vec{Y}\right) = \left(\rho_{WY}\sigma_W\sigma_Y,\ \rho_{WZ}\sigma_W\sigma_Z\right)$;

$\sigma_W^2, \sigma_Y^2, \sigma_Z^2$ – дисперсии случайных величин *W*, *Y*, *Z*;

$\rho_{WY} = \rho_{YW}, \rho_{WZ} = \rho_{ZW}, \rho_{YZ} = \rho_{ZY}$ – соответствующие коэффициенты корреляции.

Условное математическое ожидание вектора $\vec{Y} = (Y,Z)$, при условии, что $W = w$:

$$E\left(\vec{Y}|W=w\right) = \bar{\mu} + \frac{cov\left(W,\vec{Y}\right)^T}{\sigma_W^2}\left(w-\mu_E\right) = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix} + \begin{pmatrix} \rho_{YW}\dfrac{\sigma_Y}{\sigma_W}\left(w-\mu_W\right) \\ \rho_{ZW}\dfrac{\sigma_Z}{\sigma_W}\left(w-\mu_W\right) \end{pmatrix} = \begin{pmatrix} \mu_Y + \rho_{YW}\dfrac{\sigma_E}{\sigma_W}\left(w-\mu_W\right) \\ \mu_Z + \rho_{ZW}\dfrac{\sigma_Z}{\sigma_W}\left(w-\mu_W\right) \end{pmatrix}. \quad (7)$$

Условная ковариационная матрица при условии, что $W = w$:

$$COV\left(\vec{Y}|W=w\right) = COV - \frac{cov\left(W\vec{Y}\right)^T \times cov\left(W,\vec{Y}\right)}{\sigma_W^2} = \begin{pmatrix} \sigma_Y^2 & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{ZY}\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix} - \begin{pmatrix} \rho_{YW}^2\sigma_Y^2 & \rho_{YW}\rho_{ZW}\sigma_Y\sigma_Z \\ \rho_{YW}\rho_{ZW}\sigma_Y\sigma_Z & \rho_{ZW}^2\sigma_Z^2 \end{pmatrix} = \begin{pmatrix} \sigma_Y^2\left(1-\rho_{YW}^2\right) & \sigma_Y\sigma_Z\left(\rho_{YZ}-\rho_{YW}\rho_{ZW}\right) \\ \sigma_Y\sigma_Z\left(\rho_{YZ}-\rho_{YW}\rho_{ZW}\right) & \sigma_Z^2\left(1-\rho_{ZW}^2\right) \end{pmatrix}. \quad (8)$$

Пусть известна цена предложения $V = v$. В соответствии с введенными выше обозначениями $W = \ln(V)$, $w = \ln(v)$. Наиболее вероятная комбинация площади улучшений *SB* и площади земельного участка *SP* при фиксированной цене предложения рассчитывается по формуле:

$$Mode(\vec{Y}|V=v) = \exp\left(\bar{\mu} + \frac{cov\left(W,\vec{Y}\right)^T}{\sigma_W^2}\left(w-\mu_W\right) - COV + \frac{cov\left(W\vec{Y}\right)^T \times cov\left(W,\vec{Y}\right)}{\sigma_W^2}\right).$$

В *таблице 4* представлены расчетные наиболее вероятные комбинации площади улучшений *SB* и площади земельного участка *SP* для некоторых значений цены предложения.

Следует отметить, что при каждом значении цены предложения такая наиболее вероятная пара значе-

Таблица 4.

Наиболее вероятные комбинации площади улучшений *SB* и площади земельного участка *SP*, значения коэффициента плотности застройки земельного участка для некоторых значений цены предложения

Цена предложения в руб. за 1 кв.м. приведенной площади	7 000	12 000	21 000	28 000	40 000	60 000	80 000	100 000
Наиболее вероятная пара:								
площадь улучшений в кв.м.	619	634	650	659	669	682	691	698
площадь земельного участка в кв.м.	630	878	239	1 479	1 843	2 365	2 824	3 240
коэффициент плотности застройки	0,98	0,72	0,52	0,45	0,36	0,29	0,24	0,22

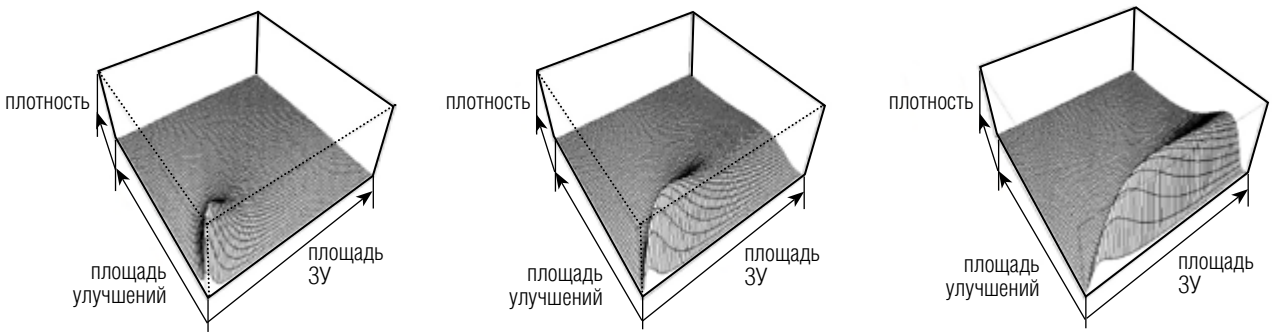


Рис. 2. Двумерные распределения *SP* (площадь земельного участка) и *SB* (площадь улучшений)

ний *SB*, *SP* является единственной (коэффициент плотности застройки в этом случае отвечает наиболее вероятной паре *SB*, *SP*). Попытки представить как наиболее удобную, с какой-либо точки зрения, другую пару компонент *SB* и *SP* означают выбор точки, для которой существует множество других равновероятных точек, с плотностью меньше максимальной, и для которых коэффициенты плотности застройки, очевидно, будут разными. На *рисунке 2* представлены изображения двумерных распределений *SB* и *SP* для цены предложения 7 000 руб., 28 000 руб., 100 000 руб. (слева направо) за 1 кв. м . площади, приведенной к площади улучшений.

Хорошо видно, что любая другая точка плоскости *SP*, *SB* имеет множество равновероятных значений. Области таких значений при цене предложения 28 000 руб. за 1 кв. м улучшений единого комплекса (земельный участок плюс улучшения) представлены на *рисунке 3.*

3. Коэффициент плотности застройки (отношение площади улучшений к площади земельного участка)

Предположим, что цена предложения (сделки) известна. Требуется оценить, каким должен быть коэффициент плотности застройки при данной цене и известной площади земельного участка. Обратимся к формулам (7) и (8). При фиксированной цене предложения ($V = v$) формулы (7) и (8) дают расчетные значения условных математических ожиданий логарифмов площади улучшений ($SB|V = v$), площади земельного участка ($SP|V = v$) и условной ковариационной матрицы. Теперь предположим, что площадь земельного участка тоже известна. Введем новые обозначения для условных логарифмов площади улучшений ($SB|V = v$) и площади земельного участка ($SP|V = v$):

$$\mu_{condSB} = \mu_Y + \rho_{YW}\frac{\sigma_Y}{\sigma_W}\left(w - \mu_W\right),$$

$$\mu_{condSP} = \mu_Z + \rho_{ZW}\frac{\sigma_Z}{\sigma_W}\left(w - \mu_W\right),$$

$$\sigma^2_{condSB} = \sigma^2_Y\left(1 - \rho^2_{YW}\right),$$

$$\sigma^2_{condSP} = \sigma^2_Z\left(1 - \rho^2_{ZW}\right),$$

$$\rho = \sigma_Y\sigma_Z\left(\rho_{YZ} - \rho_{YW}\rho_{ZW}\right)$$

(части нижних индексов “cond” означают «условное»). Рассмотрим двумерный случайный вектор ($SB|V = v$, $SP|V = v$) с указанными параметрами. При заданном значении площади земельного участка и заданном значении цены (в приведенном примере цены предложения) ($SP = sp$, $V = v$), условная мода величины *SB* (площади улучшений) равна (по аналогии с доказательством, приведенным в работе [32]):

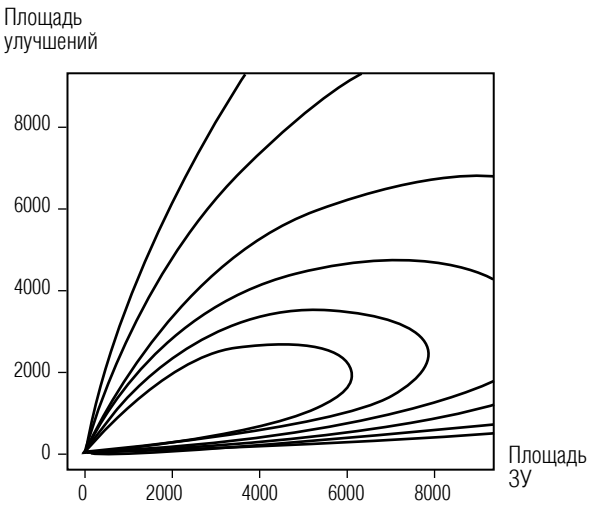


Рис. 3. Линии уровней равновероятных точек на плоскости с координатами *SP* (площадь земельного участка), *SB* (площадь улучшений)

$$Mode(SP|\mathrm{SP}=\mathrm{sp},\mathrm{V}=\mathrm{v}) = \exp(\mu_{condSP} + \\ +\rho\times\frac{\sigma^2_{condSP}}{\sigma^2_{condSB}}\left(\ln\left(sb-\mu_{condSB}\right)\right)-\sigma^2_{condSP}\left(1-\rho^2\right)). \quad (9)$$

Условная медиана величины *SP* равна:

$$Median(SP\,|\,\mathrm{SP}=\mathrm{sp},\mathrm{V}=\mathrm{v}) = \\ =\exp(\mu_{condSP}+\rho\times\frac{\sigma^2_{condSP}}{\sigma^2_{condSB}}\left(\ln\left(sb-\mu_{condSB}\right)\right).$$

Условное математическое ожидание величины *SP* равно:

$$E(SP\,|\,\mathrm{SP}=\mathrm{sp},\mathrm{V}=\mathrm{v}) = \exp(\mu_{condSP} + \\ +\rho\times\frac{\sigma^2_{condSP}}{\sigma^2_{condSB}}\left(\ln\left(sb-\mu_{condSB}\right)\right)+\frac{1}{2}\sigma^2_{condSP}\left(1-\rho^2\right)).$$

Предположим, что требуется оценить коэффициент застройки в группе объектов нижней, средней или верхней ценовой категории. Такие оценки могут быть построены в зависимости от площади земельного участка как по модальным, так и по медианным или средним значениям. Однако общий вид поверхностей, показанных на *рисунке 2*, говорит о том, что наиболее консервативными будут оценки по модальным значениям. Оценки по медианным или средним значениям выглядят завышенными (при $V = 28\,000$ руб./1 кв. м – примерно в 1,4 и 1,7 раза, *рисунок 4*). Предположим, что нас интересует следующий вопрос: если цена предложения 28 000 руб. за 1 кв. м приведенной площади, а площадь земельного участка равна 30 000 кв. м, то какую площадь улучшений (и, соответственно, какой коэффициент плотности застройки) следовало бы считать адекватным такой цене и площади земельного участка. Под коэффициентом застройки будем понимать отношение оценочного значения площади улучшений к площади земельного участка, т.е.

$$\frac{Mode\left(SB|SP=sp\right)}{sp}$$

(как вариант, $\frac{Median\left(SB|SP=sp\right)}{sp}$, $\frac{E\left(SB|SP=sp\right)}{sp}$).

На *рисунке 4* показано, что оценки по модальным, медианным и средним значениям могут значительно отличаться. Применение формулы (9) к результатам расчета условных параметров при цене 28 000 руб./кв. м и значению площади земельного участка равного 30 000 кв. м дает результат 7 165 кв. м улучшений, и тогда коэффициент плотности застройки составляет 7 165 / 30 000 = 0,24. Таким образом, исходя из данных примера (*таблица 1*), другой коэффициент площади застройки при цене 28 000 руб./кв. м, площади земельного участка 30 000 кв. м может пониматься как не отвечающий выставленной цене. Этот же результат можно было получить применением формулы, аналогичной формуле (1). В этом разделе последовательный учет условий (сначала цены $V = v$, затем площади земельного участка $SP = sp$) применен для того, чтобы показать, что коэффициент плотности застройки не является константой внутри одной ценовой группы и даже для одной конкретной цены, и имеет зависимость степенного характера от площади земельного участка. В левой части *рисунка 4* показаны линии модальных, медианных и средних значений площади улучшений, в зависимости от площади

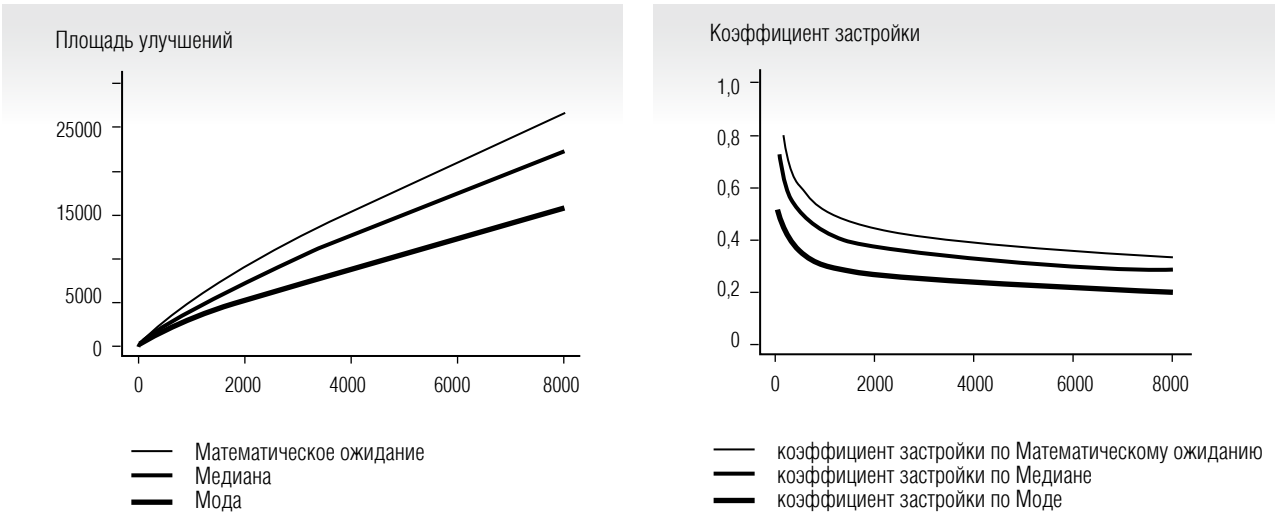


Рис. 4. Значения площади улучшений и коэффициентов застройки в зависимости от площади земельного участка

земельного участка для случая, когда цена предложения равна 28 000 руб./кв. м площади существующих улучшений, в правой части – линии коэффициентов застройки для соответствующих оценок площади улучшений. Из *рисунка 4* видно, что при заданной цене (ценовой группе) коэффициент застройки с приемлемой для целей оценки точностью может быть оценен константой только при достаточно большой площади земельного участка. Для участков с малой площадью коэффициент застройки не может быть оценен константой и должен изучаться индивидуально с учетом площади участка.

4. Замечание об общем виде совместного логарифмически нормального распределения вектора (*V*, *SB*, *SP*)

Многомерное распределение компонент вектора (*W*, *Y*, *Z*) = (ln(*V*), ln(*SB*), ln(*SP*)) совместно нормально и обладает симметрией. Облака рассеяния эмпирических наблюдений будут приобретать вид трехмерных эллипсоидов. Точка максимума плотности имеет координаты, равные средним значениям компонент *W*, *Y*, *Z*. Распределение компонент вектора (*V*, *SB*, *SP*) ассиметрично, точка максимума плотности не является центром симметрии и может быть рассчитана (см. Приложение) по следующим формулам:

$$V_{max} = \exp(\mu_W - \sigma_W^2 - \rho_{WY}\sigma_W\sigma_Y - \rho_{WZ}\sigma_W\sigma_Z),$$
$$SB_{max} = \exp(\mu_Y - \sigma_Y^2 - \rho_{YW}\sigma_W\sigma_Y - \rho_{YZ}\sigma_Y\sigma_Z),$$
$$SP_{max} = \exp(\mu_Z - \sigma_Z^2 - \rho_{ZY}\sigma_Z\sigma_Y - \rho_{ZW}\sigma_W\sigma_Z).$$

На *рисунке 5* показаны: рассеяние исходных данных и рассеяние логарифмов исходных данных, точка максимальной плотности в пространстве (*V*, *SB*, *SP*) с координатами V_{max} = 20 004 руб./кв.м, SB_{max} = 649 кв.м, SP_{max} = 1 202 кв.м, а также точка максимальной плотности в логарифмированном пространстве (*W*, *Y*, *Z*) = (ln(*V*), ln(*SB*), ln(*SP*)) с координатами μ_W = 10,30; μ_Y = 8,45; μ_Z = 9,35. На рисунке черным цветом отмечены точки максимальной плотности: слева – в пространстве (*V*, *SB*, *SP*), справа – в пространстве (*W*, *Y*, *Z*) = (ln(*V*), ln(*SB*), ln(*SP*)).

Если сгенерировать 1000 трехмерных случайных векторов с теми же параметрами, то полученное рассеяние будет иметь вид, показанный на *рисунке 6.* Слева на рисунке показано рассеяние для 1000 сгенерированных случайных векторов, компоненты которых распределены совместно логарифмически нормально, с теми же параметрами, что и параметры исходной выборки (по которым проводилось тестирование на совместную нормальность логарифмов). Справа показано рассеяние логарифмов компонент 1000 сгенерированных случайных векторов. Черным отмечены точки максимальной плотности: слева – в пространстве (*V*, *SB*, *SP*), справа – в пространстве (*W*, *Y*, *Z*) = (ln(*V*), ln(*SB*), ln(*SP*)).

Очевидно, что (см. Приложение) точка максимума плотности многомерного вектора (мода), логарифмы компонент которого распределены совместно нормально – единственна. Всем остальным значениям плотности отвечают множества, описываемые в логарифмическом измерении полыми трехмерными эллипсоидами, а в исходных координатах множества, отвечающие одному значению плотности, представляют собой результат

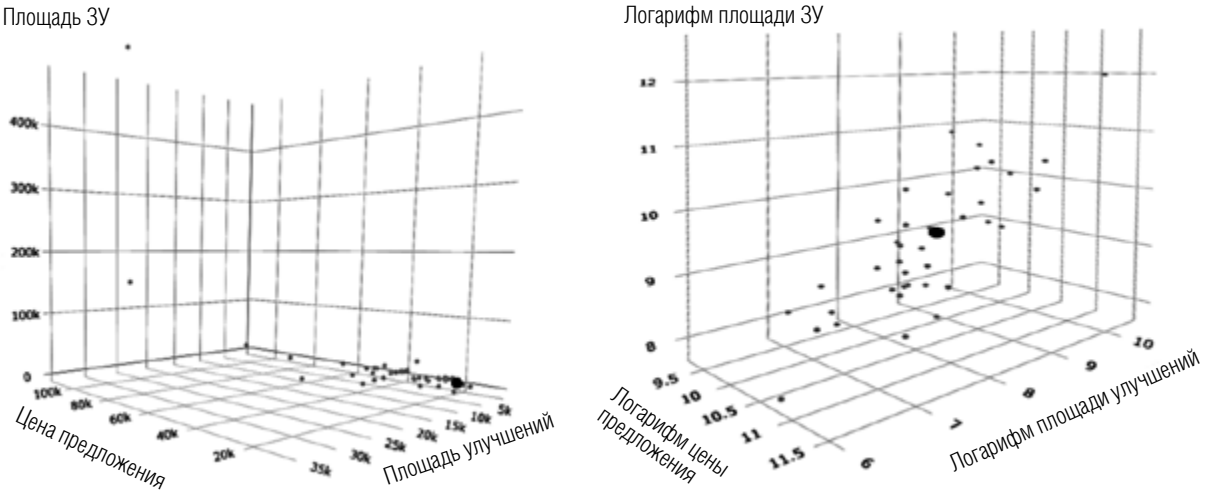


Рис. 5. Рассеяние исходных данных (слева), рассеяние логарифмов исходных данных (справа)

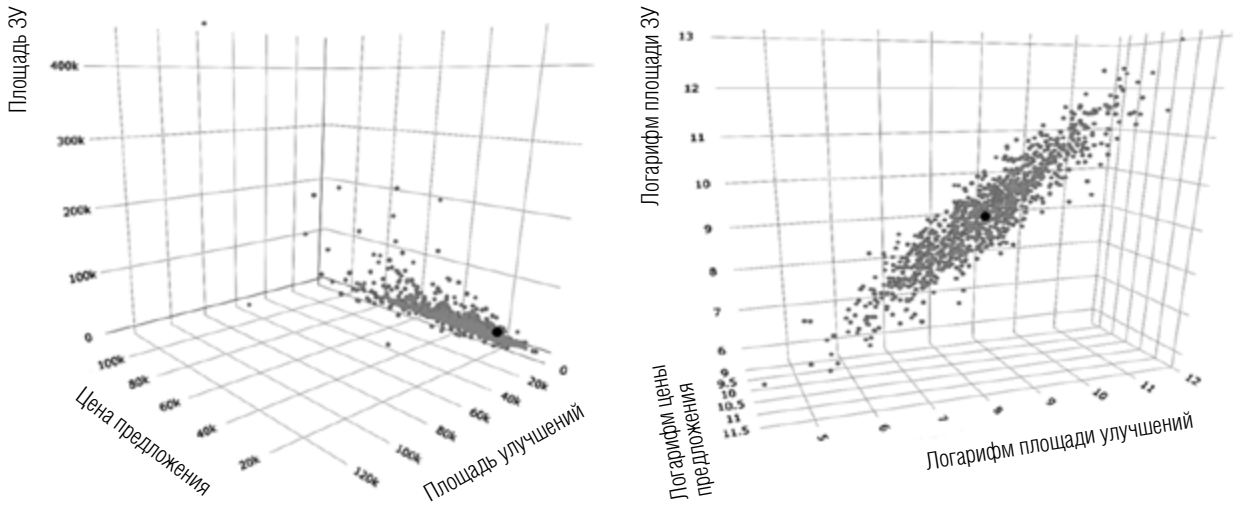


Рис. 6. Результат для 1000 сгенерированных случайных векторов

искажения (растяжения) полых эллипсоидов при обратном экспоненциальном преобразовании пространства. Таким образом, именно модальная оценка рыночной стоимости должна приводить к корректному, не создающему конфликтных ситуаций результату. Все остальные (не модальные) оценки рыночной стоимости потенциально являются источником постоянных споров относительно рыночной стоимости объекта оценки.

Заключение

Рассмотрение цен объектов сравнения и значений ценообразующих факторов как многомерных случайных величин открывает новые возможности в оценке недвижимого имущества. Часто оказывается, что эмпирические наблюдения цен и соответствующих им значений ценообразующих факторов хорошо приближаются логарифмически нормальным законом распределения, в том числе многомерным, что позволяет вывести расчетные формулы для различных задач оценки. Громоздкость этих формул компенсируется возможностями современных прикладных статистических пакетов (в частности, R). Кроме того возможность сведения расчетов к хорошо изученному многомерному нормальному закону логарифмированием компонент делает такой выбор модельного распределения предпочтительным.

Условные распределения цен при известных значениях ценообразующих факторов дают возможность оценить рыночную стоимость в полном соответствии с ее определением, закрепленным в российском законодательстве и зарубежных стандартах, как точку максимума плотности условного распределения цен.

Условные распределения ценообразующих факторов при заданной цене предложения позволяют оценить адекватность цены предложения с точки зрения набора ценообразующих факторов.

Вряд ли следует ожидать, что практикующие оценщики готовы применять приведенные в настоящей статье формулы в ежедневной практике оценки и бизнес-анализа. Этого и не требуется. Однажды написанный и отлаженный скрипт (в статистическом пакете R или в других специализированных пакетах) позволит с легкостью решать подобные задачи практически в режиме реального времени. Следует признать, что в период цифровой трансформации экономики и бизнес-анализа оценочному бизнесу пора переходить к продвинутым статистическим пакетам и автоматической обработке данных. ■

Приложение

Утверждение. Абсолютный максимум (мода) плотности случайного логарифмически нормального вектора $\vec{x}$ достигается в точке с координатами $\exp(\vec{\mu} - \Sigma \times \mathbf{1})$, где $\vec{\mu}$ – вектор математических ожиданий логарифмов компонент, Σ – ковариационная матрица логарифмов компонент, 1 – вектор, состоящий из единиц.

Доказательство. Рассмотрим плотность многомерного нормального распределения центрированного случайного вектора $\vec{y}$:

$$f(\vec{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt[2]{\det \Sigma}} \exp\left(-\frac{1}{2}\left(\Sigma^{-1}\vec{y}, \vec{y}\right)\right).$$

При замене переменных $\vec{y} = \ln(\vec{x})$ плотность логнормального распределения случайного вектора $\vec{x}$ будет:

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt[2]{\det \Sigma}} \times \frac{1}{\prod_{i=1}^{n} x_i} \times$$

$$\times \exp\left(-\frac{1}{2}\left(\Sigma^{-1}\overline{\ln(x)}, \overline{\ln(x)}\right)\right), \text{ где}$$

$\frac{1}{\prod_{i=1}^{n} x_i}$ – Якобиан преобразования координат,

Σ – ковариационная матрица, $\overline{\ln(x)}$ – центрированный случайный вектор. В точке абсолютного максимума плотности совместного логарифмически нормального распределения производная по любому направлению должна равняться нулю, что означает равенство нулю всех частных производных.

$$\frac{\partial f(\vec{x})}{\partial x_j} = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt[2]{\det \Sigma}} \times \frac{1}{\prod_{i=1}^{n} x_i}\left(-\frac{1}{x_j}\right) \times$$

$$\exp\left(-\frac{1}{2}\left(\Sigma^{-1}\overline{\ln(x)}, \overline{\ln(x)}\right)\right) +$$

$$\frac{1}{2\pi)^{\frac{n}{2}} \sqrt[2]{\det \Sigma}} \times \frac{1}{\prod_{i=1}^{n} x_i} \exp\left(-\frac{1}{2}\left(\Sigma^{-1}\overline{\ln(x)} \;\; \overline{\ln(x)}\right)\right) \times$$

$$\times\left(-\frac{2}{2}\,\Sigma^{-1}\overline{\ln(x)}\right) \times \frac{1}{x_j} = 0$$

После вынесения за скобки общих множителей остается условие:

$$-\mathbf{1} + \left(-\Sigma^{-1}\overline{\ln(x)}\right) = 0 \text{ или } \left(-\Sigma^{-1}\overline{\ln(x)}\right) = \mathbf{1},$$

где $\mathbf{1}$, $-\mathbf{1}$ – векторы размерности n, состоящие из единиц/минус единиц).

Умножим последнее равенство слева на Σ:

$$\Sigma\Sigma^{-1}\overline{\ln(x)} = -\Sigma \times \mathbf{1},$$

$$\boldsymbol{E} \times \overline{\ln(x)} = -\Sigma \times \mathbf{1}.$$

Здесь $\boldsymbol{E}$ – единичная матрица (на главной диагонали – единицы, остальные элементы равны нулю), $\mathbf{1}$ – вектор, состоящий из единиц. Т.е. значения вектора $\overline{\ln(x)}$, в котором все частные производные равны нулю, равны построчным суммам ковариационной матрицы, взятым с обратным знаком.

Остается вспомнить, что $\vec{y} = \ln(\vec{x})$ – центрированный случайный вектор. Если вектор математических ожиданий $\vec{\mu}$ содержит ненулевые значения, то окончательное решение:

$$\overline{\ln(x)} = \vec{\mu} - \Sigma \times \mathbf{1} \text{ или } \vec{x} = \exp(\vec{\mu} - \Sigma \times \mathbf{1}).$$

С учетом отрицательной определенности квадратичной формы, составленной из вторых частных производных в точке $\vec{x} = \exp(\vec{\mu} - \Sigma \times \mathbf{1})$ (автор опускает эту громоздкую запись, ее результат очевиден из геометрических соображений), точка $\vec{x} = \exp(\vec{\mu} - \Sigma \times \mathbf{1})$ является точкой максимума плотности случайного логарифмически нормального вектора $\vec{x}$.

Утверждение доказано.

Литература

1. Rusakov O.V., Laskin M.B., Jaksumbaeva O.I. Pricing in the real estate market as a stochastic limit. Log Normal approximation // International Journal of Mathematical Models and Methods in Applied Sciences. 2016. No 10. P. 229–236.
2. Aitchinson J., Brown J.A.C. The Lognormal distribution with special references to its uses in economics. Cambridge: University Press, 1963.
3. Ohnishi T., Mizuno T., Shimizu C., Watanabe T. On the evolution of the house price distribution // Columbia Business School. Center of Japanese Economy and Business. Working Paper Series. 2011. No 296.
4. Anselin L., Lozano-Gracia N. Errors in variables and spatial effects in hedonic house price models of ambient air quality // Empirical Economics. 2008. Vol. 34. No 1. P. 5–34. DOI: 10.1007/s00181-007-0152-3.
5. Benson E.D., Hansen J.L., Schwartz Jr. A.L., Smersh G.T. Pricing residential amenities: The value of a view // Journal of Real Estate Finance and Economics. 1998. Vol. 16. No 1. P. 55–73. DOI: 10.1023/A:1007785315925.
6. Debrezion G., Pels E., Rietveld P. The impact of rail transport on real estate prices: an empirical analysis of the Dutch housing market // Urban Studies. 2011. Vol. 48. No 5. P. 997–1015. DOI: 10.1177/0042098010371395.
7. Jim C.Y., Chen W.Y. Impacts of urban environmental elements on residential housing prices in Guangzhou (China) // Landscape and Urban Planning. 2006. Vol. 78. No 4. P. 422–434. DOI: 10.1016/j.landurbplan.2005.12.003.
8. Rivas R., Patil D., Hristidis V., Barr J.R., Srinivasan N. The impact of colleges and hospitals to local real estate markets // Journal of Big Data. 2019. Vol. 6. No 1. Article No 7 (2019). DOI: 10.1186/s40537-019-0174-7.
9. Wena H., Zhanga Y., Zhang L. Assessing amenity effects of urban landscapes on housing price in Hangzhou, China // Urban Forestry & Urban Greening. 2015. No 14. P. 1017–1026. DOI: 10.1016/j.ufug.2015.09.013.
10. Отчет об определении кадастровой стоимости объектов недвижимости на территории Санкт-Петербурга. 2018. №1. [Электронный ресурс]: http://www.ko.spb.ru/interim-reports/ (дата обращения: 05.06.2019).

11. Peterson S., Flanagan A.B. Neural network hedonic pricing models in mass real estate appraisal // Journal of Real Estate Research. 2009. Vol. 31. No 2. P. 147–164.

12. Rafiei M.H., Adeli H. Novel machine-learning model for estimating construction costs considering economic variables and indexes // Journal of Construction Engineering and Management. 2018. Vol. 144. No 12. Article No 04018106. DOI: 10.1061/(asce)co.1943-7862.0001570.

13. Antipov E.A., Pokryshevskaya E.B. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics // Expert Systems with Applications. 2012. No 39. P. 1772–1778. DOI: 10.1016/j.eswa.2011.08.077.

14. Kontrimas V., Verikas A. The mass appraisal of the real estate by computational intelligence // Applied Soft Computing. 2011. No 11. P. 443–448. DOI: 10.1016/j.asoc.2009.12.003.

15. Park B., Baem J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data // Expert Systems with Applications. 2015. No 42. P. 2928–2934. DOI: 10.1016/j.eswa.2014.11.040.

16. Case K.E., Shiller R.J. Prices of single-family Homes since 1970: New indexes for four cities // New England Economic Review. 1987. September. P. 45–56. DOI: 10.3386/w2393.

17. Englund P., Quigley J.M., Redfearn C.L. The choice of methodology for computing housing price indexes: comparison of temporal aggregation and sample definition // Journal of Real Estate Finance and Economics. 1999. Vol. 19. No 2. P. 91–112. DOI: 10.1023/A:1007846404582.

18. Epley D. Assumptions and restrictions on the use of repeat sales to estimate residential price appreciation // Journal of Real Estate Literature. 2016. Vol. 24. No 2. P. 275–286. DOI: 10.5555/0927-7544.24.2.275.

19. Malpezzi S. Hedonic pricing models: A selective and applied review // Housing economics and public policy: Essays in honor of Duncan Maclennan (T. O'Sullivan, K. Gibb, eds.). Oxford, UK: Blackwell Science, 2002. P. 67–89. DOI: 10.1002/9780470690680.ch5.

20. Case B., Quigley J.M. The dynamics of real estate prices // Review of Economics and Statistics. 1991. Vol. 73. No 1. P. 50–58.

21. Englund P., Quigley J.M., Redfearn C.L. Improved price indexes for real estate: Measuring the course of Swedish housing prices // Journal of Urban Economics. 1998. Vol. 44. No 2. P. 171–196.

22. Jones C. House price measurement: The hybrid hedonic repeat-sales method // Economic Record. 2010. Vol. 86. No 272. P. 95–97. DOI: 10.1111/j.1475-4932.2009.00596.x.

23. Wang F., Zheng X. The comparison of the hedonic, repeat sales, and hybrid models: Evidence from the Chinese paintings // Cogent Economics & Finance. 2018. No 6. P. 1–19. DOI: 10.1080/23322039.2018.1443372.

24. Brunnermeier M.K. Bubbles // The new Palgrave dictionary of Economics (L.E. Blume, S.N. Durlauf, eds.). New York: Palgrave Macmillan, 2009.

25 Fabozzi F.J., Xiao K. The timeline estimation of bubbles: The case of real estate // Real Estate Economics. 2019. Vol. 47. No 2. P. 564–594. DOI: 10.1111/1540-6229.12246.

26. Fernandez-Kranz D., Hon M.T. A cross-section analysis of the income elasticity of housing demand in Spain: Is there a real estate bubble? // Journal of Real Estate Finance and Economics. 2006. Vol. 32. No 4. P. 449–470. DOI: 10.1007/s11146-006-6962-9.

27. Phillips P.C.B., Shi S. P., Yu J. Testing for multiple bubbles: Historical episodes of exuberance // International Economic Review. 2015. Vol. 56. No 4. P. 1043–1078. DOI: 10.1111/iere.12132.

28. Phillips P.C.B., Shi S. P., Yu J. Testing for multiple bubbles: Limit theory of real time detectors // International Economic Review. 2015. Vol. 56. No 4. P. 1079–1134. DOI: 10.1111/iere.12131.

29. Федеральный закон от 29.07.1998 № 135-ФЗ (ред. от 29.07.2017) «Об оценочной деятельности в Российской Федерации». [Электронный ресурс]: http://www.consultant.ru/document/cons_doc_LAW_19586/ (дата обращения: 14.03.2020).

30. Слуцкий А.А., Слуцкая И.А. Модифицированный метод выделения и обобщенный модифицированный метод выделения. Применение для анализа сегмента рынка, к которому относится объект оценки. 2020 [Электронный ресурс]: http://tmpo.su/sluckij-a-a-sluckaya-i-a-mmv-i-ommv-primenenie-dlya-analiza-rynka-3/ (дата обращения: 14.03.2020).

31. Ласкин М.Б. Логарифмически нормальное распределение цен и рыночная стоимость на рынке недвижимости // Известия Санкт-Петербургского государственного технологического института. 2014. № 25 (51). С. 102–106.

32. Ласкин М.Б. Корректировка рыночной стоимости по ценообразующему фактору «площадь объекта» // Имущественные отношения в Российской Федерации. 2017. № 8 (191), с. 86–99.

Об авторе

Ласкин Михаил Борисович

кандидат физико-математических наук, доцент;

старший научный сотрудник, Санкт-Петербургский института информатики и автоматизации, Российская академия наук (СПИИРАН),
199178, Россия, г. Санкт-Петербург, 14 линия, д. 39;

E-mail: laskinmb@yahoo.com

ORCID: 0000-0002-0143-4164

Multidimensional log-normal distribution in real estate appraisals

Michael B. Laskin
E-mail: laskinmb@yahoo.com

St. Petersburg Institute for Informatics and Automation, Russian Academy of Sciences
Address: 39, 14 Line, St. Petersburg 199178, Russia

Abstract

The purpose of the research was to develop a market value appraisal methodology based on a set of a joint logarithmically normal distribution of price-forming factors. Joint logarithmically normal distribution means random vector component logarithms are distributed together jointly normally. This article suggests a method for appraising the real estate market value based on the statistical hypothesis of a joint logarithmically normal distribution and conditional distribution of prices with fixed values of pricing factors. The article suggests a method of offer price analysis from the point of view of its relevance to pricing factor values. We consider the features of the coefficient of development depending on the area of the land plot. Additional arguments are given in favor of estimating market value as a mode of conditional laws of price distribution. An example of a multidimensional log-normal distribution of prices and pricing factors such as the area of the improvements (improvements mean buildings and constructions) area and the land area in real data, i.e. for the case of a three-dimensional random vector. We present a formula for determining the absolute maximum density point of a multidimensional logarithmically normal random vector. The proof is given in the Appendix. The results obtained can be used to create information systems to support decision-making in valuation activities for real estate properties.

Key words: market value; logarithmically normal law of price distribution; multidimensional logarithmically normal distribution, valuation of real estate.

Citation: Laskin M.B. (2020) Multidimensional log-normal distribution in real estate appraisals. *Business Informatics*, vol. 14, no 2, pp. 48–63. DOI: 10.17323/2587-814X.2020.2.48.63

References

1. Rusakov O.V., Laskin M.B., Jaksumbaeva O.I. (2016) Pricing in the real estate market as a stochastic limit. Log Normal approximation. *International Journal of Mathematical Models and Methods in Applied Sciences*, no 10, pp. 229–236.
2. Aitchinson J., Brown J.A.C. (1963) *The Lognormal distribution with special references to its uses in economics*. Cambridge: University Press.
3. Ohnishi T., Mizuno T., Shimizu C., Watanabe T. (2011) *On the evolution of the house price distribution*. Columbia Business School. Center of Japanese Economy and Business. Working Paper Series, no 296.
4. Anselin L., Lozano-Gracia N. (2008) Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, vol. 34, no 1, pp. 5–34. DOI: 10.1007/s00181-007-0152-3.
5. Benson E.D., Hansen J.L., Schwartz Jr. A.L., Smersh G.T. (1998) Pricing residential amenities: The value of a view. *Journal of Real Estate Finance and Economics*, vol. 16, no 1, pp. 55–73. DOI: 10.1023/A:1007785315925.
6. Debrezion G., Pels E., Rietveld P. (2011) The impact of rail transport on real estate prices: an empirical analysis of the Dutch housing market. *Urban Studies*, vol. 48, no 5, pp. 997–1015. DOI: 10.1177/0042098010371395.
7. Jim C.Y., Chen W.Y. (2006) Impacts of urban environmental elements on residential housing prices in Guangzhou (China). *Landscape and Urban Planning*, vol. 78, no 4, pp. 422–434. DOI: 10.1016/j.landurbplan.2005.12.003.
8. Rivas R., Patil D., Hristidis V., Barr J.R., Srinivasan N. (2019) The impact of colleges and hospitals to local real estate markets. *Journal of Big Data*, vol. 6, no 1, article no 7 (2019). DOI: 10.1186/s40537-019-0174-7.
9. Wena H., Zhanga Y., Zhang L. (2015) Assessing amenity effects of urban landscapes on housing price in Hangzhou, China. *Urban Forestry & Urban Greening*, no 14, pp. 1017–1026. DOI: 10.1016/j.ufug.2015.09.013.
10. Saint Petersburg State Budget Department "Cadastral Valuation City Department" (2018) *Report on determining the cadastral value of real estate objects on the territory of Saint Petersburg*, no 1. Available at: http://www.ko.spb.ru/interim-reports/ (accessed 05 June 2019).
11. Peterson S., Flanagan A.B. (2009) Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, vol. 31, no 2, pp. 147–164.

12. Rafiei M.H., Adeli H. (2018) Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, vol. 144, no 12, article no 04018106. DOI: 10.1061/(asce)co.1943-7862.0001570.

13. Antipov E.A., Pokryshevskaya E.B. (2012) Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, no 39, pp. 1772–1778. DOI: 10.1016/j.eswa.2011.08.077.

14. Kontrimas V., Verikas A. (2011) The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, no 11, pp. 443–448. DOI: 10.1016/j.asoc.2009.12.003.

15. Park B., Baem J.K. (2015) Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, no 42, pp. 2928–2934. DOI: 10.1016/j.eswa.2014.11.040.

16. Case K.E., Shiller R.J. (1987) Prices of single-family Homes since 1970: New indexes for four cities. *New England Economic Review*, September, pp. 45–56. DOI: 10.3386/w2393.

17. Englund P., Quigley J.M., Redfearn C.L. (1999) The choice of methodology for computing housing price indexes: comparison of temporal aggregation and sample definition. J*ournal of Real Estate Finance and Economics*, vol. 19, no 2, pp. 91–112. DOI: 10.1023/A:1007846404582.

18. Epley D. (2016) Assumptions and restrictions on the use of repeat sales to estimate residential price appreciation. *Journal of Real Estate Literature*, vol. 24, no 2, pp. 275–286. DOI: 10.5555/0927-7544.24.2.275.

19. Malpezzi S. (2002) Hedonic pricing models: A selective and applied review. *Housing economics and public policy: Essays in honor of Duncan Maclennan* (T. O'Sullivan, K. Gibb, eds.). Oxford, UK: Blackwell Science, pp. 67–89. DOI: 10.1002/9780470690680.ch5.

20. Case B., Quigley J.M. (1991) The dynamics of real estate prices. *Review of Economics and Statistics*, vol. 73, no 1, pp. 50–58.

21. Englund P., Quigley J.M., Redfearn C.L. (1998) Improved price indexes for real estate: Measuring the course of Swedish housing prices. *Journal of Urban Economics*, vol. 44, no 2, pp. 171–196.

22. Jones C. (2010) House price measurement: The hybrid hedonic repeat-sales method. *Economic Record*, vol. 86, no 272, pp. 95–97. DOI: 10.1111/j.1475-4932.2009.00596.x.

23. Wang F., Zheng X. (2018) The comparison of the hedonic, repeat sales, and hybrid models: Evidence from the Chinese paintings. *Cogent Economics & Finance*, no 6, pp. 1–19. DOI: 10.1080/23322039.2018.1443372.

24. Brunnermeier M.K. (2009) Bubbles. *The new Palgrave dictionary of Economics* (L.E. Blume, S.N. Durlauf, eds.). New York: Palgrave Macmillan.

25. Fabozzi F.J., Xiao K. (2019) The timeline estimation of bubbles: The case of real estate. *Real Estate Economics*, vol. 47, no 2, pp. 564–594. DOI: 10.1111/1540-6229.12246.

26. Fernandez-Kranz D., Hon M.T. (2006) A cross-section analysis of the income elasticity of housing demand in Spain: Is there a real estate bubble? *Journal of Real Estate Finance and Economics*, vol. 32, no 4, pp. 449–470. DOI: 10.1007/s11146-006-6962-9.

27. Phillips P.C.B., Shi S.-P., Yu J. (2015) Testing for multiple bubbles: Historical episodes of exuberance. I*nternational Economic Review*, vol. 56, no 4, pp. 1043–1078. DOI: 10.1111/iere.12132.

28. Phillips P.C.B., Shi S. P., Yu J. (2015) Testing for multiple bubbles: Limit theory of real time detectors. *International Economic Review*, vol. 56, no 4, pp. 1079–1134. DOI: 10.1111/iere.12131.

29. The Federal Law of 29.07.1998 No 135-FZ (edition of 29.07.2017) "*About assessment activity in the Russian Federation*". Available at: http://www.consultant.ru/document/cons_doc_LAW_19586/ (accessed 14 March 2020).

30. Slytsky A.A., Slytskaya I.A. (2020) *The modified extraction method and the generalized modified method of allocation. Use for analyzing the market segment that the item is being evaluated belongs to*. Available at: http://tmpo.su/sluckij-a-a-sluckaya-i-a-mmv-i-ommv-primenenie-dlya-analiza-rynka-3/ (accessed 14 March 2020).

31. Laskin M.B. (2014) Logarithmically normal distribution of prices and market value in the real estate market. *Saint Petersburg State Technological Institute Review*, no 25 (51), pp.102–106.

32. Laskin M.B. (2017) Market value adjustment for the pricing factor "square". *Property Relations in the Russian Federation*, no 8 (191), pp. 86–99.

About the author

Michael B. Laskin

Cand. Sci. (Phys.-Math.), Associate Professor;

Senior Researcher, St. Petersburg Institute for Informatics and Automation,
Russian Academy of Sciences (SPIIRAS),
39, 14 Line, St. Petersburg 199178, Russia

E-mail: laskinmb@yahoo.com

ORCID: 0000-0002-0143-4164

DOI: 10.17323/2587-814X.2020.2.64.83

Спрос на навыки на рынке труда в сфере информационных технологий

А.А. Терников
E-mail: aternikov@hse.ru

Е.А. Александрова
E-mail: ea.aleksandrova@hse.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: 194100, г. Санкт-Петербург, ул. Кантемировская, д. 3, корп. 1, лит. А

Аннотация

Сфера информационных технологий является одной из наиболее динамично развивающихся на рынке труда. Предъявляемый спрос на навыки имеет значительную вариацию в зависимости от отрасли и сферы деятельности организаций, специфики рабочего места и организации труда. Для формирования навыков, достаточных для успешного трудоустройства выпускников, со стороны системы образования необходим качественный мониторинг спроса, предъявляемый работодателями. В статье представлен алгоритм, позволяющий определить ключевые комбинации навыков, которые необходимы компаниям в сфере информационных технологий в зависимости от профессиональных групп. Использованные в статье подходы, TF-IDF и *n*-граммы, позволили извлечь и структурировать знания, умения и навыки, полученные из неструктурированной базы данных интернет-вакансий на рынке труда. В результате были найдены ключевые комбинации профессиональных навыков для отдельных профессиональных групп. Предложенный алгоритм позволяет определить и стандартизировать ключевые навыки, которые могут быть использованы для создания системы российских классификаторов по профессиям и навыкам. Кроме того, алгоритм формирует списки ключевых комбинаций навыков, которые высоко востребованы компаниями в каждой конкретной профессиональной группе в сфере информационных технологий.

Ключевые слова: вакансии в сфере информационных технологий; онлайн-вакансии; анализ неструктурированных данных; рынок труда; спрос на навыки; комбинации знаний, умений и навыков работников.

Цитирование: Терников А.А., Александрова Е.А. Спрос на навыки на рынке труда в сфере информационных технологий // Бизнес-информатика. 2020. Т. 14. № 2. С. 64–83. DOI: 10.17323/2587-814X.2020.2.64.83

Введение

Совокупность знаний, умений и навыков (англ. *skills*, далее – ЗУН), предъявляемых работодателями, является одним из наиболее информативных показателей для оценивки спроса на рынке труда, представляя собой описание компетенций, требуемых в различных профессиональных областях (англ. *occupations*). Однако комбинации требуемых ЗУН динамично меняются как с течением времени, так и в зависимости от специфики отрасли, организации, конкретной вакансии. Данные изменения связаны с конъюнктурными колебаниями экономики и реструктуризацией рынка труда. Более того, профессиональные стандарты, формируемые на основе специфики системы образования, теряют свою гибкость к данным изменениям и довольно быстро устаревают. Отдельный интерес представляет собой проблема выявления ключевых наборов ЗУН на рынке труда для различных профессиональных областей в сфере информационных технологий (далее – ИТ), чему и посвящена настоящая статья.

Исследования в области идентификации комбинаций ЗУН, востребованных на рынке труда, интересны именно применительно к области информационных технологий, по нескольким причинам. Во-первых, сфера ИТ крайне динамична, и требования, предъявлемые к ЗУН, интенсивно меняются с течением времени [1–8]. Во-вторых, ЗУН достаточно ясно и просто классифицируются по профессиональным областям вследствие наличия точных формулировок языков программирования, стека технологий, графических интерфейсов пользователя и т.п. [9–11]. В-третьих, внедрение новых технологий сопровождается появлением новых рабочих задач, что требует изменений в части комбинаций ЗУН работников [12–16].

ИТ имеют высокий спрос на рынке труда: технические специалисты с определенными наборами компетенций востребованы в областях, связанных с экономикой, управлением, торговлей и т.д. Таким образом, работодатели формируют спрос на комбинации ЗУН, определяя задачи, предъявляемые в организации к конктретной должности. Не исключено, что уникальные комбинации ЗУН в отдельных сферах деятельности могут формироваться в системе образования не только в ИТ-специальностях, но и будут весьма востребованы среди экономистов, маркетологов, менеджеров, инженеров. Переосмысление образовательной политики в части формирования ЗУН не может происходить без соотнесения со спросом на рынке труда, что требует эффективных инструментов идентификации комбинаций профессиональных ЗУН, востребованных работодателями.

Настоящее иследование посвящено разработке алгоритма, направленного на извлечение и структурирование информации по ключевым ЗУН в разрезе профессиональных областей в сфере ИТ. Целью работы является идентификация тех ЗУН, которые востребованы компаниями-работодателями.

Статья имеет следующую структуру. В первом разделе представлен обзор смежных работ и методов, используемых в целях классификации и кластеризации данных о рынке труда, полученных из открытых интернет-источников. Во втором разделе представлен алгоритм, позволяющий извлечь и классифицировать информацию из неструктурированных объявлений о работе в части ЗУН, а также его применение на данных регионального рынка труда. Третий раздел содержит результаты работы, а именно: списки ключевых ЗУН и их комбинаций для различных профессиональных областей в сфере ИТ. Заключение содержит рекомендации и дискуссию в отношении применения представленного алгоритма.

1. Обзор методов анализа спроса на навыки

Большинство исследований, посвященных анализу спроса на навыки, используют в качестве исходной информации объявления о вакансиях, полученных из открытых интернет-источников [17–25]. Данные источники содержат достаточно обширную информацию о ЗУН и их комбинациях, но зачастую в неструктурированном виде. Основные методы обработки такой информации базируются на техниках обработки естественного языка (англ. *natural language processing*), в частности, на анализе меры TF-IDF (англ. *Term Frequency – Inverse Document Frequency*) и *n*-грамм (последовательности из *n* элементов), а также использовании алгоритмов классификации и кластеризации [17–25].

Онлайн-базы данных вакансий, как правило, имеют неструктурированные текстовые поля,

которые содержат информацию о профессии и необходимых компетенциях. Такие поля заполняются вручную представителями компаний, поэтому требуется применение процедур подготовки данных и алгоритмических методов для извлечения соответствующей информации в стандартизированной форме. Некоторые исследования решают задачу классификации профессиональных должностей по их описанию в онлайн-вакансиях по широко используемыми классификациями профессий и ЗУН, таких как ISCO[1], ESCO[2] и O*NET[3] [17–20]. Другие исследования применяют модели классификации на основе данных, размеченных экспертами [2, 4, 11, 13, 21]. Иначе говоря, выборка анализируется и размечается экспертами в конкретной области, после чего эта информация используется для корректировки алгортмов классификации. Кроме того, исследователи используют подходы кластеризации для определения профессий и навыков при подготовке данных, что позволяет сформировать и разделить профессиональные группы [19, 21–25]. Таким образом, сочетание различных подходов и алгоритмов подготовки и стандартизации данных позволяет заложить основу для анализа спроса на навыки. Краткое описание данных, подходов и алгоритмов, которые используются в смежных работах, представлено в *таблице 1.*

Представленная в таблице информация позволяет обобщить и системтизировать подходы к организации данных, их обработке и методам анализа, выбору критериев для идентификации комбинаций ЗУН.

Все авторы представляют свои алгоритмы извлечения и систематизации информации на основе онлайн-вакансий. Однако способы их реализации различаются в зависимости от поставленной исследовательской задачи. Например, если основная цель исследования связана с процессом сопоставления неструктурированных текстовых полей из объявлений о работе с официальным классификатором по профессиям и навыкам [17, 18, 20, 21], то алгоритмы классификации реализуются на основе нахождения сходных закономерностей в названии должности, их описания и расширенной текстовой информации из официальных классификаторов, включая значительный объем данных ручной разметки экспертов.

Другой подход основывается на ручной корректировке данных после запуска алгоритмов кластеризации [19, 22–25]. Несмотря на различия в целях исследований, в целом применяются общие методы подготовки данных и извлечения стандартизированной информации. Все авторы используют подход TF-IDF и токенизацию (включая удаление стоп-слов и стемминга) для того, чтобы обрабатывать большое количество неструктурированной текстовой информации. Кроме того, *n*-граммы используются для извлечения более чем одного слова. В результате получается набор унифицированных шаблонов (например, профессий и ЗУН). Однако авторы не предоставляют обобщенного алгоритма для сопоставления разных вариантов написания одного и того же шаблона в процессе обработки данных.

Выбор профессиональных областей и ЗУН зависит от информации из официальных классификаций и объема данных. Уровень обобщения таких групп зависит от экспертных оценок, основанных на характеристиках данных. В целом в смежных исследованиях объем данных доступен за однолетний период, и поиск подходящих шаблонов для неструктурированных полей упрощен только для объявлений о работе, опубликованных на одном языке.

Исследователи обращаются к разным метрикам оценки моделей классификации и кластеризации. Авторы используют токенизацию для необработанных текстов и *n*-граммы для построения набора терминов. Анализ сходства различных терминов и описаний с шаблоном реализуется при помощи различных индексов. В случае сохранения порядка слов используется расстояние Левенштайна (англ. *Levenshtein*), но если только пересечение одинаковых терминов является ценным для обнаружения сходства – предпочтительным является индекс Жаккара (англ. *Jaccard*).

[1] International Standard Classification of Occupations, https://www.ilo.org/public/english/bureau/stat/isco/

[2] European Skills/Competences, Qualifications and Occupations, https://ec.europa.eu/esco/portal/home

[3] The Occupational Information Network, https://www.onetonline.org//

Таблица 1.

Смежные работы в области анализа онлайн-вакансий

Основное направление	Данные				Извлеченные группы		Методы обработки данных				Объем ручной разметки данных	Меры схожести	Авторы
	Объем	Период	Источники	Язык	Профессиональные области	ЗУН	TF-IDF	*n*-граммы	Кластеризация	Классификация			
Кластеризация профессиональных областей	1460	4 месяца (апрель – июль 2018)	LinkedIn	Английский	8	96	+	+	Unweighted Pair Group Mean Average method	–	>900	Jaccard	[24]
	12849	7 месяцев (июль –ноябрь 2015 и октябрь – ноябрь 2016)	5 источников	Английский	69	НД*	+	–	Latent Semantic Indexing, Singular Value Decomposition	–	750	Cosine	[19]
Классификация профессиональных областей	75546	27 месяцев (февраль 2013 – апрель 2015)	WollyBI	Итальянский	9	НД	+	+	–	SVM (linear & RBF Kernel); Random Forest; NN	57740	Levenshtein, Jaccard, Sørensen–Dice	[17]
	40000	1 месяц (год не указан)	12 источников	Итальянский	62	542	+	–	Weighted Word Pairs (WWP) extraction	LinearSVC & Perceptron classifier.	412	Levenshtein	[21]
Кластеризация профессиональных областей и извлечение ЗУН	2786	3 месяца (осень 2015)	10 источников	Английский	4	180	+	+	Latent Dirichlet Analysis	–	180	Centrality degree	[22]
	2638	3 месяца (май – июль 2018)	Indeed.com	Английский	48	480	+	–	Latent Dirichlet Analysis	–	480	% встречаемости	[23]
	1050	5 месяцев (июль – ноябрь 2017)	6 источников	Английский	2	2335	+	–	Latent Class Analysis, Singular Value Decomposition	–	НД	VARIMAX Rotation	[25]
Классификация профессиональных областей и извлечение ЗУН	6222	4 месяца (июнь – сентябрь 2015)	3 источников	Итальянский	6	НД	+	+	–	SVM (linear & RBF Kernel); Random Forest; NN	1007	Random–Forest importance	[20]
	~2 млн	24 месяца (2016–2017)	WollyBI	Итальянский	22	8	+	+	–	SVM	НД	Levenshtein, Jaccard, Sørensen–Dice	[18]

*НД – «Нет данных»

2. Алгоритм анализа спроса на навыки и описание используемых данных

Предлагаемый алгоритм, позволяющий реализовать анализ спроса на навыки, организован для данных онлайн-вакансий. Эти данные получены из интерфейса прикладного программирования с открытым исходным кодом HeadHunter[4] – крупнейшей российской онлайн-платформы по подбору персонала[5]. Типичная структура онлайн-вакансии представлена в *таблице 2.*

Наряду с представленной структурой данных и методами, использованными в смежных работах, основной интерес данной статьи заключается в организации процесса извлечения ЗУН из неструктурированных данных. Таким образом, определение наиболее востребованных ЗУН в профессиональных группах может быть реализовано досточно точно. Несмотря на использование алгоритмов классификации для агрегирования профессий в смежных работах, текущий набор данных уже кодифицирован производителем данных. Таким образом, на предварительном этапе анализа предположим, что профессиональные группы уже присвоены и существуют. Для того чтобы организовать описание алгоритма, необходимо формализовать и упростить несколько концепций.

Определение 1. Онлайн-вакансия. Пусть I – набор числовых кодов вакансий; H – набор числовых кодов специализаций; S – набор уникальных ЗУН. Пусть $V = \{v_1, \ldots, v_n\} : n \in \mathbb{N}$ – набор вакансий. Тогда онлайн-вакансия v представляет собой 6-элементный кортеж $v = (i, C, d, p, g, K)$, где $i \in \mathrm{I}$, $C \subseteq H : |C| \in \{\overline{1,6}\}$ – подмножество числовых кодов специализаций, d – дата публикации вакансии, p – наименование вакансии, g – описание вакансии, $K \subseteq S : |K| \leq 30$ – подмножество ЗУН для данной вакансии.

Настоящее исследование сосредоточено на секторе ИТ, а предлагаемый подход протестирован на региональном рынке труда (г. Санкт-Петербург). Онлайн-вакансии из сферы ИТ в Санкт-Петербурге были собраны с 2015 по 2019 годы (для этого был использован официальный классификатор HeadHunter для получения вакансий в сфере информационных технологий по числовому коду специализации). Каждое наблюдение для конкретной вакансии (v) содержит числовой код вакансии (i), числовой код специализации по кслассификации HeadHunter (C) и список ЗУН, требуемых в данной вакансии (K). Основные задачи исследования сосредоточены на процессе извлечения и структурирования ЗУН. Таким образом, используется та часть данных, где поле, со-

Таблица 2.

Структура типового объявления портала HeadHunter

Поле	Тип поля		Описание
	Структурированное	Неструктурированное	
Vacancy ID	+		Числовой код вакансии
Specialization ID	+		Набор (от 1 до 6 включительно) числовых кодов специализаций[6]
Published date	+		Длинный формат даты публикации вакансии
Position Name		+	Текст (наименование вакансии)
Job description		+	Текст (описание вакансии)
Key skills		+	Набор текстов (30 – максимум: каждый не более 100 символов), содержащих ЗУН

[4] HeadHunter API, https://dev.hh.ru

[5] SimilarWeb: websites ranking, https://www.similarweb.com/top-websites/russian-federation/category/jobs-and-career/jobs-and-employment

[6] HeadHunter API: Specializations, https://api.hh.ru/specializations

Таблица 3.

Распределение вакансий по кодам специализаций HeadHunter в выборке

Код специализации HeadHunter	Доля, %	Наименование
1.221	20,37	Программирование, Разработка
1.82	7,99	Инженер
1.9	4,90	Web инженер
1.89	4,52	Интернет
1.10	4,18	Web мастер
1.327	3,83	Управление проектами
1.225	3,42	Продажи
1.137	3,21	Маркетинг
1.272	3,11	Системная интеграция
1.295	3,04	Телекоммуникации
1.211	2,99	Поддержка, Helpdesk
1.117	2,90	Тестирование
1.270	2,86	Сетевые технологии
1.273	2,73	Системный администратор
1.25	2,69	Аналитик
1.172	2,62	Начальный уровень, Мало опыта
1.50	2,25	Системы управления предприятием (ERP)
1.400	2,11	Оптимизация сайта (SEO)
1.536	2,10	CRM системы
1.474	2,04	Стартапы
1.359	1,75	Электронная коммерция
1.116	1,58	Контент
1.475	1,51	Игровое ПО
1.113	1,50	Консалтинг, Аутсорсинг
1.246	1,42	Развитие бизнеса
1.420	1,39	Администратор баз данных
1.395	1,04	Банковское ПО
1.203	1,01	Передача данных и доступ в интернет
1.110	0,98	Компьютерная безопасность
1.161	0,86	Мультимедиа
1.296	0,67	Технический писатель
1.274	0,66	Системы автоматизированного проектирования
1.3	0,64	CTO, CIO, Директор по IT
1.277	0,61	Сотовые, Беспроводные технологии
1.30	0,34	Арт–директор
1.232	0,18	Продюсер

держащее ЗУН, является непустым. Полученная выборка состоит из 63869 вакансий, опубликованных с мая 2015 года по сентябрь 2019 года. Каждая вакансия включает от одного до шести профессиональных кодов (специализаций HeadHunter). Распределение по 36 направлениям (внутри группы сферы ИТ) представлено в *таблице 3.*

Несмотря на наличие представленного распределения кодов специализаций HeadHunter, некоторые сферы могут быть удалены или объединены в одну большую подгруппу. В соответствии с классификацией, введенной в работе [20], мы определяем семь групп ИТ-специалистов. После перегруппировки вакансий и удаления некоторых кодов специализаций получено 56000 наблюдений (вакансий). Доля удаленных профессиональных областей составляет 10,2%. Новое распределение среди оставшихся агрегированных профессиональных групп вакансий и их наименования представлены в *таблице 4.*

Таблица 4.

Распределение вакансий в сфере ИТ по укрупненным профессиональным группам

Наименование	Сокращение	Доля, %	Коды специализаций HeadHunter
ИТ–специалисты высокого уровня квалификации	high	13,10	1.327, 1.272, 1.25, 1.113, 1.3
ИТ–специалисты низкого уровня квалификации	low	3,66	1.172, 1.296
Инженеры	engineers	16,18	1.82, 1.295, 1.117, 1.277
Разработчики программного обеспечения	soft	22,67	1.221
Веб и мультимедиа разработчики	web	20,13	1.9, 1.89, 1.10, 1.400, 1.475, 1.161
Системные администраторы и администраторы баз данных	admin	19,30	1.211, 1.270, 1.273, 1.50, 1.536, 1.420, 1.395, 1.203, 1.110
Другие	others	4,96	1.474, 1.359, 1.274

Определение 2. Профессиональная область (группа). Пусть H – набор кодов специализаций HeadHunter, а O – упорядоченный набор укрупненных профессиональных групп. Тогда профессиональная область $o \in O$ представляет собой 2-элементный кортеж $o = (L, a)$, где $L \subseteq H$ – подмножество кодов специализаций, соответствующих укрупненной профессиональной группе с сокращенным наименованием a в текстовом формате.

Для упрощения дальнейшего анализа извлечения ключевых ЗУН для определенных профессиональных групп (O, где $|O| = 7$) процесс обработки данных организован на всей выборке (в течение 5-летнего периода). В качестве допущения для такой агрегации используется относительное распределение вакансий по профессиональным группам (*рисунок 1*). Таким образом, относительное распределение вакансий по укрупненным профессиональным группам меняется незначительно во временной перспективе.

Для целей настоящего исследования ключевые ЗУН и их комбинации должны быть унифицированы и извлечены из набора вакансий (V). Однако, прежде чем вводить алгоритм извлечения ЗУН, каждая онлайн-вакансия, которая может относиться к нескольким профессиональным областям, должна быть сопоставлена с новой структурой данных (онлайн-вакансия ИТ).

Определение 3. Онлайн-вакансия ИТ.

Пусть $J \subseteq V$ – набор онлайн-вакансий ИТ, где $J = \{j_1, ..., j_m\}$: $m \le n$, $m \in \mathbb{N}$. Пусть $\mathbf{c}$ обозначает коды специализаций онлайн-вакансии: $\mathbf{c} = (c_1, ..., c_z) \in C$, где $z \le 6$. Пусть $\mathcal{L}$ –упорядоченный набор меток с отношением $(L, a) \in O \mapsto \mathcal{L}$ в следующей форме: $\mathcal{L} = (\lambda_1, ..., \lambda_q)$, где $q = |O|$. Введем функцию

$$f(\mathbf{c}, \lambda_q) = \begin{cases} 1, \mathbf{c} \subseteq L_q \\ 0 \end{cases},$$

которая ставит в соответствие код специализации o и код укрупненной профессиональной области из $\mathcal{L}$. Введем отношение $H : C \mapsto \mathcal{L}$, которое обеспечивает классификацию по нескольким меткам и отображает набор укрупненных профессиональных областей $\mathcal{L}$ на основе кодов специализаций: $\tilde{O} = H(\mathbf{c}) = \{\lambda \in \mathcal{L} \; f(\mathbf{c}, \lambda) = 1\}$, где $\tilde{O}$ – набор сокращенных наименований укрупненных профессиональных групп. Таким образом, онлан-вакансия ИТ j – это 3-элементный кортеж $j = (i, \tilde{O}, K)$.

В *таблице 5* представлено распределение полученных вакансий по укрупненным группам профессий. Полученное распределение профессиональных областей не является однородным. Иначе говоря, только относительно небольшая доля вакансий (6–30%) относится к одной укрупненной профессиональной группе. Другие вакансии связаны с несколькими такими группами. Таким образом, в последующем анализе необходимо разделение ЗУН, непосредственно связанных с конкретной профессиональной областью.

Для предоставления результатов поиска ключевых ЗУН по профессиональным группам разработан алгоритм извлечения ЗУН из неструктурированных онлайн-вакансий. Здесь и далее алгоритм извлечения ЗУН применен для выборки онлайн-вакансий сектора ИТ.

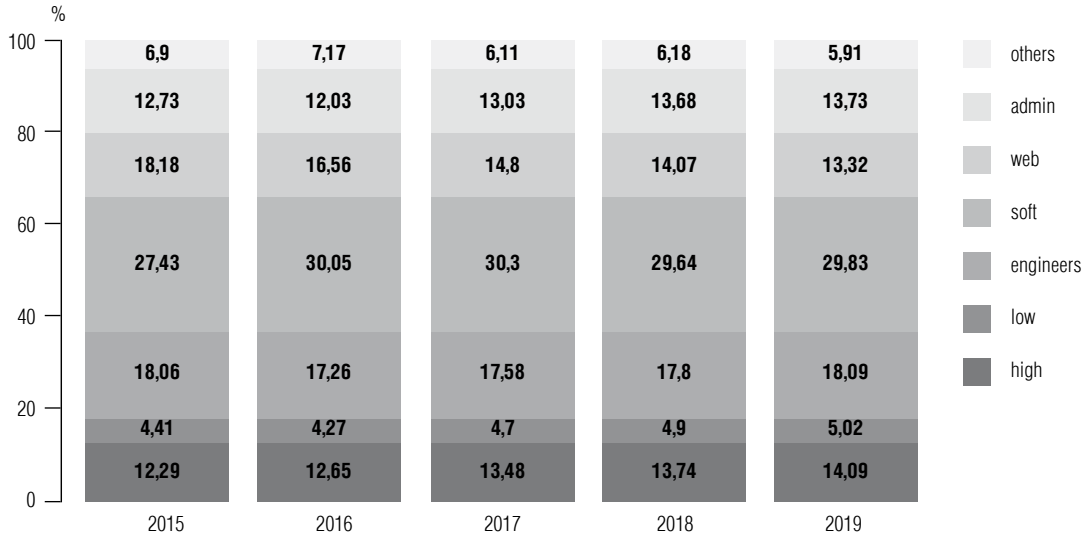


Рис. 1. Распределение количества вакансий по укрупненным профессиональным группам в динамике

Таблица 5.

Распределение укрупненных профессиональных групп в выборке

Сокращение	Количество вакансий	Доля вакансий, соответствующих только одной укрупненной группе профессий, %
high	19266	16,28
low	5383	17,28
engineers	23787	10,15
soft	33333	29,78
web	19312	9,29
admin	20825	6,18
others	7293	15,80

Вход: онлайн-вакансия ИТ (J).

Выход: набор стандартизированных ЗУН ($\tilde{K}$), сопоставленных с J.

1. Пусть $\tilde{S} \subseteq S$ набор уникальных текстовых описаний ЗУН из J
2. Пусть B 2-элементный кортеж: текстовое описание ЗУН и его частота встречаемости в J
3. **foreach** $\tilde{s}_i \in \tilde{S}$ **do**
4. $b_i \leftarrow$ частота встречаемости $\tilde{s}_i$ в J
5. $B[i] = (\tilde{s}_i , b_i)$
6. **end foreach**
7. сортируем B в убывающем порядке по b
8. **procedure** FrequentTerms(h, t)
9. $\tilde{h} \leftarrow$ подмножество h, если $h_i > t, \forall\, h_i \in h$
10. **return** $\tilde{h}$
11. **end procedure**
12. вводим пороговое значение t
13. $\tilde{B} \leftarrow$ FrequentTerms($B(b)$, t)
14. $T \leftarrow$ 3-элементный кортеж ЗУН после ручной разметки $T = (u, x, xs)$, где u идентификатор стандартизированного ЗУН, x – наименование в текстовом формате, xs – набор синонимов в текстовом формате для пары (u, x)
15. **function** Tokenizer(j)
16. нормализация пробельных символов
17. удаление знаков пунктуации
18. строчный регистр
19. стемминг (на английском и русском языках)
20. удаление стоп-слов (на английском и русском языках)
21. **end function**
22. **procedure** NGrams(J, n)
23. **for** j in J **do**
24. $G \leftarrow$ n-граммы размера n для Tokenizer(j)
25. положим G в ngramterms
26. **end for**
27. **return** ngramterms

end procedure
вводим пороговые значения t_1, t_2, t_3
ngram1:= FrequentTerms(NGrams($B(s)$, 1), t_1)
ngram2:= FrequentTerms(NGrams($B(s)$, 2), t_2)
ngram3:= FrequentTerms(NGrams($B(s)$, 3), t_3)
для полученных баз n-грамм проводится ручная разметка (чистка неинформативных терминов)
каждое наблюдение в данных базах n-грамм имеет набор идентификаторов (i, $\tilde{s}$)
procedure MatchTerms(X, Y)
Пусть L набор уникальных комбинаций из X и Y, где $L = \{l_1 | l_1 \in X, l_2 | l_2 \in Y\}$; l_1, l_2 наборы идентификаторов (i, $\tilde{s}$)
for l_1, l_2 in L **do**
$M \leftarrow$ индекс Жаккара: $\frac{|l_1(i) \cap l_2(i)|}{|l_1(i) \cup l_2(i)|}$
if M >0.5 **do**
положим (l_1, l_2) в termsmatched
end if
end for
return termsmatched
end procedure
для каждой комбинации из: ngram1, ngram2, ngram3 проводим MatchTerms(X, Y) → M1, M2, M3
для каждой комбинации из: T, M1, M2, M3 проводим MatchTerms($\tilde{X}$, $\tilde{Y}$), где $\tilde{X} := T$, $\tilde{Y} := T$ {M1,M2,M3}
$\tilde{K} \leftarrow X$ left-join Y
$\tilde{K}$ – полученная база данных со стандартизированными ЗУН, их синонимами и наборами n-грамм

Применение алгоритма к выборке данных онлайн-вакансий ИТ и извлечение ЗУН происходит следующим образом. В выборке представлены 13347 неструктурированных уникальных ЗУН. Описания таких ЗУН не унифицированы. Иначе говоря, каждая компания может ввести свою собственную текстовую строку от 1 до 100 символов. Например, запись одного и того же ЗУН может содержать одно слово/фразу или предложение, содержащее такие слова, разделенные символами пунктуации или пробелами. Для того, чтобы автоматизировать извлечение определенных ЗУН и унифицировать различные формы обозначения одного и того же термина, используются методы анализа и обработки текстовой информации.

В соответствии с [20], на первом этапе предварительной обработки данных можно построить n-граммы слов. Из векторного корпуса (TF-IDF) навыков, представленных в описаниях вакансий, были созданы уни-, би- и триграммы с использованием следующей токенизации: удаление всех знаков препинания и лишних пробелов, замена прописных букв строчными, стемминг слов на английском и на русском языках, удаление стоп-слов. В рамках данных терминов исходная структура и формулировки ЗУН были сохранены. На первом этапе были удалены неинформативные термины. Для базы данных униграмм получено 348 записей из 5234 неуникальных терминов; для биграмм – 577 из 1090; для триграмм – 110 из 303. Следующий этап – создание базы данных синонимов для уже полученных шаблонов. HeadHunter API (рекомендации по ключевым навыкам) позволяет получить часть синонимов для последующей ручной обработки полученных терминов. После получения данных синонимов была сформирована матрица терминов из 707 элементов (1296 записей для ручной проверки). Третий этап заключается в добавлении и корректировке терминов на основе опроса Stack Overflow Developer Survey[7] (108 терминов-наи-

[7] Stack Overflow Annual Developer Survey, https://insights.stackoverflow.com/survey/

менований наиболее популярных ИТ-технологий) и окончательном исправлении соответствующих терминов (база данных с исходными терминами и их синонимами).

В результате было получено 435 стандартизированных терминов (ЗУН) и 420 синонимов для них. Такой набор данных содержит как технические («жесткие»), так и нетехнические («мягкие») навыки для данного сектора. На последнем этапе обработки данных производится пересечение необработанных терминов (кодифицированных при помощи уникальных идентификационных кодов), точно совпадающих с конкретными терминами, полученными из скорректированного вручную списка синонимов HeadHunter и результатов, полученных из матриц TF-IDF (для уни-, би-, триграмм). В результате получаются пары «идентификатор ЗУН – термин» в текстовом формате. Для того, чтобы автоматически определить схожесть нескольких терминов (на основе уникального набора идентификаторов для каждого термина) и сопоставить остальные данные с заданными стандартизированными терминами, используется индекс Жаккара. Например, сходство между двумя наборами слов (терминов) *A* и *B* можно найти при помощи следующей формулы:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Эта мера подходит для сравнения категориальных данных, и ее значение находится в диапазоне от 0 до 1 включительно. Тем не менее, выбор порога отсечения зависит от структуры данных и целей исследования. В качестве порога для идентификации близких терминов используется уровень выше 0,5 (после ручной обработки полученных данных). Таким образом, 53672 вакансий (95,8% от исходной выборки) содержат, по крайней мере, один из стандартизированных ЗУН из ранее полученного набора данных терминов и их синонимов. Процентное распределение ТОП-20 стандартизированных ЗУН (по частоте встречаемости в выборке) среди всего набора данных представлено в *таблице 6.*

Анализ ключевых востребованных (со стороны работодателя) ЗУН (и их комбинаций) основывается на определении специфичных ЗУН для конкретной профессиональной области, которые в то же время достаточно часто встречаются в соответствующих вакансиях.

После подготовки данных набора вакансий получена следующая структура данных: 305217 наблюдений (определенный ЗУН из вакансии), где каждое наблюдение имеет идентификатор стандартизированного ЗУН, идентификатор вакансии и код укрупненной профессиональной группы. Для того, чтобы обеспечить классификацию ЗУН в соответствии с указанными группами вакансий, для каждой из групп вакансий был проведен поиск пар и триплетов ЗУН. Из полученных пар и триплетов ЗУН (отличных от нуля по индексу схожести Жаккара) были извлечены ЗУН с высокой степенью совпадения. Общая схема предложенного алгоритма, реализованного для предложенного набора данных, представлена на *рисунке 2.*

На первом этапе по 435 стандартизованным ЗУН, их парам (C^2_{435} = 94435) и триплетам (C^3_{435} = 1362345) были найдены индексы сходства Жаккара для каждой из семи укрупненных групп вакансий. Далее,

Таблица 6.

ТОП-20 ЗУН по частоте встречаемости в выборке ИТ сектора

ЗУН	Частота присутствия в базе данных ЗУН, %
HTML/CSS	6,73
JavaScript	4,69
1С	3,48
SQL	3,25
PHP	2,63
Git	2,53
Linux	2,32
Java	2,28
MySQL	1,86
Навыки переговоров	1,61
Sales Skills	1,52
Деловая коммуникация	1,51
English	1,45
Testing Framework	1,43
Python	1,40
jQuery	1,36
C/C++	1,30
ООП	1,29
C#	1,28
.net	1,27

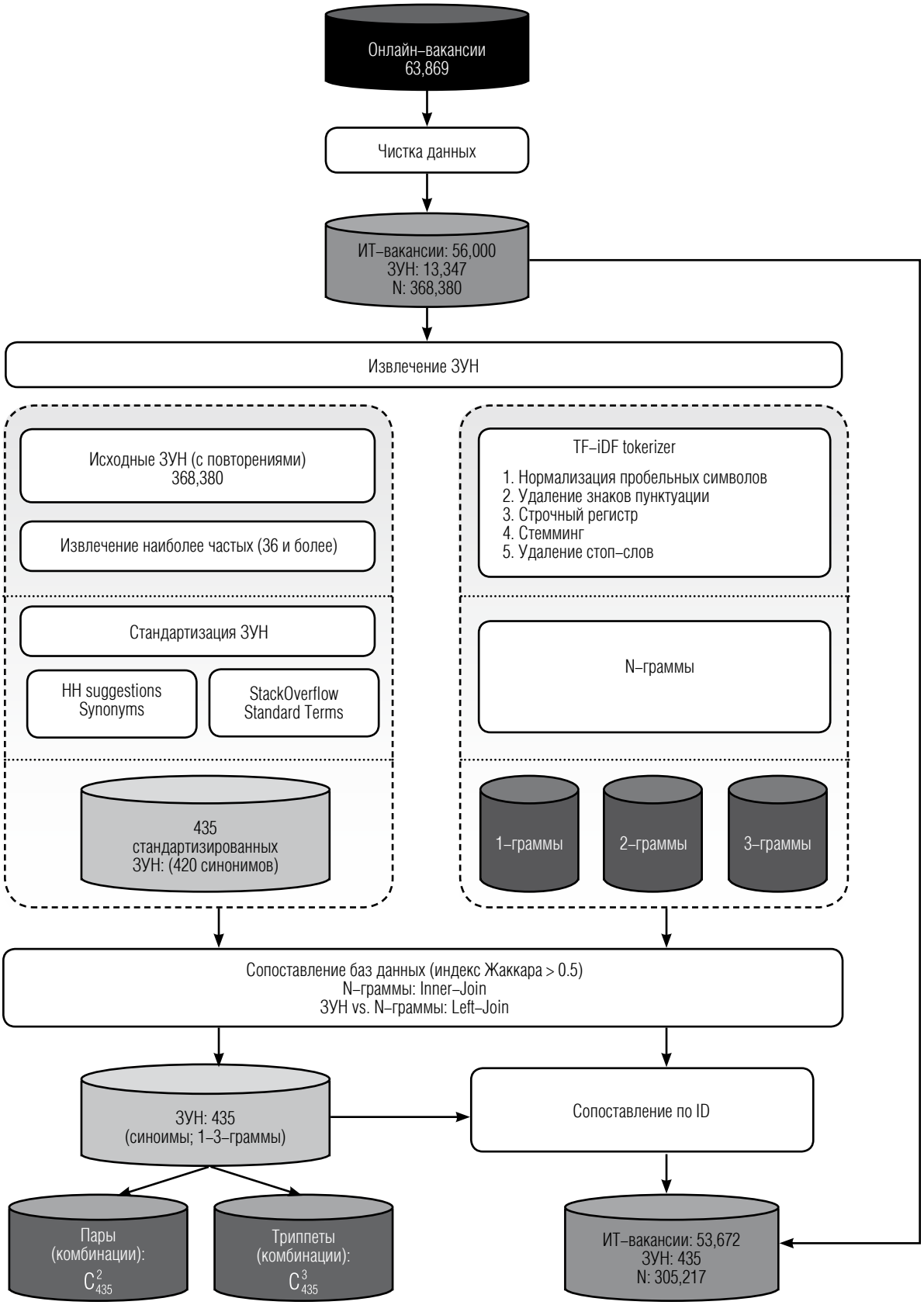


Рис. 2. Схема реализации алгоритма извлечения ЗУН

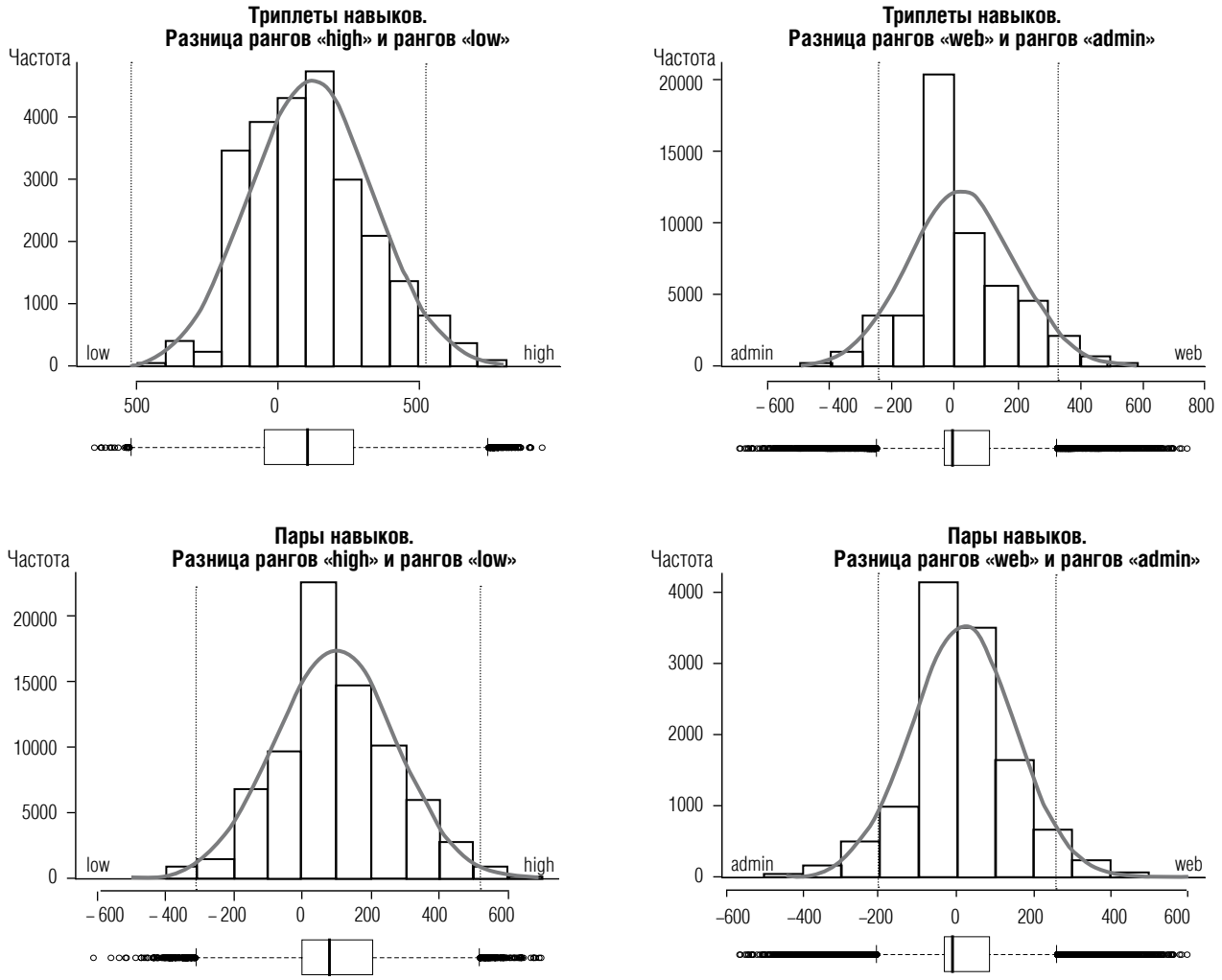


Рис. 3. Распределение разностей рангов индексов схожести Жаккара по укрупненным группам вакансий

используя пары и триплеты ЗУН (уникальные сочетания без повторений), каждый набор ЗУН был проранжирован по индексу Жаккара в пределах их квантильного распределения (с шагом в 0,1% для пар и триплетов).

Для каждой пары и триплета ЗУН такой показатель рассчитывался исходя из количества вакансий, в которые входят те или иные комбинации ЗУН. На следующем этапе, были найдены разности в рангах для каждой пары из укрупненных групп вакансий. Далее, для выявления специфичных наборов ЗУН для каждой профессиональной области были найдены выбросы в предложенном распределении разностей рангов. С точки зрения статистического аппарата, близость распределения разностей рангов к нормальному позволяет обеспечить соответствующий отбор специфичных и при этом ключевых ЗУН, которые определяют различные укрупненные группы вакансий. Например, несколько пар таких групп представлены на *рисунке 3* (в качестве границ для отсечения ЗУН используются границы усов из ящика с усами: 1,5 IQR[8] ниже и выше для подходящей разности в квантильных рангах).

Для извлеченных пар и триплетов применяется процедура добавления уникальных наборов ЗУН для каждой комбинации профессиональных групп. Таким образом, на пересечении ЗУН извле-

[8] IQR – межквартильный размах

каются только значения, составляющие более 95% (по разнице в квантилях) для того, чтобы обнаружить ключевые и специфичные наборы ЗУН. На следующем этапе для каждой пары групп вакансий ($C_7^2 = 21$) были определены ключевые определяющие ЗУН и их комбинации. Таким образом, к ЗУН, которые являются уникальными в определенной группе вакансий в последнем дециле, были добавлены уникальные ЗУН, исходя из их пересечения с другими группами вакансий. В итоге были получены три матрицы 7×7, где на пересечении *i*-й строки и *j*-го столбца ($i \neq j$) находятся уникальные наборы ЗУН (кодированные по-отдельности для уникальных ЗУН, их пар и триплетов) и уникальные ЗУН из *i*-й группы (превышающие порог в 95%) по сравнению с *j*-й группой вакансий. Наконец, данные ЗУН были распределены следующим образом: наличие основных ЗУН, которые определяют, что каждая профессиональная группа была уже установлена, использовалось для пересечения строк (для каждой из заданных матриц), что позволило получить список ключевых и специфиных ЗУН для данной профессиональной группы. Используются следующие пороговые значения: для пар пороговое значение не менее 2/3 отличий от других групп (4 и более отличных групп из оставшихся шести); для триплетов – 100% отличий (6 из 6). Используя приведенную выше последовательность действий, списки таких навыков были получены для каждой конкретной профессиональной группы вакансий, которые, в свою очередь, представляют собой ключевые (наиболее востребованные) навыки присущие конкретной профессиональной группе.

3. Результаты

На этапе определения ключевых и различных ЗУН для разных профессиональных групп в сфере ИТ извлекаются наиболее популярные из них. В соответствии с группами профессий, такие ЗУН могут быть представлены в виде облаков слов по ТОП-50 ЗУН для каждой профессиональной группы по количеству в описании вакансий (*рисунок 4*).

Однако при наличии вакансий, связанных с несколькими профессиональными группами, некоторые ЗУН дублируются. Таким образом, на этапе извлечения пар и триплетов ЗУН, используя пересечение специфичных ЗУН, такое дублирование нивелируется. Наиболее востребованные и в то же время спепцифичные профессиональные пары ЗУН представлены в *таблице 7*, триплеты – в *таблице 8*[9].

Наборы востребованных ЗУН могут быть идентифицированы для разных профессиональных групп. Кроме того, используя пары и триплеты ЗУН, становится возможным получить конкретные их комбинации. Таким образом, предлагаемые методы обработки и извлечения ЗУН могут быть полезны для более широкого понимания и анализа спроса на рынке труда, а также предоставляют дополнительную информацию для системы образования с точки зрения поддержания актуальности и обновления образовательных стандартов в части ЗУН.

Заключение

Применимость представленного в статье алгоритма имеет ряд особенностей.

Используемая база данных имеет предопределенный набор профессиональных групп. С одной стороны, это упрощает реализацию алгоритма. С другой стороны, полученные результаты свидетельствуют о значительном пересечении одинаковых ЗУН для разных групп. Не исключено, что детерминированная структура профессиональных групп не является однозначной, что создает излишнюю вариативность при идентификации уникальных комбинаций ЗУН для укрупненных профессиональных групп. Другими словами, введение классификации или кластеризации для выявления профессиональных групп может улучшить общие результаты. Тем не менее, предоставленный список комбинаций ЗУН построен с точки зрения сохранения специфики наборов ЗУН для конкретной группы профессий.

В одной конкретной вакансии могут быть заявлены совершенно разные профессиональные группы. В таком случае группировка ЗУН (например, на «мягкие» и «жесткие» навыки) может быть использована в качестве дополнительного фактора при классификации. Более того, существуют ЗУН, связанные как с конкретной технологией, так и программными средами для ее реализации, которые не могут быть отделены друг от друга.

Введение большей последовательности слов в *n*-граммах (4 и более) может предоставить больше

[9] Полные списки полученных пар и триплетов могут быть предоставлены авторами по запросу

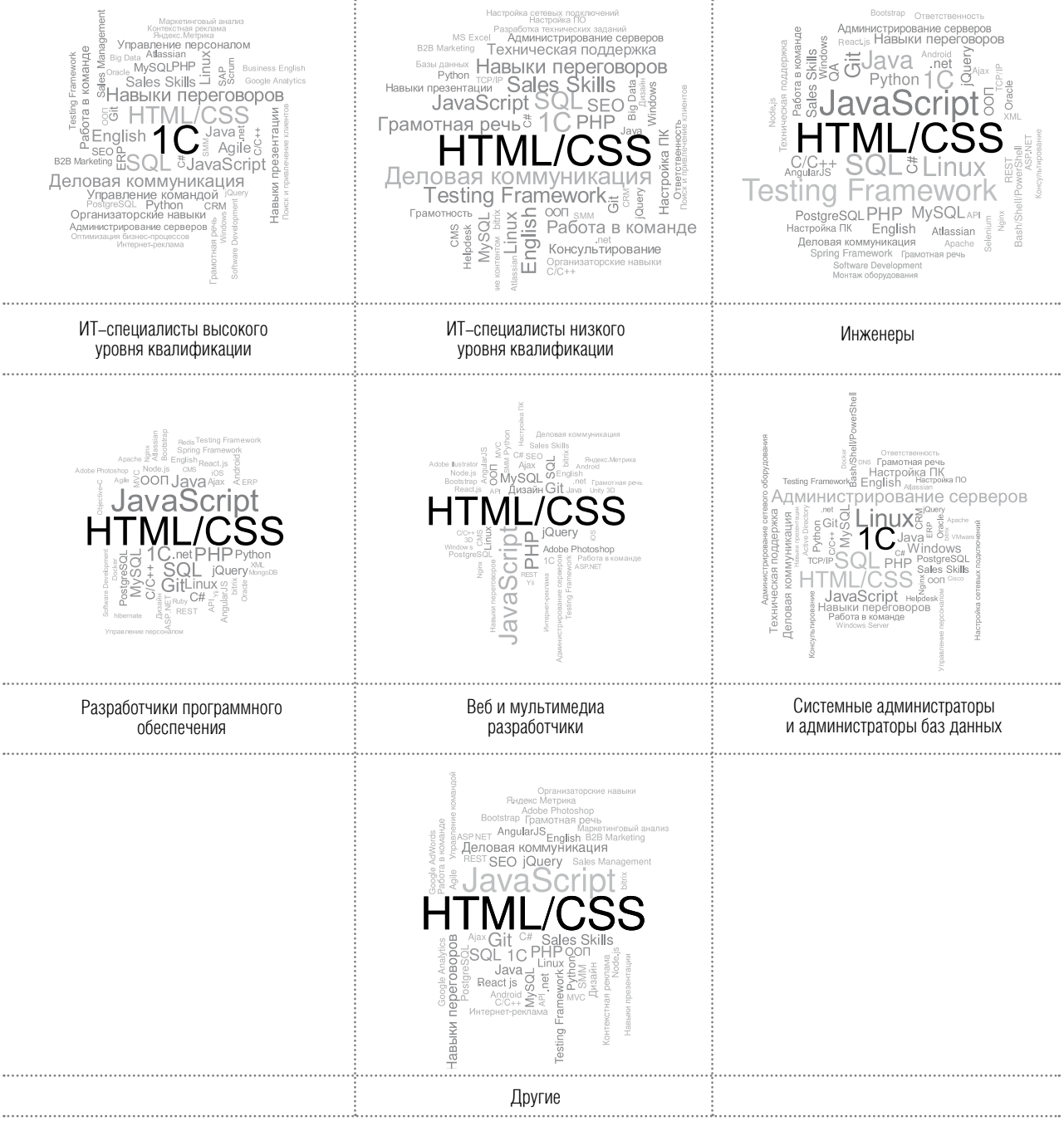


Рис. 4. ТОП–50 ЗУН по профессиональным группам

информации для извлечения ЗУН из неструктурированных баз данных. Однако временная сложность для вычисления таких алгоритмов может быть чрезвычайно высокой, что может потребовать упрощения вычисления метрик подобия (например, с использованием хеш-функций и приближенных формул).

Расширенная реализация алгоритма может быть нацелена на выявление ключевых ЗУН для других секторов рынка труда, учет изменений во времени и организацию межрегиональных сравнений. Результаты могут быть агрегированы с использованием официальной статистики (если цели работы будут направлены на решение общеэкономических вопросов).

Таблица 7.

Ключевые пары ЗУН в профессиональных группах

ЗУН №1	ЗУН №2	Индекс Жаккара	Группа
Dart	Flutter	0,167	high
Billing	Solaris	0,136	high
Arduino	Raspberry Pi	0,125	high
Технические средства информационной защиты	Assembly	0,073	high
Средства криптографической защиты информации	Assembly	0,065	high
Автоматизация производства	САПР	0,125	low
CCNA	OSPF	0,125	low
A/B тесты	Mobile Marketing	0,100	low
Arduino	ARM	0,083	low
Оптимизация бизнес–процессов	Citrix	0,038	low
Системы мониторинга сетей	Google Cloud Platform	0,071	engineers
Cordova	Xamarin	0,065	engineers
Управление персоналом	Яндекс.Метрика	0,014	engineers
Elasticsearch	Node.js	0,011	engineers
MS SharePoint	Windows	0,010	engineers
Firebase	Google Cloud Platform	0,083	soft
Грамотная речь	Составление договоров	0,015	soft
Elasticsearch	Yii	0,011	soft
Scrum	TFS	0,010	soft
Контекстная реклама	Поиск и привлечение клиентов	0,006	soft
Business Intelligence Systems	Olap	0,063	web
3D	Altium Designer	0,022	web
SPA	Unit Testing	0,018	web
Написание статей	Google AdWords	0,017	web
API	Mercurial	0,016	web
Технический аудит сайта	Технический перевод	0,074	admin
Аналитические исследования	Системный анализ	0,033	admin
Apache	Windows Server	0,029	admin
REST	Xsd	0,022	admin
API	Xsd	0,016	admin
Корректура текстов	Adobe Lightroom	0,111	others
Мобильность	Billing	0,111	others
Pandas	Wifi networks	0,100	others
Технический аудит сайта	SMO	0,091	others
A/B тесты	Business Analysis	0,080	others

Таблица 8.

Ключевые триплеты ЗУН в профессиональных группах

ЗУН №1	ЗУН №2	ЗУН №3	Индекс Жаккара	Группа
ARM	GCC	Raspberry Pi	0,019	high
Медиа–планирование	Планирование маркетинговых кампаний	Facebook	0,018	high
CentOS	EJB	NetBeans	0,017	high
Обработка видео	Adobe Premier Pro	SketchUp	0,013	high
Бизнес–планирование	Mobile Marketing	Product Marketing	0,012	high
Автоматизация производства	КИПиА	САПР	0,050	low
Автоматизация технологических процессов	КИПиА	САПР	0,048	low
Автоматизация производства	Автоматизация технологических процессов	САПР	0,043	low
Debian	OSPF	VLAN	0,043	low
Аналитические исследования	Business Analysis	Product Marketing	0,038	low
Обработка видео	Обработка изображений	Adobe Lightroom	0,020	engineers
Ведение переписки на иностранном языке	Написание пресс–релизов	Технический перевод	0,018	engineers
Написание пресс–релизов	Письменный перевод	Технический перевод	0,015	engineers
Обработка изображений	Adobe After Effects	Adobe Lightroom	0,014	engineers
FreeBSD	OSPF	VLAN	0,014	engineers
Поисковая оптимизация сайтов	Работа с биржами	Технический аудит сайта	0,021	soft
Математический анализ	MATLAB	R	0,017	soft
Математическая статистика	MATLAB	R	0,016	soft
Корректура текстов	Написание пресс–релизов	Подготовка презентаций	0,013	soft
Работа с биржами	Технический аудит сайта	SMO	0,013	soft
Microsoft Azure	TensorFlow	Torch/PyTorch	0,020	web
Баннерная реклама	Обработка видео	Adobe Premier Pro	0,016	web
Ведение отчетности	Налоговая отчетность	Billing	0,016	web
Корректура текстов	Написание пресс–релизов	Рерайтинг	0,015	web
Microsoft Azure	Spark	TensorFlow	0,014	web
Математический анализ	Olap	VBA	0,019	admin
Математический анализ	A/B тесты	R	0,018	admin
Chef	LDAP	Wifi networks	0,015	admin
BGP	Chef	LDAP	0,013	admin
Chef	LDAP	OSPF	0,013	admin
Внутренняя оптимизация сайта	Технический аудит сайта	SMO	0,063	others
Внутренняя оптимизация сайта	Российские поисковые системы	SMO	0,048	others
Flask	Pandas	Wifi networks	0,037	others
Мобильность	Электронный документооборот	Billing	0,031	others
Внутренняя оптимизация сайта	Лидогенерация	SMO	0,027	others

В итоге результаты настоящего исследования позволяют выделить несколько значимых моментов. Во-первых, предложенный алгоритм позволяет выявить и стандартизировать ключевые ЗУН, которые могут быть применимы для создания системы русскоязычного классификатора профессий, знаний, умений и навыков. Во-вторых, алгоритм позволяет предоставлять списки наиболее популярных (ключевых) комбинаций ЗУН, которые высоко востребованы компаниями и работодателями для каждой конкретной вакансии. Наконец, гибкость алгоритма позволяет сочетать его с методами классификации и кластеризации данных, которые могут быть полезны для исследований рынка труда. ■

Литература

1. Autor D.H., Levy F., Murnane R.J. The skill content of recent technological change: An empirical exploration // The Quarterly Journal of Economics. 2003. Vol. 118. No 4. P. 1279–1333. DOI: 10.1162/003355303322552801.
2. Bensberg F., Buscher G., Czarnecki C. Digital transformation and IT topics in the consulting industry: A labor market perspective // Advances in consulting research: Recent findings and practical cases / V. Nissen (Ed.). Cham, Switzerland: Springer, 2019. P. 341–357.
3. Christoforaki M., Ipeirotis P.G. A system for scalable and reliable technical-skill testing in online labor markets // Computer Networks. 2015. No 90. P. 110–120. DOI: 10.1016/j.comnet.2015.05.020.
4. Florea R., Stray V. Software tester, we want to hire you! An analysis of the demand for soft skills // 19th International Conference on Agile Processes in Software Engineering and Extreme Programming (XP 2018), Porto, Portugal, 21–25 May 2018. P. 54–67.
5. Goles T., Hawk S., Kaiser K.M. Information technology workforce skills: The software and IT services provider perspective // Information Systems Frontiers. 2008. Vol. 10. No 2. P. 179–194.
6. Johnson K.M. Non-technical skills for IT professionals in the landscape of social media // American Journal of Business and Management. 2016. Vol. 4. No 3. P. 102–122. DOI: 10.11634/216796061504668.
7. Skills for success at different stages of an IT professional's career / L. Kappelman [et al.] // Communications of the ACM. 2016. Vol. 59. No 8. P. 64–70. DOI: 10.1145/2888391.
8. Litecky C.R., Arnett K.P., Prabhakar B. The paradox of soft skills versus technical skills in is hiring // Journal of Computer Information Systems. 2004. Vol. 45. No 1. P. 69–76.
9. Havelka D., Merhout J.W. Toward a theory of information technology professional competence // Journal of Computer Information Systems. 2009. Vol. 50. No 2. P. 106–116.
10. Hussain W., Clear T., MacDonell S. Emerging trends for global DevOps: A New Zealand perspective // IEEE 12th International Conference on Global Software Engineering, Buenos Aires, Argentina, 22–23 May 2017. Vol. 1. P. 21–30. DOI: 10.1109/ICGSE.2017.16.
11. Wowczko I. Skills and vacancy analysis with data mining techniques // Informatics. 2015. Vol. 2. No 4. P. 31–49. DOI: 10.3390/informatics2040031.
12. Bailey J., Mitchell R.B. Industry perceptions of the competencies needed by computer programmers: Technical, business, and soft skills // Journal of Computer Information Systems. 2006. Vol. 47. No 2. P. 28–33.
13. Brooks N.G., Greer T.H., Morris S.A. Information systems security job advertisement analysis: Skills review and implications for information systems curriculum // Journal of Education for Business. 2018. Vol. 93. No 5. P. 213–221.
14. Casado-Lumbreras C., Colomo-Palacios R., Soto-Acosta P. A vision on the evolution of perceptions of professional practice // International Journal of Human Capital and Information Technology Professionals. 2015. Vol. 6. No 2. P. 65–78. DOI: 10.4018/IJHCITP.2015040105.
15. Föll P., Thiesse F. Aligning is curriculum with industry skill expectations: A text mining approach // 25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, 5–10 June 2017. P. 2949–2959.
16. Stal J., Paliwoda-Pękosz G. Fostering development of soft skills in ICT curricula: A case of a transition economy // Information Technology for Development. 2019. Vol. 25. No 2. P. 250–274. DOI: 10.1080/02681102.2018.1454879.
17. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. Classifying online job advertisements through machine learning // Future Generation Computer Systems. 2018. Vol. 86. P. 319–328.
18. Colombo E., Mercorio F., Mezzanzanica M. AI meets labor market: Exploring the link between automation and skills // Information Economics and Policy. 2019. No 47. P. 27–37. DOI: 10.1016/j.infoecopol.2019.05.003.
19. Data mining approach to monitoring the requirements of the job market: A case study / I. Karakatsanis [et al.] // Information Systems. 2017. No 65. P. 1–6. DOI: 10.1016/j.is.2016.10.009.
20. Lovaglio P.G., Cesarini M., Mercorio F., Mezzanzanica M. Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies // Statistical Analysis and Data Mining. 2018. Vol. 11. No 2. P. 78–91. DOI: doi.org/10.1002/sam.11372.
21. Challenge: Processing web texts for classifying job offers / F. Amato [et al.] // IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, California, USA, 7–9 February 2015. P. 460–463.
22. De Mauro A., Greco M., Grimaldi M., Ritala P. Human resources for Big Data professions: A systematic classification of job roles and required skill sets // Information Processing & Management. 2018. Vol. 54. No 5. P. 807–817. DOI: 10.1016/j.ipm.2017.05.004.

23. Gurcan F., Cagiltay N.E. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling // IEEE Access. 2019. No 7. P. 82541–82552.

24. Pejic-Bach M., Bertoncel T., Meško M., Krstić Ž. Text mining of industry 4.0 job advertisements // International Journal of Information Management. 2020. No 50. P. 416–431.

25. Radovilsky Z., Hegde V., Acharya A., Uma U. Skills requirements of business data analytics and data science jobs: A comparative analysis // Journal of Supply Chain and Operations Management. 2018. Vol. 16. No 1. P. 82–101.

Об авторах

Терников Андрей Александрович

аспирант (аспирантская школа по экономике), преподаватель департамента менеджмента,
Санкт-Петербургская школа экономики и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»,
194100, г. Санкт-Петербург, Кантемировская ул., д. 3, корп., 1, лит. А;

E-mail: aternikov@hse.ru

ORCID: 0000-0003-2354-0109

Александрова Екатерина Александровна

кандидат экономических наук;

директор, Международный центр экономики, управления и политики в области здоровья;
доцент департамента экономики, Санкт-Петербургская школа экономики и менеджмента;
доцент департамента финансов, Санкт-Петербургская школа экономики и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»,
194100, г. Санкт-Петербург, Кантемировская ул., д. 3, корп., 1, лит. А;

E-mail: ea.aleksandrova@hse.ru

ORCID: 0000-0001-7067-5087

Demand for skills on the labor market in the IT sector

Andrei A. Ternikov
E-mail: aternikov@hse.ru

Ekaterina A. Aleksandrova
E-mail: ea.aleksandrova@hse.ru

National Research University Higher School of Economics
Address: 3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia

Abstract

One of the most dynamically changing parts of the labor market relates to information technologies. Skillsets demanded by employers in this sphere vary across different industries, organizations and even certain vacancies. The educational system in the most cases lags behind such changes, so that obsolete skillsets are being taught. This article proposes an algorithm of skillsets identification that allows us to extract skills that are needed by companies from different occupational groups in the information technologies sector. Using the unstructured online-vacancies database for the Russian regional labor market, skills are extracted and unified with the use of TF-IDF and *n*-grams approaches. As a result, key specific skillsets for various occupations are found. The proposed algorithm allows us to identify and standardize key skills which might be applicable to create a system of Russian classification for occupations and skills. In addition, the algorithm allows us to provide lists of the key combinations of skills that are in high demand among companies inside each particular occupation.

Key words: job vacancies in IT sector; online vacancies; unstructured data analysis; labor market; demand on skills of job candidates; combinations of skillsets.

Citation: Ternikov A.A., Aleksandrova E.A. (2020) Demand for skills on the labor market in the IT sector. *Business Informatics*, vol. 14, no 2, pp. 64–83. DOI: 10.17323/2587-814X.2020.2.64.83

References

1. Autor D.H., Levy F., Murnane R.J. (2003) The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, vol. 118, no 4, pp. 1279–1333. DOI: 10.1162/003355303322552801.
2. Bensberg F., Buscher G., Czarnecki C. (2019) Digital transformation and IT topics in the consulting industry: A labor market perspective. *Advances in consulting research: Recent findings and practical cases* (ed. V. Nissen). Cham, Switzerland: Springer, pp. 341–357.
3. Christoforaki M., Ipeirotis P.G. (2015) A system for scalable and reliable technical-skill testing in online labor markets. *Computer Networks*, vol. 90, pp. 110–120. DOI: 10.1016/j.comnet.2015.05.020.
4. Florea R., Stray V. (2018) Software tester, we want to hire you! An analysis of the demand for soft skills. Proceedings of the *19th International Conference on Agile Processes in Software Engineering and Extreme Programming (XP 2018), Porto, Portugal, 21–25 May 2018* (eds. J. Garbajosa, X. Wang, A. Aguiar), pp. 54–67.
5. Goles T., Hawk S., Kaiser K.M. (2008) Information technology workforce skills: The software and IT services provider perspective. *Information Systems Frontiers*, vol. 10, no 2, pp. 179–194.
6. Johnson K.M. (2016) Non-technical skills for IT professionals in the landscape of social media. *American Journal of Business and Management*, vol. 4, no 3, pp. 102–122. DOI: 10.11634/216796061504668.
7. Kappelman L., Jones M.C., Johnson V., McLean E.R., Boonme K. (2016) Skills for success at different stages of an IT professional's career. *Communications of the ACM*, vol. 59, no 8, pp. 64–70. DOI: 10.1145/2888391.
8. Litecky C.R., Arnett K.P., Prabhakar B. (2004) The paradox of soft skills versus technical skills in is hiring. *Journal of Computer Information Systems*, vol. 45, no 1, pp. 69–76.
9. Havelka D., Merhout J.W. (2009) Toward a theory of information technology professional competence. *Journal of Computer Information Systems*, vol. 50, no 2, pp. 106–116.
10. Hussain W., Clear T., MacDonell S. (2017) Emerging trends for global DevOps: A New Zealand perspective. Proceedings of the *IEEE 12th International Conference on Global Software Engineering, Buenos Aires, Argentina, 22–23 May 2017* (ed. R. Bilof), vol. 1, pp. 21–30. DOI: 10.1109/ICGSE.2017.16.
11. Wowczko I. (2015) Skills and vacancy analysis with data mining techniques. *Informatics*, vol. 2, no 4, pp. 31–49. DOI: 10.3390/informatics2040031.
12. Bailey J., Mitchell R.B. (2006) Industry perceptions of the competencies needed by computer programmers: Technical, business, and soft skills. *Journal of Computer Information Systems*, vol. 47, no 2, pp. 28–33.
13. Brooks N.G., Greer T.H., Morris S.A. (2018) Information systems security job advertisement analysis: Skills review and implications for information systems curriculum. *Journal of Education for Business*, vol. 93, no 5, pp. 213–221.
14. Casado-Lumbreras C., Colomo-Palacios R., Soto-Acosta P. (2015) A vision on the evolution of perceptions of professional practice. *International Journal of Human Capital and Information Technology Professionals*, vol. 6, no 2, pp. 65–78. DOI: 10.4018/IJHCITP.2015040105.
15. Föll P., Thiesse F. (2017) Aligning is curriculum with industry skill expectations: A text mining approach. Proceedings of the *25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, 5–10 June 2017* (eds. I. Ramos, V. Tuunainen, H. Krcmar), pp. 2949–2959.
16. Stal J., Paliwoda-Pękosz G. (2019) Fostering development of soft skills in ICT curricula: A case of a transition economy. *Information Technology for Development*, vol. 25, no 2, pp. 250–274. DOI: 10.1080/02681102.2018.1454879.
17. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. (2018) Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, vol. 86, pp. 319–328.
18. Colombo E., Mercorio F., Mezzanzanica M. (2019) AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, no 47, pp. 27–37. DOI: 10.1016/j.infoecopol.2019.05.003.
19. Karakatsanis I., AlKhader W., MacCrory F., Alibasic A., Omar M.A., Aung Z., Woon W.L. (2017) Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, no 65, pp. 1–6. DOI: 10.1016/j.is.2016.10.009.
20. Lovaglio P.G., Cesarini M., Mercorio F., Mezzanzanica M. (2018) Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining*, vol. 11, no 2, pp. 78–91. DOI: doi.org/10.1002/sam.11372
21. Amato F., Boselli R., Cesarini M., Mercorio F., Mezzanzanica M., Moscato V., Picariello A. (2015) Challenge: Processing web texts or classifying job offers. Proceedings of the *2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, California, USA, 7–9 February 2015* (eds. M.S. Kankanhalli, T. Li, W. Wang), pp. 460–463.
22. De Mauro A., Greco M., Grimaldi M., Ritala P. (2018) Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, vol. 54, no 5, pp. 807–817. DOI: 10.1016/j.ipm.2017.05.004.
23. Gurcan F., Cagiltay N.E. (2019) Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access*, no 7, pp. 82541–82552.
24. Pejic-Bach M., Bertoncel T., Meško M., Krstić Ž. (2020) Text mining of industry 4.0 job advertisements. *International Journal of Information Management*, no 50, pp. 416–431.
25. Radovilsky Z., Hegde V., Acharya A., Uma U. (2018) Skills requirements of business data analytics and data science jobs: A comparative analysis. *Journal of Supply Chain and Operations Management*, vol. 16, no 1, pp. 82–101.

About the authors

Andrei A. Ternikov

Doctoral Student, Doctoral School on Economics;

Lecturer, Department of Management, St. Petersburg School of Economics and Management,
National Research University Higher School of Economics,
3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia;

E-mail: aternikov@hse.ru

ORCID: 0000-0003-2354-0109

Ekaterina A. Aleksandrova

Cand. Sci. (Econ.);

Director, International Centre for Health Economics, Management, and Policy;
Associate Professor, Department of Economics, St.Petersburg School of Economics and Management;
Associate Professor, Department of Finance, St.Petersburg School of Economics and Management,
National Research University Higher School of Economics,
3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia;

E-mail: ea.aleksandrova@hse.ru

ORCID: 0000-0001-7067-5087

DOI: 10.17323/2587-814X.2020.2.84.92

О возможности определения префикса и суффикса слова по подсловам фиксированной длины

Г.Н. Жукова[a]
E-mail: galinanzhukova@gmail.com

Ю.Г. Сметанин[b]
E-mail: smetanin.iury2011@yandex.ru

М.В. Ульянов[c]
E-mail: muljanov@mail.ru

[a] Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

[b] Федеральный исследовательский центр «Информатика и управление» Российской академии наук
Адрес: 119333, г. Москва, ул. Вавилова, д. 40

[c] Институт проблем управления им. В.А. Трапезникова Российской академии наук
Адрес: 117997, г. Москва, ул. Профсоюзная, д. 65

Аннотация

В прикладных задачах бизнес-информатики, связанных с анализом данных (в частности, при анализе и прогнозировании временных рядов, при исследовании лог-файлов бизнес-процессов) возникают задачи качественного анализа. Методы качественного анализа достаточно часто используют символьное кодирование как способ представления информации об исследуемых процессах. В ряде ситуаций, обусловленных фрагментарностью таких описаний, возникает задача реконструкции полного символьного описания процесса (слова) по его последовательным фрагментам (подсловам). По мультимножеству всех подслов достаточно большой длины исходное слово восстанавливается однозначно. В случае недостаточно длинных подслов возможно множество различных реконструкций исходного неизвестного слова. Число допустимых реконструкций можно сократить, если определить суффикс и префикс реконструируемого слова. Предложен метод определения префикса и суффикса слова над конечным алфавитом, состоящих из $k-1$ символов каждый, на основании мультимножества V подслов фиксированной длины, равной k. Принимается гипотеза о том, что это мультимножество порождено смещением на один символ окна фиксированной длины k по неизвестному слову. Метод определения префикса и суффикса основан на построении и анализе матрицы, образованной записанными по строкам в произвольном порядке подсловам из V и использовании оператора, действующего на мультимножества символов алфавита, образованных соседними столбцами этой матрицы. Метод позволяет определить префикс $a_1a_2 \ldots a_{k-1}$ и суффикс $b_1b_2 \ldots b_{k-1}$ неизвестного слова в случае, если $a_i \neq b_i$ для любых i от 1 до $k-1$. В случае, если $a_i \neq b_i$ только для некоторых значений i, в префиксе и суффиксе определяются символы в соответствующих позициях, а для остальных символов выполняется условие $a_j = b_j$. В худшем случае метод констатирует, что $a_i = b_i$ для всех i от 1 до $k-1$, но не определяет сами символы. Это ситуация, при которой префикс и суффикс совпадают, но не могут быть определены.

Ключевые слова: реконструкция слова; префикс; суффикс; мультимножество подслов; подслова фиксированной длины; оператор сдвига.

Цитирование: Жукова Г.Н., Сметанин Ю.Г., Ульянов М.В. О возможности определения префикса и суффикса слова по подсловам фиксированной длины // Бизнес-информатика. 2020. Т. 14. № 2. С. 84–92.
DOI: 10.17323/2587-814X.2020.2.84.92

Введение

В прикладных областях бизнес-информатики, связанных с анализом данных, таких как анализ и прогнозирование временных рядов [1–6], исследование лог-файлов бизнес-процессов [7] и др. возникают задачи качественного анализа. В этом случае одним из часто используемых способов представления информации о процессах является символьное кодирование [8]. При этом описание поведения временного ряда или бизнес-процесса кодируется словом над конечным алфавитом, которое и является объектом дальнейшего исследования. Однако в ряде случаев, в том числе при анализе бизнес-процессов и временных рядов, исследователи получают не само слово целиком, а множество подслов, которые являются последовательными фрагментами некоторого слова. Поскольку при этом позиции подслов в исходном слове неизвестны, возникает задача реконструкции – восстановления неизвестного слова по исходному множеству подслов [9–17]. Эта задача содержательно относится к специальному разделу дискретной математики – комбинаторике слов [18]. Объектами исследования в комбинаторике слов являются слова над произвольными алфавитами, а предметом исследований – изучение комбинаторных свойств различных множеств слов, как конечных, так и бесконечных. В реальных прикладных задачах информация о словах часто оказывается неполной. Например, такая ситуация неизбежна при анализе бесконечных временных рядов, измеряемых на протяжении конечных интервалов времени.

Заметим, что одной из важных областей практического применения методов комбинаторики слов является область биомолекулярных моделей и процессов. При этом работа с фрагментарной информацией характерна для ряда задач биоинформатики и геномики. Например, задача секвенирования геномов [19, 20] по сути является задачей реконструкции слов в условиях сильных ограничений, подразумевающей однозначность реконструкции.

Задачи восстановления слов над конечным алфавитом имеют различные постановки, отличающиеся как объемом имеющейся информацией, так и ограничениями на допустимые решения [21–23]. Обычно эти задачи, как задачи с неполной информацией, являются сложными, и получение какой-либо дополнительной информации, очевидно, позволяет сократить рассматриваемое множество возможных решений.

При качественном анализе временных рядов [24, 25] кодирование значений наблюдаемой величины может осуществляться в некотором алфавите, например, (A, B, C, D, E, F), символами которого могут быть именованы полусегменты значений наблюдаемой величины в порядке их возрастания: A – имя полусегмента наименьших значений, F – наибольших. Поскольку фиксация наблюдений ведется в дискретном времени, описание значений временного ряда по именам полусегментов есть слово над алфавитом имен. Если наблюдаемый процесс характеризуется резкими выбросами значений наблюдаемой величины (до уровня F) относительно базального уровня (A, B) за один дискрет времени, равно как и резкими спадами (от F до B), то получаемые кодовые слова временного ряда не будут содержать подслов CDE и EDC. Если при этом исходные данные представляют собой подслова – разрозненные фрагменты наблюдений, то задача реконструкции слова по подсловам есть задача восстановления всего описания временного ряда в предположении об особенностях его поведения.

Аналогичная ситуация возникает при реконструкции лог-файлов бизнес-процессов при наличии фрагментарной информации. При описании бизнес-процессов аппаратом теории графов [7] модель (граф бизнес-процесса) может быть представлена следующим образом: состояния процесса кодируются именованными вершинами, а переходы состояний – ребрами, отождествленными с этапами бизнес-процесса. Тогда запись конкретной реализации бизнес-процесса есть некоторое слово над алфавитом имен вершин, отражающее порядок перехода состояний. Если процесс физически

распределен между различными организациями и исполнителями, то, скорее всего, мы получим информацию о его полном прохождении в виде набора подслов. При этом запрещенные подслова могут быть интерпретированы как нарушения модели – регламента бизнес-процесса. Возникающая задача реконструкции без запрещенных подслов содержательно означает возможность полной реконструкции всего процесса, соответствующего теоретической модели.

Таким образом, представляет интерес подробное изучение различных вариантов задачи реконструкции слов по некоторому множеству подслов меньшей длины, интерпретируемых как множество последовательных фрагментов неизвестного слова. При этом интерес представляет как случай, когда реконструируемое слово не содержит заранее заданного запрещенного подслова, так и случай с наличием запрещенных подслов. Один из возможных вариантов решения этой задачи на основе подслов фиксированной длины в гипотезе сдвига один предложен в работах [26, 27]. Однако множество возможных реконструкций может быть достаточно велико и возникает задача о возможном сокращении числа претендентов на «правильное» реконструируемое слово. Мы хотим получить дополнительную информацию из исходного множества подслов, которая будет полезна при редукции полученного множества реконструкций. Речь идет о возможности восстановления и/или определения шаблона префикса и суффикса неизвестного слова, что в рамках процедуры редукции приведет к рассмотрению только тех слов, которые обладают полученными шаблонами префикса и суффикса. Именно эта задача и является предметом настоящей статьи.

1. Терминология и обозначения

Далее в тексте статьи будут использоваться следующие обозначения:

$\Sigma = \{s_1, s_2, \ldots s_l\}$ – алфавит, s_i – i-ый символ алфавита;

Σ^k – k-я декартова степень множества Σ (множество k-элементных кортежей);

$\Sigma^* = \bigcup_{k=0}^{\infty} \Sigma^k$ – транзитивное замыкание Σ (множество всех возможных кортежей);

w – слово (над алфавитом) – последовательность символов алфавита, при этом собственно символы алфавита есть слова по определению;

$L(\cdot) : L(C) = W$, где $C \subseteq \Sigma^*$ – множество кортежей, W – множество слов. Оператор $L(\cdot)$ есть оператор создания множества слов, состоящих из символов алфавита Σ, действующий на множество кортежей;

a_i – i-ый символ слова w, $a_i \in \Sigma$;

$w = a_1 a_2 \ldots a_n \in L\left(\Sigma^n\right)$ – произвольное слово из n символов над алфавитом Σ;

$|w| = n$ – длина слова, определяемая как число элементов в порождающем кортеже;

$L_k = L\left(\Sigma^k\right) = \left\{w \,|\, |w| = k\right\}$ – множество всех слов длины k над алфавитом Σ.

Пусть $w = a_1 a_2 \ldots a_n \in L\left(\Sigma^n\right)$, тогда при $k < n$:

$v = a_{i_1} a_{i_2} \ldots a_{i_k}, 1 \le i_1, i_2 = i_1 + 1, i_k = i_{k-1} + 1 \le n$ – подслово слова w длины k;

$Q(w, i, k)$ – оператор выделения подслова длины k в слове w, начиная с символа в позиции i. Пусть $|w| = n$, тогда оператор определен при $i + k - 1 \le n$

$$Q(a_1 a_2 \ldots a_n, i, k) = a_i a_{i+1} \ldots a_{i+k-1},$$

$$Q(w, i, k) \in L_k;$$

Для следующих двух операторов полагаем, что $|w| = n \ge 2$ и $1 \le k < n$:

$P(w, k) = Q(w, 1, k) = a_1 a_2 \ldots a_k \in L_k$ – префикс длины k слова w;

$S(w, k) = Q(w, n - k + 1, k) = a_{n-k+1} \ldots a_n \in L_k$ – суффикс длины k слова w;

$SH1(w, k)$ – оператор сдвига один. Определенный при $|w| > k$ оператор порождает мультимножество подслов длины k мощности $|w| > k + 1$, выполняя сдвиг на единицу окна длины k по слову w, начиная с крайней левой позиции слова w:

$SH1(w, k) = \{v_j | j = 1, |w| - k + 1; v_j = Q(w, i, k)\}$.

2. Постановка задачи

В дальнейшем мы считаем заданными: длину подслова – k, число подслов – m, а также исходное мультимножество подслов V над алфавитом Σ, рассматриваемое как базис реконструкции некоторого неизвестного слова w:

$$V = \left\{v_i \,|\, i = \overline{1,m}; v_i = a_{i1} a_{i2} \ldots a_{ik} \in L_k\right\}.$$

Принимаемая авторами гипотеза сдвига один состоит в том, что мы рассматриваем V как мультимножество подслов сдвига один относительного некоторого неизвестного слова w, при этом $|w| = n = m + k - 1$:

$$V = SH1(w,k) = \left\{ v_j \mid j = 1, n-k+1; v_j = Q(w,j,k) \right\}.$$

Содержательная постановка: В условиях гипотезы сдвига один относительно мультимножества V возможно ли определить префикс и суффикс длины $k-1$ неизвестного слова w, или получить какую-либо содержательную информацию о его префиксе и суффиксе?

Математическая постановка: По данному мультимножеству V с длиной подслова k и числом подслов m определить префикс $P(w, k-1)$ и суффикс $S(w, k-1)$ длины $k-1$ исходного слова $w = a_1 a_2 ... a_n$, а также указать условия, при которых решение возможно.

3. Метод определения префикса и суффикса

Предварительно отметим, что основная проблема (и в аспекте задачи реконструкции, и в аспекте задачи определения суффикса и префикса) заключается в том, что нам исходно дано мультимножество подслов V, а не кортеж подслов. При этом основная трудность связана именно с потерей порядка на исходных подсловах, полученных оператором сдвига один.

Решение поставленной задачи начнем с построения матрицы A, состоящей из m строк и k столбцов, строками которой являются исходные слова v_i из множества V. Слова из множества V представимы в виде $v_i = a_{i1}, a_{i2}, ... a_{ik}$, и элементами матрицы A являются символы алфавита $\sum - A = (a_{ij})$, где a_{ij} – символ алфавита на j-й позиции в i-м слове мультимножества V в порядке их перечисления.

Запишем явно матрицу A в прямой последовательности окна сдвига один. Очевидно, что в реальности в порядке перечисления по мультимножеству V мы будем наблюдать некоторую перестановку слов прямой последовательности, и, следовательно, соответственную перестановку строк матрицы A:

$$A = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{m-1} \\ v_m \end{pmatrix} = \begin{pmatrix} a_1, a_2, ..., a_k \\ a_2, a_3, ..., a_{k+1} \\ a_3, a_4, ..., a_{k+2} \\ \vdots \\ a_{n-k}, a_{n-k+1}, ..., a_{n-1} \\ a_{n-k+1}, a_{n-k+2}, ..., a_n \end{pmatrix}.$$

Содержательно решение поставленной задачи опирается на анализ соседних столбцов этой матрицы. Рассмотрим первый и второй столбцы. В каждом из них при любой перестановке строк будет символ, стоящий на втором месте в неизвестном слове $w - a_2$, и символ, стоящий на третьем месте – a_3, и т.д. Если из этих двух столбцов вычеркнуть совпадающие пары символов, то останутся только символы a_1 и a_{n-k+2}. При условии, что они различны, мы получаем их конкретные значения. Если же a_1 и a_{n-k+2} совпадают, то будут вычеркнуты все символы в этих столбцах, и мы получаем информацию о том, что в соответствующих позициях префикса и суффикса находятся неизвестные, но совпадающие символы. Такой анализ может быть продолжен для всех $k-1$ пар соседних столбцов матрицы A. При условии, что после вычеркивания пар совпадающих символов у нас всегда остается не совпадающая пара, мы восстанавливаем префикс и суффикс длины $k-1$ неизвестного слова w.

Опишем метод формально.

Введем в рассмотрение кортеж всех символов алфавита, для которого разрешены кратности элементов

$$C = \left(s_1^{(\alpha 1)}, s_2^{(\alpha 2)}, ..., s_l^{(\alpha l)} \right),$$

при этом кратность 0 приводит к пустому множеству в данной позиции $s_i^{(0)} = \varnothing$. Определим оператор G действующий на i-й столбец матрицы A и создающий кортеж C_i, содержащий для всех символов алфавита их кратности в соответствии с числом символов, находящихся в этом столбце

$$GC(A, i) = C_i = \left(s_1^{(\alpha 1)}, s_2^{(\alpha 2)}, ..., s_l^{(\alpha l)} \right).$$

Применим оператор G к двум столбцам матрицы A, и обозначим:

$$GC(A, i) = C_i = \left(s_1^{(\alpha 1)}, s_2^{(\alpha 2)}, ..., s_l^{(\alpha l)} \right),$$
$$GC(A, k) = C_k = \left(s_1^{(\beta 1)}, s_2^{(\beta 2)}, ..., s_l^{(\beta l)} \right).$$

Введем в рассмотрение оператор получения символа GS, действующий на два кортежа столбцов матрицы A по следующему правилу:

$$GS(A,i,k) = \begin{cases} \bigcup_{j=1}^{l} s_j^{(\alpha_j - \beta_j)}, \left(s_j^{(\alpha_j)} \in GC(A,i) \right), \\ \qquad s_j^{(\beta_j)} \in GC(a,k), \\ s_j^{(\alpha j - \beta_j)} = \varnothing, \text{если } a_j - \beta_j \le 0. \end{cases}$$

Теперь применим оператор GS к двум последовательным столбцам матрицы A. В силу описанной выше структуры последовательных столбцов матрицы A результатом оператора GS будет

или символ, или пустое множество. Отметим, что если $GS(A, i, i+1) \neq \varnothing$, то и $GS(A, i+1, i) \neq \varnothing$. В этом случае мы определяем i-й символ префикса $a_i = GS(A, i, i+1)$ и $n-k+i$-й символ неизвестного слова $a_{n-k+i} = GS(A, i+1, i)$, который является i-м символом суффикса длины $k-1$.

Например, если $GS(A, 1, 2) = s_i$, то нам становится известен первый символ неизвестного слова w (первый символ префикса) – $a_1 = s_i$. В этой ситуации значение $GS(A, 2, 1)$ обязательно не пусто. Пусть $GS(A, 2, 1) = s_j$, в результате мы получаем первый символ суффикса $a_{n-k+2} = s_j$. Если же $GS(A, 1, 2) \neq \varnothing$, то очевидно, что и $GS(A, 2, 1) = \varnothing$, и мы получаем информацию о том, что $a_1 = a_{n-k+2}$. Однако при этом сам символ алфавита на этих позициях остается нам неизвестен.

Поскольку мы имеем $k-1$ последовательных пар столбцов, то если для каждой последовательной пары столбцов оператор GS возвращает непустое множество, то используя операцию «+» для обозначения конкатенации символов, мы получаем решение:

$$P(w, k-1) = a_1 a_2 \ldots a_{k-1} = \sum_{i=1}^{k-1} GS(A, i, i+1),$$

$$S(w, k-1) = a_{n-k+2} \ldots a_n = \sum_{i=1}^{k-1} GS(A, i+1, i).$$

Если для каждой пары оператор GS возвращает пустое множество, то символы префикса и суффикса остаются неизвестными, но при этом мы получаем информацию об их равенстве как подслов:

$$P(w, k-1) = S(w, k-1).$$

В общем случае мы получаем информацию о символах префикса и суффикса в виде некоторого шаблона, причем если это конкретные символы, то они расположены в одинаковых позициях префикса и суффикса, а если символы не удается определить, то у нас есть информация о том, что на этих позициях символы префикса и суффикса совпадают.

Приведем пример для слова $w = abbaaabb$ в алфавите $\sum = \{a, b\}$ и множества подслов, полученных оператором сдвига один с шириной окна, равной трем. При этом $k = 3$, $m = 6$, $n = 8$, и матрица A имеет вид:

$$A = \begin{pmatrix} abb \\ bba \\ baa \\ aaa \\ aab \\ abb \end{pmatrix}.$$

Применение оператора G к трем столбцам матрицы A дает следующие кортежи:

$$GC(A, 1) = C_1 = (a^{(4)}, b^{(2)}),$$

$$GC(A, 2) = C_2 = (a^{(3)}, b^{(3)}),$$

$$GC(A, 3) = C_3 = (a^{(3)}, b^{(3)}).$$

В результате мы получаем $GC(A, 1, 2) = a$, $GC(A, 2, 1) = b$, и $GC(A, 2, 3) = GC(A, 3, 2) = \varnothing$, и, тем самым, шаблоны префикса слова $w = abbaaabb$ длины два $P(w, 2) = a*$ и суффикса $S(w, 2) = b*$, где символ $*$ обозначает неизвестный, но совпадающий символ в соответствующих позициях префикса и суффикса (на самом деле это символ «b»).

4. Применение к задаче реконструкции

В одной из предыдущих статей [26] авторы предложили решение задачи о полной реконструкции в условиях мультимножества подслов и гипотезы сдвига один. В ряде случаев число реконструкций, определяемых числом эйлеровых путей или циклов в соответствующем мультиорграфе де Брейна, может быть значительным [26].

Введем в рассмотрение множество возможных реконструкций слов по исходному множеству V

$$W = \left\{(w| \ |w| = m, k-1 = n, V = SH1(w, k)\right\},$$

при этом если $|W| \geq 2$, то реконструкция возможна и многозначна. Пусть $w*$ – исходное, но неизвестное нам слово, по которому получено множество $V = SH1(w*, k)$. Тогда при выборе из возможных реконструкций (т.е. из множества W) мы выбираем только те слова, которые обладают полученным оператором GS префиксом и суффиксом, с учетом шаблонов неизвестных символов:

$$\tilde{W} = \left\{ \begin{array}{l} (w| \ P(w, k-1) = \sum_{i=1}^{k-1} GS(A, i, i+1), \\ S(w, k-1) = \sum_{i=1}^{k-1} GS(A, i+1, i) \end{array} \right\},$$

при этом гарантированно $w* \in \tilde{W}$.

Это приводит к редукции полученного множества реконструкций, поскольку мы рассматриваем только те слова, которые обладают заданными шаблонами префикса и суффикса. Более того, этот подход можно применять не только для редукции

конечного множества реконструкций, а рассматривать префикс как шаблон выбора начальных дуг для эйлеровых путей в мультиорграфе де Брейна при построении реконструкции [26].

Заключение

В статье в аспекте решения задачи восстановления символьных описаний временных рядов и логов бизнес-процессов предложено решение задачи определения префикса и суффикса неизвестного слова. Решение основано на предположении о том, что исходно задано полное множество подслов фиксированной длины k, порожденное смещением окна длины k по неизвестному слову со сдвигом один. Получено решение, позволяющее получить информацию о префиксе и суффиксе неизвестного слова или некоторый шаблон для префикса и суффикса. Предложенное решение позволяет получить дополнительную информацию о возможных реконструкциях и тем самым сократить число возможных реконструкций слов по заданному множеству подслов. В лучшем случае предложенный метод позволяет определить префикс и суффикс длины k неизвестного слова, а в худшем случае – констатировать, что префикс и суффикс совпадают между собой.

Результаты могут быть использованы совместно с решением задачи реконструкции [26, 27] для редукции множества возможных реконструкций при качественном анализе в таких задачах бизнес-информатики, как анализ временных рядов и логов бизнес-процессов. ■

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00150.

Литература

1. Querying and mining of time series data: Experimental comparison of representations and distance measures / H. Ding [et al.] // Proceedings of the VLDB Endowment. 2008. Vol. 1. No 2. P. 1542–1552. DOI: 10.14778/1454159.1454226.
2. Kurbalija V., Radovanović M., Geler Z., Ivanović M. The influence of global constraints on DTW and LCS similarity measures for time-series databases // Advances in Intelligent and Soft Computing. 2011. Vol. 101. P. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
3. Wu Y.-L., Agrawal D., el Abbadi A. A comparison of DFT and DWT based similarity search in time-series databases // Ninth International Conference on Information and Knowledge Management (CIKM '00), McLean, VA, 6–11 November 2000. P. 488–495.
4. Bemdt D.J., Clifford J. Using dynamic time warping to find patterns in time series // AAAI-94 Workshop on Knowledge Discovery in Databases. 1994. P. 359–370. [Электронный ресурс]: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (дата обращения 15.03.2020).
5. Dreyer W., Dittrich A.K., Schmidt D. Research perspectives for time series management systems // SIGMOD Record. 1994. Vol. 23. No 1. P. 10–15.
6. Keogh E.J., Pazzani M.J. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback // Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, 27–31 August 1998. P. 239–241.
7. Andersen B. Business processes improvement toolbox. New York: ASQ Quality Press, 1999.
8. Lin J., Keogh E., Wei L., Lonardi S. Experiencing SAX: A novel symbolic representation of time series // Data Mining and Knowledge Discovery. 2007. Vol. 15. No 2. P. 107–144. DOI: 10.1007/s10618-007-0064-z.
9. String reconstruction from substring compositions / J. Acharya [et al.] // SIAM Journal on Discrete Mathematics. 2014. Vol. 29. No 3. P. 1340–1371.
10. Reconstruction of sequences / B. Manvel [et al.] // Discrete Mathematics. 1991. Vol. 94. No 3. P. 209–219. DOI: 10.1016/0012-365X(91)90026-X.
11. Carpi A., de Luca A. Words and special factors // Theoretical Computer Science. 2001. Vol. 259. No 1–2. P. 145–182.
12. de Luca A. On the combinatorics of finite words // Theoretical Computer Science. 1999. Vol. 218. No 1. P. 13–39.
13. Dudík M., Schulman L.J. Reconstruction from subsequences // Journal of Combinatorial Theory. Series A. 2003. Vol. 103. No 2. P. 337–348. DOI: 10.1016/S0097-3165(03)00103-1.
14. Erdös P.L., Ligeti P., Sziklai P., Torney D.C. Subwords in reverse-complement order // Annals of Combinatorics. 2006. Vol. 10. No 4. P. 415–430. DOI: 10.1007/s00026-006-0297-3.
15. Fici G., Mignosi F., Restivo A., Sciortino M. Word assembly through minimal forbidden words // Theoretical Computer Science. 2006. Vol. 359. No 1–3. P. 214–230. DOI: 10.1016/j.tcs.2006.03.006.
16. Levenshtein V.I. Efficient reconstruction of sequences from their subsequences or supersequences // Journal of Combinatorial Theory, Series A. 2001. Vol. 93. P. 310–332.
17. Piña C., Uzcátegui C. Reconstruction of a word from a multiset of its factors // Theoretical Computer Science. 2008. Vol. 400. No 1–3. P. 70–83. DOI: 10.1016/j.tcs.2008.01.052.

18. Lothaire M. Algebraic combinatorics on words. Cambridge, UK: Cambridge University Press, 2002.

19. Gusfield D. Algorithms on strings, trees, and sequences: Computer science and computational biology. Cambridge, UK: Cambridge University Press, 1997.

20. Skiena S.S., Sundaram G. Reconstructing strings from substrings // Journal of Computational Biology. 1995. Vol. 2. No 2. P. 333–353.

21. Leont'ev V.K., Smetanin Y.G. Problems of Information on the set of words // Journal of Mathematical Sciences. 2002. Vol. 108. No 1. P. 49–70. DOI: 10.1023/A:1012705332306.

22. Левенштейн В.И. Восстановление объектов по минимальному числу искаженных образцов // Доклады РАН. 1997. Т. 354. № 5. С. 593–596.

23. Krasikov I., Roditty Y. Note: On a reconstruction problem for sequences // Journal of Combinatorial Theory. Series A. 1997. No 77. P. 344–348.

24. Ульянов М.В., Сметанин Ю.Г. Подход к определению характеристик колмогоровской сложности временных рядов на основе символьных описаний // Бизнес-информатика. 2013. № 2. С. 49–54.

25. Сметанин Ю.Г., Ульянов М.В. Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика. 2014. № 3. С. 40–46.

26. Smetanin Yu.G., Ulyanov M.V. Reconstruction of a word from a finite set of its subwords under the unit Shift hypothesis. I. Reconstruction without for bidden words // Cybernetics and Systems Analysis. 2014. Vol. 50. No 1. P. 148–156.

27. Smetanin Yu.G., Ulyanov M.V. Reconstruction of a word from a finite set of its subwords under the unit Shift hypothesis. II. Reconstruction with forbidden words // Cybernetics and Systems Analysis. 2015. Vol. 51. No 1. P. 157–164. DOI: 10.1007/s10559-015-9708-y.

Об авторах

Жукова Галина Николаевна

кандидат физико-математических наук;

доцент департамента программной инженерии, факультет компьютерных наук,
Национальный исследовательский университет «Высшая школа экономики»,
101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Сметанин Юрий Геннадиевич

доктор физико-математических наук:

главный научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук,
119333, г. Москва, ул. Вавилова, д. 40;

E-mail: smetanin.iury2011@yandex.ru

ORCID: 0000-0003-0242-6972

Ульянов Михаил Васильевич

доктор технических наук, профессор;

ведущий научный сотрудник, Институт проблем управления им. В.А. Трапезникова Российской академии наук,
117997, г. Москва, ул. Профсоюзная, д. 65;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836

About the possibility of determining the prefix and suffix of a word by subwords of fixed length

Galina N. Zhukova[a]
E-mail: galinanzhukova@gmail.com

Yuri G. Smetanin[b]
E-mail: smetanin.iury2011@yandex.ru

Mikhail V. Ulyanov[c]
E-mail: muljanov@mail.ru

[a] National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia
[b] Federal Research Center "Computer Science and Control", Russian Academy of Sciences
Address: 40, Vavilova Street, Moscow 119333, Russia
[c] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
Address: 65, Profsoyuznaya Street, Moscow 117997, Russia

Abstract

In applied problems of business informatics related to data analysis (in particular, in the analysis and forecasting of time series, in the study of log files of business processes, etc.), problems of qualitative analysis arise. Qualitative analysis methods often use symbolic coding as a way of presenting information about the processes under study. In a number of situations, due to the fragmentation of such descriptions, the problem arises of reconstructing a complete symbolic description of a process (word) from its successive fragments (subwords). From the multiset of all subwords of a sufficiently large length, the original word is uniquely restored. In the case of insufficiently long subwords, several different reconstructions of the original word are possible. The number of feasible reconstructions can be reduced by determining the suffix and prefix of the reconstructed word. A method is proposed for determining the prefix and suffix of a word consisting of $k-1$ symbols each on the basis of multiset V of subwords of a fixed length equal to k. We accept the hypothesis that this multiset is generated by a window of a fixed length k of one symbol shift in an unknown word. The method for determining the prefix and suffix is based on the construction and analysis of the matrix formed by subwords from V written in rows in arbitrary order and the use of the operator acting on multisets of characters of the alphabet formed by neighboring columns of this matrix. The method is capable of determining the prefix $a_1a_2 \ldots a_{k-1}$ and suffix $b_1b_2 \ldots b_{k-1}$, if $a_i \neq b_i$ for any i from 1 to $k-1$. If in the prefix and suffix $a_i \neq b_i$ only for some values of i, the characters in the corresponding positions are determined, and $a_j = b_j$ for the remaining characters. In the worst case, the method concludes that $a_i = b_i$ for any i from 1 to $k-1$, but does not determine the characters themselves. This is a situation in which the prefix and suffix coincide but cannot be determined.

Key words: word reconstruction; prefix; suffix; multiset of subwords; subwords of fixed length; shift operator.

Citation: Zhukova G.N., Smetanin Yu.G., Ulyanov M.Yu. (2020) About the possibility of determining the prefix and suffix of a word by subwords of fixed length. *Business Informatics*, vol. 14, no 2, pp. 84–92.
DOI: 10.17323/2587-814X.2020.2.84.92

References

1. Ding H., Trajcevski G., Scheuermann P., Wang X., Keogh E. (2008) Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, vol. 1, no 2, pp. 1542–1552. DOI: 10.14778/1454159.1454226.
2. Kurbalija V., Radovanović M., Geler Z., Ivanović M. (2011) The influence of global constraints on DTW and LCS similarity measures for time-series databases. *Advances in Intelligent and Soft Computing*, vol. 101, pp. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
3. Wu Y.-L., Agrawal D., el Abbadi A. (2000) A comparison of DFT and DWT based similarity search in time-series databases. Proceedings of the *Ninth International Conference on Information and Knowledge Management (CIKM '00), McLean, VA, 6–11 November 2000*, pp. 488–495.
4. Bemdt D.J., Clifford J. (1994) Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 359–370. Available at: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (accessed 15 March 2020).
5. Dreyer W., Dittrich A.K., Schmidt D. (1994) Research perspectives for time series management systems. *SIGMOD Record*, vol. 23, no 1, pp. 10–15.
6. Keogh E.J., Pazzani M.J. (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Proceedings of the *Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, 27–31 August 1998*, pp. 239–241.
7. Andersen B. (1999) *Business processes improvement toolbox*. New York: ASQ Quality Press.
8. Lin J., Keogh E., Wei L., Lonardi S. (2007) Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, vol. 15, no 2, pp. 107–144. DOI: 10.1007/s10618-007-0064-z.
9. Acharya J., Das H., Milenkovic O., Orlitsky A., Pan S. (2014) String reconstruction from substring compositions. *SIAM Journal on Discrete Mathematics*, vol. 29, no 3, pp. 1340–1371.
10. Manvel B., Meyerowitz A., Schwenk A., Smith K., Stockmeyer P. (1991) Reconstruction of sequences. *Discrete Mathematics*, vol. 94, no 3, pp. 209–219. DOI: 10.1016/0012-365X(91)90026-X.
11. Carpi A., de Luca A. (2001) Words and special factors. *Theoretical Computer Science*, vol. 259, no 1–2, pp. 145–182.
12. de Luca A. (1999) On the combinatorics of finite words. *Theoretical Computer Science*, vol. 218, no 1, pp. 13–39.

13. Dudík M., Schulman L.J. (2003) Reconstruction from subsequences. *Journal of Combinatorial Theory. Series A*, vol. 103, no 2, pp. 337–348. DOI: 10.1016/S0097-3165(03)00103-1.

14. Erdös P.L., Ligeti P., Sziklai P., Torney D.C. (2006) Subwords in reverse-complement order. *Annals of Combinatorics*, vol. 10, no 4, pp. 415–430. DOI: 10.1007/s00026-006-0297-3.

15. Fici G., Mignosi F., Restivo A., Sciortino M. (2006) Word assembly through minimal forbidden words. *Theoretical Computer Science*, vol. 359, no 1–3, pp. 214–230. DOI: 10.1016/j.tcs.2006.03.006.

16. Levenshtein V.I. (2001) Efficient reconstruction of sequences from their subsequences or supersequences. *Journal of Combinatorial Theory, Series* A, Vol. 93, pp. 310–332.

17. Piña C., Uzcátegui C. (2008) Reconstruction of a word from a multiset of its factors. *Theoretical Computer Science*, vol. 400, no 1–3, pp. 70–83. DOI: 10.1016/j.tcs.2008.01.052.

18. Lothaire M. (2002) *Algebraic combinatorics on words*. Cambridge, UK: Cambridge University Press.

19. Gusfield D. (1997) *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge, UK: Cambridge University Press.

20. Skiena S.S., Sundaram G. (1995) Reconstructing strings from substrings. *Journal of Computational Biology*, vol. 2, no 2, pp. 333–353.

21. Leont'ev V.K., Smetanin Y.G. (2002) Problems of Information on the set of words. *Journal of Mathematical Sciences*, vol. 108, no 1, pp. 49–70. DOI: 10.1023/A:1012705332306.

22. Levenshtein V.I. (1997) Restoring objects based on the minimum number of distorted samples. *Doklady Akademii Nauk*, vol. 354, no 5, pp. 593–596 (in Russian).

23. Krasikov I., Roditty Y. (1997) Note: On a reconstruction problem for sequences. *Journal of Combinatorial Theory, Series A*, no 77, pp. 344–348.

24. Ulyanov M.V., Smetanin Yu.G. (2013) Determining the characteristics of Kolmogorov complexity of time series: An approach based on symbolic descriptions. *Business Informatics*, no 2, pp. 49–54 (in Russian).

25. Smetanin Yu.G., Ulyanov M.V. (2014) Measure of symbolical diversity: Combinatorics on words as an approach to identify generalized characteristics of time series. *Business Informatics*, no 3, pp. 40–46 (in Russian).

26. Smetanin Yu.G., Ulyanov M.V. (2014) Reconstruction of a word from a finite set of its subwords under the unit Shift hypothesis. I. Reconstruction without for bidden words. *Cybernetics and Systems Analysis*, vol. 50, no 1, pp. 148–156.

27. Smetanin Yu.G., Ulyanov M.V. (2015) Reconstruction of a word from a finite set of its subwords under the unit Shift hypothesis. II. Reconstruction with forbidden words. *Cybernetics and Systems Analysis*, vol. 51, no 1, pp. 157–164. DOI: 10.1007/s10559-015-9708-y.

About the authors

Galina N. Zhukova

Cand. Sci. (Phys.-Math.);

Associate Professor, School of Software Engineering, Faculty of Computer Science,
National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Yuri G. Smetanin

Dr. Sci. (Phys.-Math.);

Chief Researcher, Federal Research Center "Computer Science and Control", Russian Academy of Sciences,
40, Vavilova Street, Moscow 119333, Russia;

E-mail: smetanin.iury2011@yandex.ru

ORCID: 0000-0003-0242-6972

Mikhail V. Ulyanov

Dr. Sci. (Tech.);

Leading Researcher, Laboratory of Scheduling Theory and Discrete Optimization, V.A. Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences,
65, Profsoyuznaya Street, Moscow 117997, Russia;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836