Рубрика 4. ЭКОНОМИКА ТРАНСПОРТА

УДК [UDC] 338.47
DOI 10.17816/transsyst202062106-115

© А. Л. Каменков
Петербургский государственный университет путей сообщения
Императора Александра I
(Санкт-Петербург, Россия)

ПРИМЕНЕНИЕ БОЛЬШИХ ДАННЫХ ДЛЯ АНАЛИЗА ПАССАЖИРОПОТОКА НА СКОРОСТНЫХ МАГИСТРАЛЯХ РОССИЙСКОЙ ФЕДЕРАЦИИ

Обоснование: Данные, которые объединяют в себе техники и технологии, которые извлекают смысл из информации на экстремальном пределе практичности, называются «большими данными». Большие данные способствовали углублению понимания поведения особенностей многих систем, а, в частности, транспортной системы. Большие данные из различных источников с различным диапазоном масштабов данных становятся все более доступными для общества. Однако, в рамках железнодорожного транспорта, получить большие данные релевантные для пользователей, которыми являются, как частные лица, так и компании, – сложная задача, а инструменты сбора данных дефицитны и их отсутствие является препятствием для углубленного изучения закономерностей мобильности человека. Информационной основой исследования являются данные о невостребованных билетах поездов «Сапсан» по направлению Москва – Санкт-Петербург в период 19.02.2020 – 04.03.2020 гг. Информация о невостребованных билетах позволяет проанализировать поведения пользователей железнодорожного транспорта во время перемещений по маршруту Москва-Санкт – Петербург в недельном и суточном разрезе за период 19.02.2020 – 04.03.2020 гг.

Цель: Анализ поведения пассажиров ж/д транспорта на основе общедоступных больших данных с официального сайта РЖД.

Материалы и методы: Обнаруженный суточный цикл дает нам возможность сделать вывод о предпочтениях пассажиров и позволяет производить дальнейшие исследования в данном направлении, опираясь на полученные результаты. Результатом исследования можно назвать нахождение некоторых закономерностей в предпочтениях пассажиров. Данное исследование актуально для компаний, сфера деятельности которых сопряжена с железнодорожным транспортом, а также - для компаний, непосредственно участвующих в перевозке. Данные, полученные в исследовании, позволяют, при помощи отдельных методов интеллектуального анализа или прикладного программирования, производить А/Б тестирование, выстраивать маркетинговые стратегии, оптимизировать производство и решать прочие задачи, связанные с оценкой пассажиропотока и перемещения пассажиров.

Результаты: Обнаружен волнообразный суточный цикл максимальной загруженности линии, выделенный двумя временными периодами: 06.00–09.00 и 15.00–17.00. Обнаружено, что наиболее низкая заполняемость была на протяжении рабочей недели. Выявлено, что более высокая заполняемость – около 90 % – была с пятницы по воскресенье как в одном, так и в другом направлении движения. Также обнаружено

Received: 03.03.2020 **Revised: 07.05.2020** **Accepted: 30.06.2020**
Поступил: 03.03.2020 **Одобрена: 07.05.2020** **Принята: 30.06.2020**

различие заполняемости между направлениями движения поездов. Поезда, движущиеся из Москвы в Санкт-Петербург, отличается более плотной заполняемостью.

Ключевые слова: большие данные, скоростной транспорт, аналитика открытых данных, поведение пользователей.

Rubric 4. TRANSPORT ECONOMICS

© Alexander L. Kamenkov
Emperor Alexander I St. Petersburg State Transport University
(St. Petersburg, Russia)

THE APPLICATION OF BIG DATA FOR THE ANALYSIS OF PASSENGER FLOW ON THE HIGH-SPEED LINES OF THE RUSSIAN FEDERATION

Background: Data that combines techniques and technologies that make sense of information at the extreme limit of practicality is called "big data." Big data contributed to a deeper understanding of the behavior of the features of many systems, and, in particular, the transport system. Big data from various sources with a different range of data scales is becoming more accessible to society. However, in the framework of railway transport, obtaining big data relevant for users, which are both private individuals and companies, is a difficult task, and data collection tools are scarce and their absence is an obstacle to an in-depth study of the patterns of human mobility. The informational basis of the study is the data on unclaimed Sapsan train tickets in the direction Moscow - St. Petersburg during the period February 19, 2020 - March 4, 2020. Information on unclaimed tickets allows us to analyze the behavior of railway users during their travels along the Moscow - St. Petersburg route in a weekly and daily breakdown for the period February 19, 2020 - March 4, 2020.

Aim: Analysis of the behavior of passengers of railway transport based on publicly available big data from the official website of Russian Railways.

Materials and Methods: The discovered daily cycle gives us the opportunity to conclude on the preferences of passengers and allows us to carry out further research in this direction, based on the results. The result of the study can be called finding some patterns in the preferences of passengers. This study is relevant for companies whose field of activity is associated with rail transport, as well as for companies directly involved in transportation. The data obtained in the study allow, using separate methods of intellectual analysis or applied programming, to perform A / B testing, build marketing strategies, optimize production and solve other problems associated with assessing passenger flow and passenger movement.

Results: A wavy diurnal cycle of maximum line congestion was detected, distinguished by two time periods: 06.00 - 09.00 and 15.00 -17.00. It was found that the lowest occupancy rate was during the working week. It was revealed that a higher occupancy rate - about 90 % - was from Friday to Sunday in both one and the other direction of movement. A difference in occupancy between train directions was also found. Trains moving from Moscow to St. Petersburg are more densely populated.

Keywords: *Big data, high-speed transport, open data analytics, user behavior.*

Received: 03.03.2020 **Revised: 07.05.2020** **Accepted: 30.06.2020**
Поступил: 03.03.2020 **Одобрена: 07.05.2020** **Принята: 30.06.2020**

ВВЕДЕНИЕ

Каждую секунду огромное количество данных записывается с помощью социальных сетей, данных Wi-Fi, данных сотовой связи [1].

Эти, так называемые, «большие данные» содержат в себе ценную информацию о том, как люди взаимодействуют друг с другом и со своим окружением, предоставляя большие возможности для детального изучения структуры, функции и динамики человеческих систем, а также для понимания пространственно-временных закономерностей на макро- и микроэкономических уровнях [2]. В то же время развитие интернет-технологий, высокопроизводительных вычислений и возможностей хранения данных теперь позволяют нам эффективно использовать большие данные.

Большие данные о человеческой мобильности оказались особенно полезными для характеристики человеческого поведения, что привело к открытию многих закономерностей, которые не могут быть достигнуты с помощью традиционных пространственно-временных исследований и статистических данных [3].

Возникает вопрос - чем отличаются большие данные от обычных данных [4]:

1) Формат;
2) Объем данных;
3) Тип моделей решателей;
4) Распределенность данных;
5) Вычислительные ресурсы.

Различные источники данных, такие как социальные сети, отчеты о действиях пользователей на веб-сайтах были исследованы для того, чтобы понять структуру поведения пользователей в социальных сетях, общественные настроения и модели путешествий человека. Многие из этих данных могут эффективно фиксировать траектории движения людей, выявляя закономерности в мобильности человека и обеспечивая детализацию в понимании структуры и функциональности городских систем. Большие данные являются уникальным продуктом для использования производителями данных или государственных органов, что исключает проведение более глубоких исследований. Большие данные, доступные для общественности становятся все более необходимыми, так как могут быть использованы в качестве мощного инструмента для визуализации и понимания антропогенных структур и социальной динамики в различных пространственных масштабах. Быстрое развитие систем общественного транспорта позволяет восстанавливать траектории движения для каждого отдельного индивидуума [5].

Received: 03.03.2020 Revised: 07.05.2020 Accepted: 30.06.2020
Поступил: 03.03.2020 Одобрена: 07.05.2020 Принята: 30.06.2020

Также при сочетании этих данных с передовыми методами интеллектуального анализа данных, такими как применения нейросетевых моделей распознавания элементов, верификационные методы, методы детектирования и прочие – может возникнуть множество возможностей для изучения паттернов перемещения. Например, при помощи алгоритмов кластеризации и метода грубых множеств можно выявить различные группы транзитных пассажиров с различными моделями движения, можно оценить производительность транзитных сетей и характеристики пространственно-временных моделей коммутации пассажиров общественного транспорта. Задачей нашего анализа будет являться нахождение закономерностей в перемещениях пассажиров и изучение характеристик паттернов перемещения.

Чтобы решить эту задачу, в данной работе исследован источник данных о железнодорожном транспорте, который не был исследован ранее: веб-сайт по продаже билетов на поезда «Сапсан» [6]. С развитием интернета онлайн-системы продажи билетов получили все большее распространение как в России, так и во всем мире. Большое количество сайтов по продаже тех или иных билетов позволяет пассажирам запрашивать количество оставшихся билетов (количество мест, доступных для покупки в данный момент) на определенные маршруты. С помощью интерфейсов прикладного программирования (API) можно автоматически собирать эти оставшиеся данные по билетам с веб-сайтов, что позволяет объединять информацию для всех предопределенных маршрутов в течение предопределенного периода времени, тем самым создавая вид больших данных. Однако, в случае РЖД возникла сложность с оценкой данных, так как они не имеют открытого API, в следствии чего появилась необходимость сделать инструмент, который собирает данные по конкретному поезду в автоматическом режиме. Информация о конкретных пассажирах отсутствует, поэтому произвести анализ индивидуальных пассажирских траекторий не представляется возможным. В этой статье будет показано, как информацию о билетах можно использовать для характеристики перемещающихся паттернов. Объектом нашего исследования будут являться данные о железнодорожном поезде «Сапсан», курсирующем по направлению Москва – Санкт-Петербург. Российские высокоскоростные поезда пока не дошли до точки развития, сопоставимой с европейскими или азиатскими аналогами, однако шаги, проделанные в последнее время, говорят о том, что государство, как основной инвестор, очень заинтересовано в повышении скоростей передвижения. Мы ожидаем, что наша работа привлечет внимание к этому уникальному типу открытых данных для изучения других аналогичных транспортных систем.

Received: 03.03.2020 Revised: 07.05.2020 Accepted: 30.06.2020
Поступил: 03.03.2020 Одобрена: 07.05.2020 Принята: 30.06.2020

МЕТОДЫ РАСЧЕТА

В настоящее время в России только формируется система высокоскоростных магистралей. Ожидается, что к 2030 году протяженность ВСМ РФ будет равно 4300 км, что в настоящее время составило бы около 5 % общей протяженности высокоскоростных железных дорог в мире. По прогнозам, ВСМ РФ к 2030 году охватит 100 млн. человек населения, что сейчас составляет более 65 процентов [7]. В настоящее время функционируют несколько скоростных магистралей в РФ: Санкт-Петербург – Москва, Москва – Нижний Новгород. В данной работе мы будем рассматривать только СМ (скоростная магистраль) Санкт-Петербург – Москва. Средняя скорость движения поездов по этой магистрали – 180 км в час, что позволяет более 1000 пассажирам одновременно в течении 3–4 часов перемещаться между двумя крупнейшими городами РФ – Москвой и Санкт-Петербургом.

Сайт продажи билетов для СМ «Сапсан» [6] предоставляет_удобный интерфейс, который позволяет пассажирам в любой момент времени запрашивать количество оставшихся билетов между любыми двумя заданными станциями на маршруте СМ. Мы разработали специальный API интерфейс, используя язык программирования Python, чтобы сделать такой запрос автоматически. Это позволило нам получить количество оставшихся билетов на данный регулярный поезд непосредственно перед отправлением. Мы показываем, что, по крайней мере, два типа полезной информации могут быть получены следующим образом:

1) «Сырые» данные оставшихся билетов. Эти данные могут точно измерить остаточный или неиспользованный транзитный потенциал СМ.
2) Количество пассажиров в каждом конкретном поезде и заполняемость. Количество пассажиров можно оценить по количеству неиспользованных билетов, которые вычисляются, как разница между общим количеством билетов и количеством оставшихся билетов в последнюю секунду перед отправлением. Коэффициент заполняемости рассчитывается как соотношение общего числа пассажиров к общему числу мест в поезде.

Мы считаем, что сайт продажи билетов заслуживает доверия, так как это единственный официальный сайт российских железных дорог и считаем, что нет признаков манипуляции данными.

Received: 03.03.2020 Revised: 07.05.2020 Accepted: 30.06.2020
Поступил: 03.03.2020 Одобрена: 07.05.2020 Принята: 30.06.2020

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Мы провели два типа анализа, чтобы проиллюстрировать, как вышеупомянутые типы полезной информации могут быть использованы для понимания моделей перемещения на высокоскоростных магистралях и скоростных магистралях. Наши анализы описываются на участке Санкт-Петербург – Москва, представленном на Рис. 1. Санкт-Петербург и Москва – это основные города РФ, обладающие высоким уровнем развития агломерации и высокой плотностью населения. В настоящее время этот участок СМ является самым активным в РФ [8]. Как правило, поездка занимает около 4-х часов, а расстояние между городами равно около 700 км.

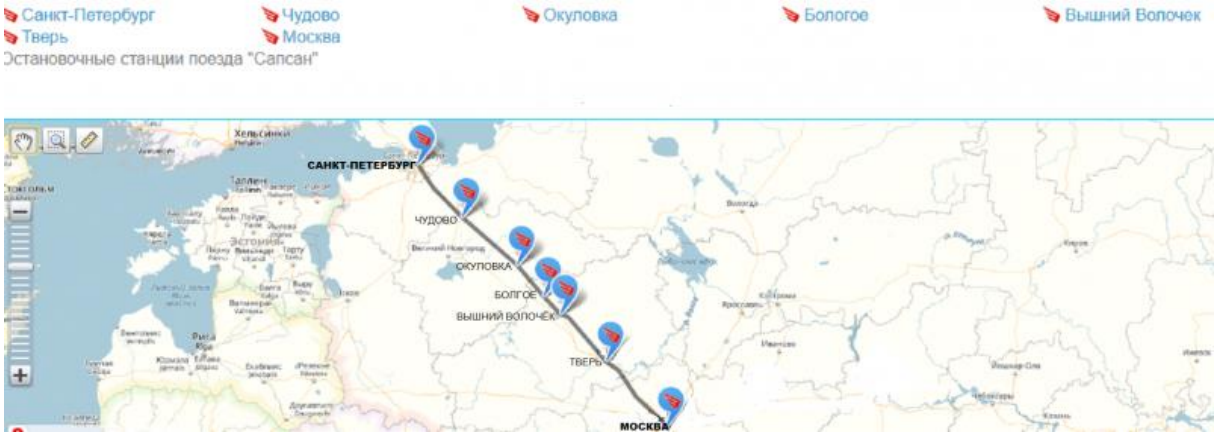

Рис.1. Расположение станций на СМ между Санкт-Петербургом и Москвой [8]

Первый тип анализа – анализ неиспользованного транзитного потенциала СМ

Мы собрали данные по наличию билетов для всех поездов СМ в направлении Санкт-Петербург – Москва на протяжении периода времени, равного 14 дней: с 19 февраля 2020 по 4 марта 2020. Рис. 2 представляет собой регулярную структуру оставшихся билетов. Одной из заметных особенностей является волнообразный суточный цикл, выделенный двумя временными циклами: 06.00–09.00 и 15.00–17.00. Именно в эти два отрезка времени было доступно наименьшее количество билетов. Эту особенность можно объяснять привязанностью к рабочему графику. Пользователи СМ, выезжающие с утра, предпочитают завершить свою поездку до полудня и иметь весь день в своем распоряжении, что предоставляет возможность продолжить работу, но уже в другом городе. В то время пользователи, отъезжающие после обеда, предпочитают добраться до точки назначения до ночи и иметь возможность воспользоваться прочим общественным

Received: 03.03.2020 Revised: 07.05.2020 Accepted: 30.06.2020
Поступил: 03.03.2020 Одобрена: 07.05.2020 Принята: 30.06.2020

транспортом, который еще будет функционировать во время прибытия поезда [9]. Еще одно примечательное наблюдение состоит в том, что в пятницу и в воскресенье почти недоступны билеты на вечерние поезда. Это объясняется тем, что пользователи СМ перемещаются между работой и домом в выходные дни. Также, можно предположить, что влияние на данный спрос оказывают пассажиры, использующие СМ с целью путешествия или с прочей социальной целью. Особенно ярко это заметно в выходные дни.

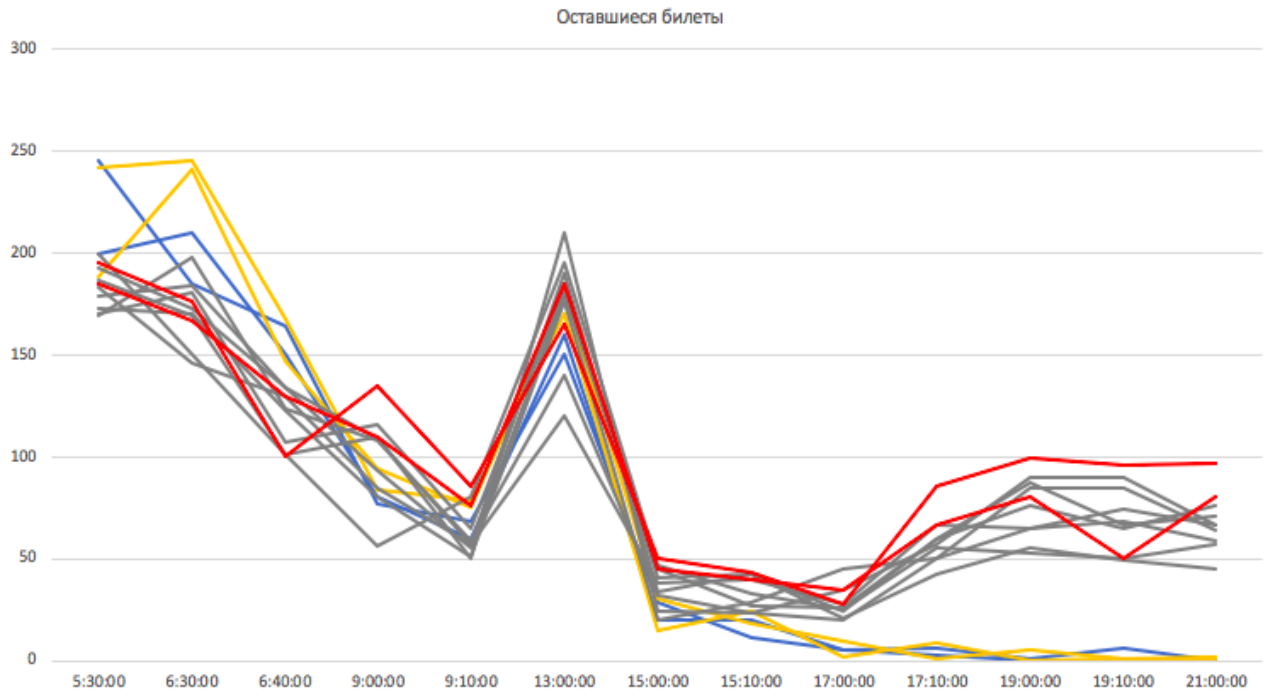

Рис 2. Количество оставшихся билетов

На данном графике отражены линии количества оставшихся билетов в разрезе по временному промежутку и по дням недели, где:

- Серый цвет – понедельник-четверг;
- Желтый цвет – пятница;
- Красный цвет – суббота;
- Синий цвет – воскресенье.

Анализ показывает, что коллективное поведение обладает повышенной регулярностью. Более того, данный факт может послужить отправной точкой для решения интересных вопросов, таких как изменение мобильности, изменение спроса на перемещения и другие.

Received: 03.03.2020
Поступил: 03.03.2020
Revised: 07.05.2020
Одобрена: 07.05.2020
Accepted: 30.06.2020
Принята: 30.06.2020

Второй тип анализа: количество пассажиров и заполняемость

Каждый день из Санкт-Петербурга отправляется 13 поездов «Сапсан». Для понимания того, выполняют ли эти поезда свою основную функцию в плане перевозок, мы собрали ежедневный показатель заполняемости. Мы сравнили два показателя заполняемости: по направлению Санкт-Петербург – Москва и в обратном направлении. Ознакомиться с данными показателями можно на Рис. 3.

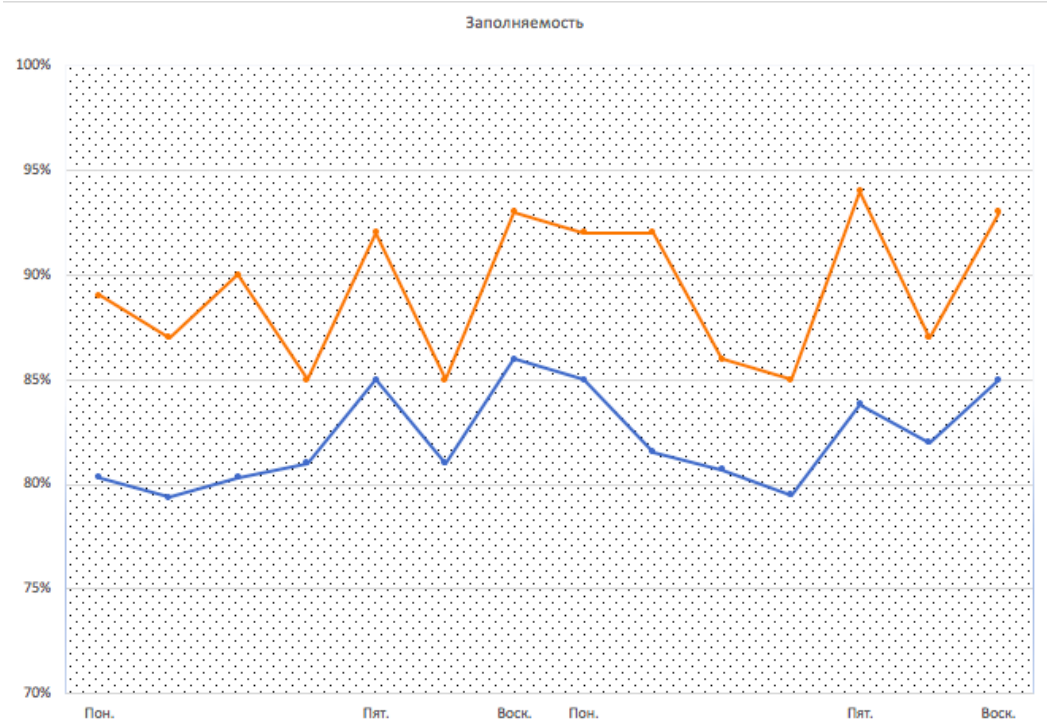

Рис. 3. Заполняемость поездов Сапсан

Синей линией отражается заполняемость поезда, следующего по направлению Санкт-Петербург – Москва, а оранжевой – по направлению Москва – Санкт-Петербург. Как можно увидеть из графика, наиболее низкая заполняемость была на протяжении рабочей недели. Более высокая заполняемость – около 90 % – была с пятницы по воскресенье как в одном, так и в другом направлении движения. Также явно заметно различие заполняемости между направлениями движения поездов. Направление Москва–Санкт-Петербург отличается более плотной заполняемостью. Это обуславливается тем, что уровень жизни в Москве, в среднем, более высокий, и жители столицы предпочитают проводить время в Санкт-Петербурге [11]. Подобно необработанным данным из первого анализа, анализ уровня заполняемости на станциях в диапазоне пространственных и временных масштабов может дать ценное представление о различных моделях поведения во время передвижения. Анализ показал, что данные об объеме потока могут способствовать лучшему пониманию системы СМ.

Received: 03.03.2020 **Revised: 07.05.2020** **Accepted: 30.06.2020**
Поступил: 03.03.2020 **Одобрена: 07.05.2020** **Принята: 30.06.2020**

ЗАКЛЮЧЕНИЕ

Данное исследование является результатом анализа данных открытых источников о доступности билетов поездов СМ Москва-Санкт – Петербург и оценки их заполняемости. Данный анализ используется в качестве примера, позволяющего обосновать новые аналитические процедуры исследования больших данных и применения их в целях роста эффективности пассажирских перевозок. Как и любой другой тип больших данных, данные о доступных билетах далеки от совершенства [10]:

1) Они генерируются на уровне отдельных поездов, поэтому не могут быть использованы для получения траекторий движения на уровне отдельных пассажиров;
2) Исторические данные не всегда доступны на транспортных сайтах;
3) Данные не могут быть точно получены для всех станций;
4) Валидация все еще необходима для оценки непредвзятости данных.

Приведенный анализ показал заметный волнообразный суточный цикл количества оставшихся билетов, выделенный пиком дефицита билетов в вечернее время суток. Была выделена несбалансированность потока на протяжении недели на поездах Москва – Санкт-Петербург и Санкт-Петербург – Москва в период 19.02.2020–04.03.2020 гг. Привлекая дополнительные массивы данных, в том числе более крупных пространственных и временных масштабов, можно получить более явные закономерности в перемещениях пассажиров, а также – использовать данные для предиктивной аналитики.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке научного руководителя, профессора кафедры «Экономика транспорта» д.э.н. Журавлевой Натальи Александровны.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК / References

1. Kitchin R. The Real-Time City? Big Data and Smart Urbanism. SSRN Electronic Journal [Internet]. Elsevier BV; 2013. Available from: http://dx.doi.org/10.2139/ssrn.2289141
2. Batty M. Big data and the city. *Built Environment.* 42(3):321-337. doi: 10.2148/benv.42.3.321
3. Gonza´lez MC, Hidalgo CA, Baraba´si A-L. Understanding individual human mobility patterns. *Nature.* 2008:779-782. doi: 10.1038/nature06958
4. Павлов А.И. Большие данные в фотограмметрии и геодезии // Образовательные ресурсы и технологии. – 2015. – № 4(12). – С. 96–100. [Pavlov AI. Bol'shie dannye v fotogrammetrii i geodezii. *Obrazovatel'nye resursy i tekhnologii.* 2015;4(12):96-100. (In Russ.)]. Доступно по: https://www.muiv.ru/vestnik/pp/chitatelyam/poisk-

Received: 03.03.2020 Revised: 07.05.2020 Accepted: 30.06.2020
Поступил: 03.03.2020 Одобрена: 07.05.2020 Принята: 30.06.2020

po-statyam/8556/46166/. Ссылка активна на: 16.04.2020.
5. Travel card and Oyster [Internet]. [16 April 2020] Available from: https://www.nationalrail.co.uk/times_fares/ticket_types/46575.aspx
6. Официальный сайт РЖД [Officialnii sait RZD [Internet]. (In Russ.)]. Доступно по: https://www.rzd.ru/. Ссылка активна на: 16.04.2020.
7. ВСМ в России [VSM v Rossii (In Russ.)]. Доступно по: http://www.hsrail.ru/info/vsmr/. Ссылка активна на: 16.04.2020
8. Гид по маршруту поезда «Сапсан» [Gid po marshrutu poezda "Sapsan" [Internet]. (In Russ.)]. Доступно по: https://peterburg.center/ln/gid-po-marshrutu-poezda-sapsan-sankt-peterburg-moskva.html. Ссылка активна на: 16.04.2020
9. David Gordon - Phoenix - Therapeutic Patterns of Milton H. Erickson, META Publications; 1st edition (June 1, 1981) p. 105-142
10. Franks B. The Analytics Revolution. John Wiley & Sons, Inc.; 2014 Sep 17; [16 April 2020] Available from: http://dx.doi.org/10.1002/9781118936672.
11. Уровень жизни в регионах России [The standard of living in the regions of Russia [Internet]. (In Russ.)]. Доступно по: https://gks.ru/folder/13397. Ссылка активна на: 16.04.2020.

Сведения об авторе:
Каменков Александр Леонидович, аспирант кафедры «Экономика транспорта»
eLibrary SPIN: 5385-2508; ORCID: 0000-0001-7052-5353
E-mail: sashaskotch@gmail.com

Information about the author:
Alexander L. Kamenkov, graduate student of the Department of Transport Economics
eLibrary SPIN: 5385-2508; ORCID: 0000-0001-7052-5353
E-mail: sashaskotch@gmail.com

Цитировать:
Каменков А.Л. Применение больших данных для анализа пассажиропотока на скоростных магистралях Российской Федерации // Транспортные системы и технологии. – 2020. – Т. 6. – № 2. – С. 106–115. doi: 10.17816/transsyst202062106-115

To cite this article:
Kamenkov AL. The application of Big Data for the Analysis of Passenger Flow on the High-Speed Lines of the Russian Federation. *Transportation Systems and Technology*. 2020;6(2):106-115. doi: 10.17816/transsyst202062106-115

Received: 03.03.2020 **Revised: 07.05.2020** **Accepted: 30.06.2020**
Поступил: 03.03.2020 **Одобрена: 07.05.2020** **Принята: 30.06.2020**