УДК 332.143

Регион: экономика и социология, 2018, № 4 (100), с. 69–88

Ю.П. Воронов

РЕГИОНАЛЬНАЯ СТАТИСТИКА В КОНТЕКСТЕ БОЛЬШИХ ДАННЫХ

В статье обсуждаются проблемы пространственной статистики в условиях, когда активно используются «большие данные». Приведены примеры сдвигов в зарубежной практике и практического совместного применения официальной статистики и больших данных из исследований автора. Показано, что данные региональной статистики станут другими. Например, от статистики цен на непроданные товары будет сделан переход к статистике по данным кассовых аппаратов. Изменятся и результаты расчетов по межрегиональным экономико-математическим моделям. Сделан вывод о необходимости ускоренного вовлечения больших данных в расчеты по моделям и в работу Росстата, с тем чтобы официальные статистические данные и результаты моделирования стали более полезными для решения практических и исследовательских задач.

Ключевые слова: большие данные; демография; геолокация; социальные сети; посевные площади; спрос на жилье; потребительские цены

Вначале – два определения, важных для дальнейшего изложения. *Официальная статистика* есть обязательная часть информационной системы демократического общества, работающая на нужды государства, экономики и общества с данными о демографии, экономике, социальной и экологической ситуации[1]. *Большие данные* – это потоки информации, формирующиеся сами по себе по мере потребности в них или даже при отсутствии такой потребности. Большие данные определяются как неструктурированная информация, появляющаяся

[1] URL: http://unstats.un.org/unsd/methods/statorg/FP-English.htm .

в разных сферах человеческой жизни, и задаются примерами: социальные сети, данные видеокамер, спутников, медицинские карты, трафик сотовой связи и т.д. Этот информационный ресурс по объемам многократно превышает официальную статистику, которая лучше него по системной организации собираемых ею данных. Термин «большие» расшифровывают по трем параметрам: объемы, скорость и разнообразие. По-английски – 3V: Volume, Velocity, Variety[2].

Большие данные представляют собой информационный ресурс, привлекательный для использования в экономике и социальной сфере. При сохранении систем официальной государственной и международной статистики сформировалось своеобразное двоевластие, когда два информационных ресурса сосуществуют, практически не пересекаясь. При взгляде на большие данные со стороны официальной статистики они делятся на шесть категорий[3]:

- административные: страховые полисы, банковская информация, посещения медицинских учреждений и т.д.;
- коммерческие: информация о торговых сделках, о платежах с банковских карт и через сотовые телефоны, о слияниях и поглощениях и т.д.;
- приборные: спутниковые снимки, данные видеокамер на дорогах и в городах, информация метеостанций и т.д.;
- данные геолокации (геопозиционирования), которые пока делятся на две связанные между собой группы: позиционирование сотовых телефонов и информация систем GPS и Глонасс;
- поведенческие: поиск покупателями определенных товаров, услуг и работы, посещения сайтов и отдельных страниц;
- общественное мнение: суждения, высказываемые в социальных сетях, на заседаниях, собраниях, конференциях, семинарах, митингах и т.д. Здесь встают проблемы обоснованности суждений и отделения эмоциональной составляющей высказываний.

[2] URL: http://www.gartner.com/it-glossary/big-data/ .

[3] См.: *What* does «big data» mean for official statistics? United Nations Economic Commission for Europe Conference of European Statisticians, 10 March 2013.

Новые источники информации вызывают появление новых проблем, каковые можно разделить на четыре группы:

- правовые, касающиеся прав на доступ к данным и на их использование и связанные с конфиденциальностью, т.е. с обеспечением доверия к соблюдению ограничений на доступ к данным третьих лиц;
- финансовые, когда затраты на доступ к данным сопоставляются с получаемыми выгодами;
- управленческие, т.е. касающиеся правил и практики управления данными и способами их защиты. Это особенно сложно, если информация позволяет получить новые результаты, какие при использовании официальной статистики были недостижимы;
- методические, связанные с качеством данных и наличием статистических методов, обеспечивающих их достоверность. Использование новых методов, принципиально отличающихся от принятых статистических процедур, требует привлечения новых математических и статистических алгоритмов. Большие данные – это информация, самопорождаемая и курсирующая сама по себе. Она изначально не может быть репрезентативной. Идет разработка новых методов, которые бы согласовывали между собой информацию, полученную из множества нерепрезентативных выборок.

ПЕРВЫЕ ШАГИ К ИНТЕГРАЦИИ

На соединение с большими данными двинулась официальная статистика. Бюро переписей США (U.S. Census Bureau) еще полвека назад стало продавать статистические таблицы по заказам. Чем подробнее таблица, тем она ближе к тому уровню, который допускает использование больших данных для независимой проверки на достоверность или репрезентативность данных, собранных официальной статистикой (во время переписи или в результате специального статистического обследования). В Международный год статистики (2013 г.) главами европейских статистических служб был принят Схевенингенский меморандум «Big data и официальная статистика». В мемо-

рандуме зафиксировано, что большие данные представляют собой явление, касающееся многих сфер экономики и социальной сферы, что инновации в технологиях информатики и связи во все большей мере делают экономику цифровой и это расширяет возможности статистики. С этим связаны многие европейские инициативы. Большие данные, в частности, позволят построить систему цепочек добавленной стоимости по всем странам ЕС.

В 2014 г. Статистическая комиссия ООН организовала обсуждение меморандума «Big data и модернизация официальной статистики». Можно сказать, что с тех пор большие данные и официальная статистика стали активно интегрироваться. Исходная информация интенсивно подвергается оцифровке: тексты, схемы, карты, картинки, видеоизображения, аудиозаписи теперь представлены в цифровой форме большими объемами. По мнению американских статистиков, методы анализа больших данных имеют значительный потенциал для совершенствования официальной статистики и позволяют существенно расширить возможности получения знаний о социально-экономических явлениях и процессах [6].

Поэтому официальная статистика должна осваивать новые методы обработки разнородной информации, принципиально отличающиеся от привычных. Для этого нужно привлекать ученых, преподавателей университетов, практиков, работающих во многих отраслях. Следует также провести переподготовку работников государственной статистики. В меморандуме сказано, что работа с большими данными должна обязательно включаться в годовые планы работ Евростата[4].

Преградой на этом пути является отсутствие надежных программных средств систематизации больших данных. По этой и некоторым другим причинам процесс идет медленно, хотя многие считают, что только использование больших данных как дополнительного и важного источника информации позволит совершенствовать аналитику в социально-экономической сфере. Тем самым будет защищено будущее официальной государственной и международной статистики, ко-

[4] См.: *Regulation* (EU) No 99/2013 of the European Parliament and of the Council of 15 January 2013 on the European Statistical Programme 2013-17 // OJ. – 2013. – L 39. – P. 12–29.

торое без союза с методами анализа больших данных представляется крайне сомнительным.

Многие это понимают. Так, Европейская экономическая комиссия создала он-лайн ресурс Unite Big Data Inventory Home. Это платформа, на которой любой участник может разместить массивы своих данных. Там можно найти данные органов государственного управления, данные с камер видеонаблюдения дорожного движения, информацию, собранную с «умных» приборов учета, и множество иной информации[5]. Этот шаг необходим как предварительный, поскольку прежде чем интегрировать потоки информации из разных источников, нужно научиться собирать их в одном месте.

Европейское статистическое агентство (Eurostat) поставило перед собой цель проработать два подхода к работе с большими данными. Первый подход назван «от пользователя» (user-centric), второй получил название «от сети» (web-centric). В первом идет классификация пользователей сети, во втором – классификация событий в сети. Оба подхода опираются на концепцию «объединенных открытых данных» (federated open data). В объединенные открытые данные включаются все данные официальной статистики и та часть больших данных, которые удалось с ними состыковать [11].

Наша страна в этом плане явно не находится на передовой линии союза официальной статистики и больших данных. Только с 2023 г. Росстат начнет по каналам связи получать из Федеральной государственной информационной системы «Единый государственный реестр ЗАГС» регулярную информацию о регистрации рождений, смертей, браков и разводов[6]. Это означает, что Всероссийская перепись населения 2020 г., по которой было намечено формирование показателей с использованием больших данных, не будет использовать материалы ЗАГСов. Более того, этот процесс начнется только после обработки данных переписи, т.е. спустя три года после ее проведения [3].

[5] URL: https://statswiki.unece.org/display/BDI/UNECE+Big+Data+Inventory+Home .

[6] См.: *Концепция* создания системы персонального учета населения Российской Федерации (одобрена распоряжением Правительства РФ от 9 июня 2005 г. № 748-р).

Предполагается, что пройдет долгое время, пока из отдельных региональных методик сформируются национальные методики, а только затем появится единая международная методика. Только тогда будут возможны сравнительные исследования и сопоставления по разным регионам и странам[7].

МИКРОДАННЫЕ (MICRODATA) – МОСТ МЕЖДУ СТАТИСТИКОЙ И БОЛЬШИМИ ДАННЫМИ

При использовании больших данных статистика становится более прозрачной и проверка статистической информации возможна даже для тех, кто не имеет доступа к первичной информации переписи. Так, одним из потоков больших данных является так называемая кладбищенская информация открытого доступа. Любой, кто зайдет на кладбище, может записывать информацию, нанесенную на надгробия: пол умершего, даты рождения и смерти. При межрегиональных и межпоселенческих демографических сравнениях такая статистика может быть источником очень полезной информации. Эту же информацию можно было бы получать и просто сделав запрос в администрацию кладбища или в тот же ЗАГС. Но на такой запрос, если нет соответствующих прав доступа, информации не получишь.

Мы можем заметить, что подобной информации много повсеместно: кроме сведений ЗАГСа есть еще данные о платежах за квартиры, информация налоговой инспекции, данные историй болезни и т.д. Вся эта информация получила название «микроданные» (microdata). Микроданные немногим отличаются от больших данных – всего лишь тем, что они существовали давно, еще до кардинальных изменений в информационной сфере. Так, Департамент ОЭСР[8] по науке, технологиям и инновациям создал Лабораторию микроданных для анализа и интеграции пяти баз данных, которые уже были у этой организации:

[7] Более оптимистические оценки приводятся по готовности к big data российского бизнеса. См.: *Шаль А.В.* Технологии больших данных в статистике // Учет и статистика. – 2017. – № 2 (46). – С. 81–88.

[8] Организация экономического сотрудничества и развития объединяет 37 стран, 18% населения мира.

патентной (73 млн записей), базы данных компаний Orbis (86 млн), базы данных Scopus (26 млн), базы данных о торговых марках (6 млн) и базы прав на конструктивные решения (около 1 млн)[9]. Совместное использование этих баз данных и национальной статистической информации позволяет проводить более глубокий анализ развития экономики и делать более точные прогнозы.

Наибольшие достижения в сфере микроданных имеет тройка международных ИТ-компаний: SAP, «Oracle» и «Microsoft». Они пока сосредоточены на производственных приложениях, так называемых ERP-системах. Но поскольку в этих системах уже используются микроданные, их можно применять и в масштабном статистическом анализе и включать в исходную информацию при создании межрегиональных экономико-математических моделей.

ГЕОЛОКАЦИЯ КАК БАЗА ОБЪЕДИНЕНИЯ ПРОСТРАНСТВЕННОЙ СТАТИСТИКИ С БОЛЬШИМИ ДАННЫМИ

Как было отмечено выше, как только статистические данные начинают разбивать на группы в развернутых таблицах, чаще всего учитывающих пространственные факторы, так сразу появляется возможность проверить их достоверность на основе больших данных. Из всех признаков, по которым делят статистическую информацию, наибольший интерес для альянса статистики с большими данными представляет пространственный фактор.

Большие данные, как правило, содержат информацию о событиях. А любое событие обычно привязано к определенной точке в пространстве. Информацию по отдельным городам можно проверять по потокам больших данных. Что же это за потоки?

В первую очередь это данные геолокации, т.е. регулярное установление географического положения отдельных единиц наблюдения. Существует несколько источников геолокационных данных. Ос-

[9] URL: https://communications.elsevier.com/nl/jsp/m.jsp?c=%408cA%2Bz5VPZzRLTWL6vYhSTAuj7vUEhePQ3OXoT03NgaA%3D&utm_campaign=RN_AG .

новной источник в настоящее время – данные о позициях исходящих звонков мобильных телефонов.

Установлено, что если построить уравнение регрессии, описывающее зависимость количества сотовых телефонов от численности населения, то статистическая значимость коэффициентов такой регрессии будет очень высокой. Построив такую регрессию в предположении, что количество телефонов примерно соответствует численности населения, можно получить независимую (альтернативную) оценку численности населения. В мире сейчас количество работающих сотовых телефонов примерно равно половине численности населения. Но у одного человека может быть два-три сотовых телефона, а у другого – ни одного. Поэтому корреляция между численностью населения и количеством сотовых телефонов обычно высокая – в пределах 0,8–0,85, но не бывает очень близка к единице [2].

Для повышения коррелированности при таких расчетах вводят дополнительно еще два параметра: структуры населения по возрасту и по плотности расселения. Эти параметры могут использоваться двумя способами: в составе множественной регрессии и как возможность построения частной регрессии, т.е. регрессии за вычитанием влияния плотности расселения и возраста.

Возможности позиционирования сотовых телефонов хорошо показывают себя в статистике миграции. По ним вычисляют не только миграцию внутри страны, включая маятниковую миграцию [10], но и международную миграцию в диапазоне от туристических поездок до переездов на ПМЖ [16].

В больших данных сложно использовать те критерии учета мигрантов и миграции в официальной статистике, которые используются в первую очередь в концепции обычного места жительства. Некоторые исследователи делают исходя из этого вывод, что большие данные не могут быть альтернативой традиционным источникам информации о миграции [1; 4].

Но с этим выводом вряд ли можно согласиться. Большие данные по-новому ставят проблему мобильности населения. В традиционной статистике необходимо было разрабатывать критерии, по которым постоянные жители отличались от временно проживающих. Именно

численность постоянных жителей была основой для межрегиональных сравнений. Если появляется возможность постоянного мониторинга за местопребыванием человека через геолокацию, то для понимания и мониторинга демографических событий простое деление жителей региона или населенного пункта на постоянных и временных выглядит слабым, не отражающим социально-экономические процессы во всей полноте. По большим данным возможно отделять, например, постоянно проживающих мигрантов от тех, кто приезжает на строительный сезон, а на зиму уезжает к себе на родину, что ранее было невозможно [2].

Геолокационная информация регулярно собирается самими провайдерами сотовой связи в форме анонимной статистики об исходящих и входящих звонках. Для понимания пространственных направлений развития сетей сотовой связи нужно собирать информацию об участках сети с наиболее напряженным трафиком и о пиковых нагрузках на этих участках. Поскольку такая информация может пригодиться не только провайдерам, они начали продавать ее заинтересованным партнерам. Пионером продаж в этой сфере считается американская компания-провайдер «Orange», а первыми потребителями – компании и муниципалитеты, исследующие дорожные пробки[10]. Чем больше пробка, тем больше трафик на этом участке сети. Но эта информация может быть полезна и для более масштабных исследовательских обобщений.

В Новосибирской области такую информацию по Сибири уже несколько лет собирает компания «Эксперт связи»[11]. Однако спрос на нее пока есть только у проектировщиков и отдельных провайдеров, исследующих ситуацию на рынке сотовой связи. Аналогичные компании существуют и в других регионах, и на федеральном уровне. В отличие от зарубежной практики сами провайдерские компании этим бизнесом не занимаются.

---

[10] URL: www.orangebusiness.com/mnc/press/press_releases/2012/mediamobile.html .

[11] URL: https://www.expertsvyazi.ru/ . Часть информации находится в свободном доступе.

ПОЗИЦИОНИРОВАНИЕ ПО СЕТИ ИНТЕРНЕТ

Второй источник демографической информации – выходы пользователей в сеть Интернет, дающие возможность установить их географическое положение. Кроме этого, географические позиции устанавливаются по Skype [8] и по социальным сетям, прежде всего по Facebook, Twitter [15] и LinkedIn [13]. Последняя сеть особенно ценна при изучении мест концентрации научных кадров.

Демографические процессы событийные. К числу событий, их составляющих, относятся браки и разводы. Поэтому данные из Интернета о фактах свадеб и других семейных событиях могут стать вспомогательным информационным ресурсом для демографических исследований. В данном случае требуется решить две методические проблемы. Первая – унификация исходной информации, вторая – разделение этой информации по категориям, имеющим определенное соответствие с категориями, используемыми в официальной статистике. Первая проблема в настоящее время решается методом, получившим название «вербальная аутопсия». Он состоит в запросах, которые подаются в поисковые системы в стандартной форме, с типовым набором слов. Вторая проблема решается также тщательной формулировкой запросов.

В некоторых исследованиях показано, что если в Google формировать запросы со словами типа «родился» или «беременна», то можно на несколько месяцев вперед прогнозировать не только численность новорожденных, но даже и желание их иметь [5]. Из совмещения намерений и фактических рождений появляется возможность построить индекс «склонности к рождению детей», который мог бы стать еще одним параметром выявления различий между регионами.

На основании таких данных даже разрабатываются специальные, ранее отсутствовавшие индикаторы склонности к рождениям. Так, Венский институт демографии (Австрия) разработал так называемый «Барометр фертильности», в котором большие данные согласуются с информацией о регистрации новорожденных. Но во многих странах отсутствует надежная система регистрации рождений и потому приходится использовать только данные радостных сообщений о вновь появившемся человеке.

Еще один пример построения параллельных индексов – Интернет-индекс вакансий (Internet Vacancy Index, IVI), который базируется на мониторинге сайтов четырех основных рекрутинговых агентств Австралии. Кроме него со статистическим индексом безработицы в Австралии конкурирует еще и газетный индекс вакансий (Skilled Vacancy Index, SVI)[12].

Но прямой зависимости между статистикой и данными вербальной аутопсии не существует. Доказано, например, что объем поиска на слово «аборт» обратно пропорционален числу фактических абортов на данной территории и прямо пропорционален строгости запретов на аборты на ней же [12].

Метод вербальной аутопсии используется также для анализа уровня смертности. Для этого в запросы вводятся слова «умер», «погиб», «с глубоким сожалением» и подобные. В настоящее время этот метод применяется преимущественно в развивающихся странах вследствие ограниченности и малой достоверности официальной статистики [14].

Естественным развитием идеи является отслеживание в социальных сетях сообщений о болезнях, в особенности это касается эпидемий гриппа [7]. Хотя именно по сообщениям о болезнях отмечаются наибольшие расхождения между результатами вербальной аутопсии и статистическими данными.

Новые оригинальные массивы данных формирует система мониторинга дорог и улиц Google Street View[13]. Эти данные сейчас уже научились использовать для оценки численности населения в отдельно взятом населенном пункте. Эти снимки содержат пометки «Google Карты» или «Просмотр улиц» и являются собственностью компании «Google». Лица и номерные знаки автомобилей на таких изображениях размываются автоматически с целью защиты личной информации.

[12] URL: https://www.jobs.gov.au/ . Сайт был создан Австралийским департаментом образования, занятости и поиска рабочих мест (Australian Department of Education, Employment and Workplace Relations, DEEWR). В настоящее время ликвидирован.

[13] URL: https://www.google.com/streetview/ .

Примерно в 190 странах мира каждые десять, а иногда и каждые пять лет проводятся переписи населения[14]. Они основываются преимущественно на интервью, которые проводят по месту жительства прошедшие инструктаж люди в течение ограниченного интервала времени, переходя от одной двери к другой. Этот метод дает как бы «моментный снимок» численности и состава населения. По этим моментным данным строятся прогнозы и аналитика. Дополнительным источником информации предполагаются записи актов гражданского состояния, т.е. информация событийная. В этом плане она похожа на то, что называется большими данными.

В заключение приведем примеры практического применения больших данных (помимо демографии) в тех сферах, где ранее использовались только данные официальной статистики[15].

ПРАКТИКА ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ ПОМИМО ПРОСТРАНСТВЕННОЙ ДЕМОГРАФИИ

Пример 1. Оценка посевных площадей. В советское время существовала строгая система учета посевных площадей, за нарушение которой полагалось уголовное наказание. Сегодня информация о посевных площадях собирается методом опроса, поскольку на обмеры независимыми землеустроителями средств не выделяется.

В большинстве стран мира замеры посевных площадей ведутся со спутников (remote sensing). На основании мониторинга фермерам, например, компенсируется отказ от посева тех сельскохозяйственных культур, цены на которые осенью могут упасть. В КНР только для оценки посевов пшеницы запущено четыре специализированных спутника. Даже Замбия имеет канал на французском спутнике, по которому идет оценка посевных площадей в этой стране. В Евросоюзе таких спутников шестнадцать, и они не только оценивают размеры посевных площадей, но и дают информацию для ранних прогнозов урожая. Более того, по данным спутников начисляются платежи от-

[14] См.: *UNSTATS*. Population and housing censuses, Feb. 2017. – URL: https://unstats.un.org/unsd/demographic/sources/census/alternativeCensusDesigns.htm .

[15] Примеры базируются на исследованиях автора.

дельным фермерам, если те отказались сеять, например, кукурузу на одном из своих полей. Их отказ позволяет облегчить поддержание высоких (стабильных) цен на эту культуру.

Использование данных дистанционного зондирования посевных площадей и посевов многогранно, но фундаментальным достижением для региональной (пространственной) статистики следует считать именно определение размеров посевных площадей. Если они неизвестны, то невозможно оценить урожайность и ее динамику.

В Новосибирской области в 2017 г. была отснята с беспилотника примерно четверть посевных площадей. Видеоматериалы не обработаны до сих пор. Причина не только в сложностях обработки, но и в проблеме использования будущих результатов: что с ними делать, если обнаружатся значительные расхождения с данными официальной статистики? Проведенная нами экспериментальная обработка нескольких десятков кадров показала заметные расхождения с данными официальной статистики.

Пример 2. Статистика цен. Во многих странах сейчас разрабатываются методологические подходы к использованию больших данных в статистике потребительских цен. Наряду с трудностями получения информации по каждой отдельной цене существует сложность определения системы географических точек для таких измерений[16].

Этот процесс имеет мощную пространственную составляющую. В настоящее время Росстат ведет мониторинг потребительских цен по 275 городам РФ. Если изменить набор городов, то и статистические результаты будут другими. Исследований зависимости уровня цен от состава попавших в выборку городов не проводилось. На основе же больших данных вполне вероятно построение статистики цен, не зависящей от выборки населенных пунктов. А это означает, что будут полнее учитываться цены в селах и малых городах, т.е. статистика цен станет более адекватной товарообороту.

[16] См.: *Воронов Ю.П.* Умение назначать цену // ЭКО. – 2007. – № 11; *Воронов Ю.П.* Умение назначить цену: Пособие по практическому ценообразованию. – Новосибирск: Изд-во СибАГС, 2007.

Новые методы сбора ценовой информации будут учитывать тот факт, что статистика цен в настоящее время построена как статистика цен на находящиеся на прилавках непроданные товары. Поэтому повышение цен на какую-либо группу товаров не означает, что цены повысились, а отражает тот факт, что дешевые товары из данной группы распроданы, ушли с прилавков.

Большие данные (чеки кассовых аппаратов) содержат информацию о ценах реальных продаж. Это информация другого свойства. Массачусетский технологический институт в начале 2013 г. начал разработку программного обеспечения для сбора ценовой информации через Интернет. Цель – построить новый индекс потребительских цен (Consumer Price Index, CPI). Он, совершенно очевидно, будет отличаться от того, что существует в настоящее время.

Пример 3. Структура покупателей. Сегодня появилась техническая возможность доступа к огромным массивам данных кассовых аппаратов, где фиксируются цены реально проданных товаров. Информация с них имеет относительно стандартизованную форму чека. По чекам можно получить информацию о ценах на проданные товары, которые будут отличаться от цен, собираемых официальной статистикой и маркетинговыми агентствами. Кроме того, по чекам можно выделить категории покупателей, совершающих покупки в конкретном магазине.

Так, в одном из наших исследований по крупному новосибирскому супермаркету «Посуда-центр» было обработано 1,3 млн чеков (немногим более годового цикла). По этой информации было выделено шесть типов покупателей[17].

Первый тип – «новички»: покупатель зашел в магазин, ничего не нашел и, чтобы поездка в магазин не пропадала, купил какую-то мелочевку. К данному типу, разумеется, относятся не только те, кто пришел в магазин впервые, хотя именно новички составляют основную его часть. Так что «новичок» – это не покупатель, который впервые за-

[17] Подробнее о методике выделения см.: *Воронов Ю.П.* Прикладная экспериментальная экономика. – Новосибирск: Изд-во ИЭОПП СО РАН, 2009. – С. 95–117.

шел в магазин, а тот, кто, по данным кассовых аппаратов, ведет себя как новичок.

Вторая категория – «целевики». Их представляют чеки с единственным, но дорогим товаром. Третий тип покупателей – так называемые «дарители», или покупатели, которые чаще всего покупают нечто дорогое в подарок, а затем приобретают какое-либо дешевое дополнение. В чеке должно быть два товара, один из которых дорогой, а второй – дешевый. Четвертый тип можно назвать «собирателями», ориентированными на спонтанные покупки. Он основной для магазина.

Пятый тип покупателей – мелкие оптовики. Их вычислить несложно по характеру чека. Для них характерно значительное количество разнообразных товаров в одном чеке, набор которых явно больше потребностей одной семьи. Это смешанная группа, кроме перепродавцов в нее входят и те, кто покупает товары для своих знакомых, родственников и друзей. И наконец, последний, шестой, тип покупателей – крупные оптовики. Это покупатели, приобретающие товаров заведомо больше, чем предназначено для личного потребления или для подарков.

Типология покупателей приводит к выводу, что супермаркет «Посуда-центр» представляет собой шесть относительно независимых видов бизнеса (виртуальных магазинов) в соответствии с выделенными шестью типами покупателей.

Для сравнительных межрегиональных исследований формируется уникальный источник информации – соотношение типов покупателей. Он только косвенно связан с ценами. Если в данном регионе высока доля «новичков», то это означает, что покупательная способность здесь низкая. Высокие доли «дарителей» и «собирателей» свидетельствуют об относительно высоком благосостоянии. Высокие доли покупателей категории «мелкий опт» говорят о распространенности малого предпринимательства, а категории «крупный опт» – о недостатках в товаропроводящей сети, являющихся причиной относительно высоких цен.

Пример 4. Спрос на жилье. В рамках разработки программы комплексного развития коммунальной инфраструктуры г. Бийска (Алтай-

ский край) рассчитывались суммы обременений для инвесторов, вкладывающих средства в жилищное строительство[18]. Инвесторы должны платить за подвод тепла, воды и канализации к тем домам, которые будут построены на их вложения. Чтобы обосновать такие обременения, мало было их скалькулировать. Нужно было убедить инвесторов, что средства, вложенные ими в инфраструктуру, окупятся.

Существовала статистика жилого фонда по микрорайонам города. В мэрии имелась также информация о масштабах жилищного строительства по этим же микрорайонам. Этот уровень информации можно было отнести к официальной статистике. Кроме такой информации существовали и потоки больших данных. Они представляли собой объявления о продажах квартир в конкретных микрорайонах.

Большое количество предложений о продаже квартир свидетельствует о том, что в некотором конкретном микрорайоне рынок квартир данного типа (например, трехкомнатных большой площади в кирпичном доме) близок к насыщению. На этом основании инвестору предлагалось либо изменить структуру квартир в новостройке, либо выбрать в городе другой участок для строительства жилья. Такие рекомендации позволяли инвестору более адекватно оценить перспективы будущих продаж квартир и более реально подойти к финансовым тратам на обременение.

Применительно к региональной статистике это позволяет иначе взглянуть на данные о вводе нового жилья и об уровне цен на квадратный метр. Если инвесторы не в состоянии быстро продать квартиры в домах, строящихся в том или ином микрорайоне города, они будут терпеть убытки и в зависимости от собственного оптимизма либо повышать, либо понижать цены. И этот мотив дезинформирует исследователей, занимающихся межрегиональными сопоставлениями.

Естественно, что переход к новой региональной статистике – долгий процесс. На включение подобной информации в региональные исследования потребуются годы. Но рано или поздно поток больших данных изменит содержание таких исследований.

---

[18] Работа выполнялась под руководством автора компанией «Корпус» (Новосибирск).

* * *

Официальная статистика во всех странах вынуждена корректировать действующие методы сбора информации в условиях, когда открылись возможности работать с потоками больших данных. Российские статистические органы пока только присматриваются к этому процессу.

Термин «региональная статистика» имеет еще одно значение – как статистика, ведущаяся в регионах по своим методикам. Региональной статистики в этом значении как института в нашей стране нет. Есть территориальные подразделения федеральной службы статистики, лишенные методологической, финансовой и прочей самостоятельности. Поэтому никто ничего в регионах делать не будет, пока это не войдет в федеральный план статистических работ с соответствующим финансированием и утвержденными методиками. Вполне возможно, что российский бизнес освоит работу с большими данными существенно раньше, чем официальная российская статистика.

При объединении статистики и больших данных в перспективе будут корректироваться базовые данные региональной статистики, прежде всего численность населения, размеры посевных площадей, уровни потребительских цен по регионам. Вполне возможно, что тогда изменятся и результаты расчетов по межрегиональным экономико-математическим моделям [9].

Список источников

1. *Мусин У.Р., Нусратуллин И.В.* Применение больших данных в оценке миграционных процессов // Вестник университета. – 2017. – № 7-8. – С. 188–193.

2. *Сарджвеладзе С., Савельева Н.* Власти выявят реальное население Москвы по сотовым телефонам. – URL: https://www.m24.ru/articles/01072014/48769?utm-source (дата обращения: 01.07.2018).

3. *Суринов А.Е.* Модернизация производства статистических данных в Российской Федерации // Вопросы статистики. – 2015. – № 10. – С. 3–13.

4. *Чудиновских О.С.* Большие данные и статистика миграции // Вопросы статистики. – 2018. – Т. 25, № 2. – С. 48–56.

5. *Billari F., Amuri D.F., Marcucci J*. Forecasting births using Google. Annual Meeting of the Population Association of America, 2013, New Orleans, LA.

6. *Bohon S.A*. Demography in the big data revolution: changing the culture to forge new frontiers. 2018, March. Population Research and Policy Review. – P. 1–19.

7. *Ginsberg J., Mohebbi M.H., Patel R.S. et al*. Detecting influenza epidemics using search engine query data // Nature. – 2008. – Vol. 457, No. 7232. – P. 1012–1014.

8. *Kikas R., Dumas M., Saabas A*. Explaining international migration in the Skype network: The role of social network features // Proc. SIdEWayS. – 2015. – P. 17–22.

9. *O'Neil C*. We don't need more complicated models, we need to stop lying with our models. – URL: https://mathbabe.org/2013/04/03/we-dont-need-more-complicated-models-we-need-to-stop-lying-with-our-models (дата обращения: 11.07.2018).

10. *Phithakkitnukoon S., Smoreda Z., Olivier P*. Socio-geography of human mobility: A study using longitudinal mobile phone data // PlOS One. – 2012. – No. 7 (6).

11. *Reimsbach-Kounatze C*. The Proliferation of «Big Data» and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis / OECD Digital Economy Papers. – Paris: OECD Publishing, 2015. – No. 245.

12. *Reis B.Y., Brownstein J.S.* Measuring the impact of health policies using internet search patterns: the case of abortion // BMC Public Health. – 2010. – Vol. 10, No. 1. – P. 514.

13. *State B., Rodriguez M., Helbing D., Zagheni E*. Migration of professionals to the U.S. – evidence from LinkedIn data // SocInfo. – 2014. – P. 531–543.

14. *Tamgno J.K., Faye R.M., Lishou C*. Verbal autopsies, mobile data collection for monitoring and warning causes of deaths // Advanced Communication Technology (ICACT). – 2013. – P. 495–501.

15. *Zagheni E., Garimella V.R.K., Weber I. et al*. Inferring international and internal migration patterns from twitter data // Proc. ACM Conf., 2014.

16. *Zagheni E., Weber I*. You are where your e-mail: using e-mail data to estimate international migration rates // Proc. of the 5th Annual ACM Web Science Conf. – 2012. – P. 348–351.

Информация об авторе

Воронов Юрий Петрович (Россия, Новосибирск) – кандидат экономических наук, директор ООО «Корпус» (630089, Новосибирск, мкр. Горский, 1, e-mail: corpus-cons@ngs.ru); ведущий научный сотрудник Института экономики и организации промышленного производства СО РАН (630090, Новосибирск, просп. Акад. Лаврентьева, 17).

DOI: 10.15372/REG20180403

Region: Economics & Sociology, 2018, No. 4 (100), p. 69–88

Yu.P. Voronov

SPATIAL STATISTICS IN THE CONTEXT OF BIG DATA

The article considers the problems of spatial statistics when using big data. It provides examples of changes in foreign practice and the author's practical joint implementations of statistics and big data. Regional statistics databases are to transform. For instance, a transition will be made from unsold goods price statistics to cash register data. Calculation results on economic-mathematical models are also expected to change. The article concludes with a need to accelerate big data mainstreaming into modeling and Rosstat functions so that model estimations and official statistics would become more useful in practical and research applications.

Keywords: big data; demography; geolocation; social media; sown area; demand for housing; consumer prices

References

1. *Musin, U.R.* & *I.V. Nusratullin*. (2017). Primenenie bolshikh dannykh v otsenke migratsionnykh protsessov [Application of big data in an assessment of migratory processes]. Vestnik universiteta [University Bulettin], 7-8, 188–193.

2. *Sardzhveladze, S.* & *N. Savelyeva*. (2014). Vlasti vyyavyat realnoe naselenie Moskvy po sotovym telefonam [City officials are to discover the real population of Moscow with phone numbers]. Available at: https://www.m24.ru/articles/трафик/01072014/48769 (date of access: 01.07.2018).

3. *Surinov, A.E*. (2015). Modernizatsiya proizvodstva statisticheskikh dannykh v Rossiyskoy Federatsii [Modernization of statistical production in the Russian Federation]. Voprosy statistiki [Issues of Statistics], 10, 3–13.

4. *Chudinovskikh, O.S.* (2018). Bolshie dannye i statistika migratsii [Big data and statistics on migration]. Voprosy statistiki [Issues of Statistics], Vol. 25, No. 2, 48–56.

5. *Billari, F., D.F. Amuri* & *J. Marcucci*. (2013). Forecasting births using Google. Annual Meeting of the Population Association of America, New Orleans, LA.

6. *Bohon, S.A*. (2018). Demography in the big data revolution: Changing the culture to forge new frontiers. Population Research and Policy Review, March, 1–19.

7. *Ginsberg, J., M.H. Mohebbi, R.S. Patel et al*. (2008). Detecting influenza epidemics using search engine query data. Nature, Vol. 457, No. 7232, 1012–1014.

8. *Kikas, R., M. Dumas & A. Saabas*. (2015). Explaining international migration in the Skype network: The role of social network features. In Proc. SIdEWayS, 17–22.

9. *O'Neil, C*. (2013). We don't need more complicated models, we need to stop lying with our models. Available at: https://mathbabe.org/2013/04/03/we-dont-need-more-complicated-models-we-need-to-stop-lying-with-our-models (date of access: 11.07.2018).

10. *Phithakkitnukoon, S., Z. Smoreda & P. Olivier*. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. PlOS One, 7(6).

11. *Reimsbach-Kounatze, C*. (2015). The Proliferation of «Big Data» and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis. OECD Digital Economy Papers, No. 245. Paris, OECD Publ.

12. *Reis, B.Y. & J.S. Brownstein*. (2010). Measuring the impact of health policies using internet search patterns: the case of abortion. BMC Public Health, Vol. 10, No. 1, 514.

13. *State, B., M. Rodriguez, D. Helbing & E. Zagheni*. (2014). Migration of professionals to the U.S. evidence from LinkedIn data. SocInfo, 531–543.

14. *Tamgno, J.K., R.M. Faye & C. Lishou*. (2013). Verbal autopsies, mobile data collection for monitoring and warning causes of deaths. Advanced Communication Technology (ICACT), 495–501.

15. *Zagheni, E., V.R.K. Garimella, I. Weber et al*. (2014). Inferring international and internal migration patterns from twitter data. In Proc. ACM Conf.

16. *Zagheni, E. & I. Weber*. (2012). You are where you e-mail: using e-mail data to estimate international migration rates. Proc. WebSci, 348–351.

Information about the author

Voronov, Yury Petrovich (Novosibirsk, Russia) – Candidate of Sciences (Economics), Director of OOO Korpus (1, Gorsky microdistrict, Novosibirsk, 630089, Russia, e-mail: corpus-cons@ngs.ru); Leading Researcher at the Institute of Economics and Industrial Engineering, Siberian Branch of the Russian Academy of Sciences (17, Ac. Lavrentiev av., Novosibirsk, 630090, Russia).

Рукопись статьи поступила в редколлегию 01.10.2018 г.