УДК 528.946 + 911.3:314 DOI: 10.21782/GIPR0206-1619-2020-5(203-207)

А.Н. ВОРОБЬЁВ

Институт географии им. В.Б. Сочавы СО РАН,
664033, Иркутск, ул. Улан-Баторская, 1, Россия, Tore12@yandex.ru

БОЛЬШИЕ ДАННЫЕ В ИЗУЧЕНИИ ЛОКАЛИЗАЦИИ И МОБИЛЬНОСТИ НАСЕЛЕНИЯ

Картографирование населения с высоким разрешением изображений облегчает исследование городской среды, в то время как все еще активно обсуждаются такие актуальные информационные проблемы, как слабая пространственная неоднородность альтернативных данных, качество традиционных (как правило, устаревших и неточных) данных о населении территории. Разработана методика составления карт локализации и мобильности населения. Данные, полученные с мобильных телефонов, могут отображаться в режиме реального времени, а также усиливают пространственную неоднородность пользователей. Опираясь на географическое местоположение базовых станций, можно проанализировать плотность пользователей в конкретное время в определенной точке. На основе результатов пространственного анализа данных нами получена плотность сигналов в определенный промежуток времени. Однако пользовательские ареалы, зарегистрированные по мобильному телефону базовыми станциями, не имеют фиксированной пространственной границы. Для решения этой проблемы нами были составлены полигоны Вороного по базовым станциям.

Ключевые слова: *геопространственные данные, мобильный телефон, плотность населения, данные сотовых операторов, местоположение, картографирование.*

A.N. VOROBYEV

V.B. Sochava Institute of Geography, Siberian Branch, Russian Academy of Sciences,
664033, Irkutsk, ul. Ulan-Batorskaya, 1, Russia, Tore12@yandex.ru

BIG DATA IN THE STUDY OF LOCALIZATION AND MOBILITY OF THE POPULATION

Mapping of the population with high resolution of images alleviates the investigation of the urban environment, while currently relevant information problems have been actively discussed, such as the poor spatial inhomogeneity of alternative data, and the quality of traditional (usually outdated and inaccurate) data on the population of the territory. The technique was developed for compiling population localization and mobility. Data received from mobile telephones can be displayed in real time as well as enhancing the spatial inhomogeneity of users. Based on the geographical location of base stations, it is possible to analyze the density of users at a particular time at a particular point. Using results of spatial analysis as a basis, we obtained the density of signals for a specific time interval. However, the geographical range of users recorded via mobile telephone by base stations do not have any fixed spatial boundary. To solve this problem we compiled the Voronoi polygons for the base stations.

Keywords: *geospatial data, mobile telephone, population density, data of cellular operators, location, mapping.*

ВВЕДЕНИЕ

Термин «большие данные» является буквальным переводом английского выражения big data, которое изначально применялось в области информационных технологий, но впоследствии пришло и в другие сферы жизни людей. Известно, что впервые термин big data прозвучал в 1998 г. в презентации Дж. Мэши, главного ученого компании Silicon Graphics. Однако тогда термин не получил широкого распространения, поскольку Дж. Мэши, предсказывая будущий рост данных, обращался к узкому кругу коллег [1]. Популярность словосочетание обрело в 2008 г. после выхода в свет специального выпуска журнала Nature, где были собраны материалы о феномене взрывного роста объемов обрабатываемых данных и технологических перспективах в парадигме вероятного скачка «от количества к качеству»; термин был предложен по аналогии с популярными в деловой англоязычной среде выражениями «большая нефть», «большая руда» [2]. Под «большими данными» подразумевается со-

вокупность технологий по поиску, сбору, анализу, хранению, обработке и т. п. значительных объемов информации. Основные отличия больших данных от других данных можно охарактеризовать как «три V» (volume, velocity, variety) — объем, скорость, многообразие. Часто большие данные включают прямую или косвенную ссылку на местоположение на Земле и могут называться большими геопространственными данными.

Цель исследования — разработка методики применения геоданных для изучения размещения и пространственного перемещения населения при анализе демографических процессов и структур.

МЕТОДИКА ЭКСПЕРИМЕНТА

Геоданные — пространственно-временные данные, отражающие свойства объектов, процессов и явлений, проходящих на Земле. Они содержат информацию о предметах, формах территории и инфраструктурах на поверхности Земли, причем в них должны обязательно присутствовать пространственные отношения [3].

В современном мире множество устройств (телефоны, навигаторы, смарт-часы и др.) осуществляет сбор огромных массивов данных (геоданных). В недалеком прошлом для подобного сбора требовались технически сложные, громоздкие и дорогие устройства, процесс измерения был трудоемким и требовал высокой квалификации специалистов. Однако стремительный рост информационных технологий в бытовых устройствах, таких как смартфоны и др., позволил перейти на качественно новый уровень сбора геоданных. Современные устройства способны получать геопространственную информацию на беспрецедентном уровне в отношении значительно улучшенной точности, временного разрешения и тематической детализации. Мобильные устройства малы, просты в обращении и способны непрерывно собирать информацию о пользователях с помощью датчиков и GPS-приемников даже без ведома абонента. Миллионы абонентов отслеживаются операторами сотовой связи при помощи SIM-карт. Сигнал отправляется на базовую станцию приблизительно 12 раз в час. Из этих сигналов складывается довольно точный маршрут передвижения абонентов. Данной опцией пользователь не управляет, в отличие от GPS, где достаточно отключить функцию геолокации на своем устройстве. Единственный способ не быть запеленгованным — это отказаться от устройств, имеющих SIM-карты.

Возможности датчиков слежения также распространяются на автомобили, фиксирующие расположение транспортного средства, в режиме реального времени. Миллиарды транзакций осуществляются по всему миру при помощи банковских карт и бесконтактных платежей с использованием смартфонов, каждая из которых тоже оставляет свой цифровой след.

Основной особенностью геоданных является точное отражение местоположения абонента в конкретный момент и возможность получить точные координаты, а также вероятность восстановления хронологии передвижений абонента. Появляются практические приложения геопространственных технологий для решения конкретных задач в бизнесе, рекламе, управлении, политике.

Один из примеров применения современных геопространственных технологий — сервис-стартап Geofeedia (США) как аналитическая платформа для социальных сетей, связывающая сообщения в социальных сетях с географическими точками. Большинство других аналитических инструментов оперируют данными, основанными на ключевых словах, а Geofeedia ориентирован на GPS-местоположение. Обнародована информация, что Geofeedia собирал и передавал полиции все сообщения в социальных сетях с геотегами интересующей их территории, что позволяло вести мониторинг массовых протестов [4].

Другим примером сочетания в себе всех современных технологий работы с большими данными и использования человеческих ресурсов является свободный веб-картографический сервис OpenStreetMap. К созданию карты может подключиться любой зарегистрированный пользователь, который становится поставщиком больших данных в виде GPS-треков, аэроснимков, видеозаписей и др. Такие пользователи добровольно, как волонтеры, участвуют в сборе данных. В связи с этим в картографии и геоинформатике возник новый термин — «добровольческая (волонтерская) географическая информация» (Volunteered Geographic Information — VGI)» [5].

В данной работе рассматривается получение данных с использованием мобильного телефона — наиболее распространенный гаджет среди всех возрастных категорий населения. Интерес к использованию таких данных растет стремительно. Быстрое увеличение количества функций, выполняемых мобильными устройствами, также способствует росту объема и ценности геопространственной информации о пользователях, поступающей в распоряжение сотовых операторов.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Для получения информации о локализации и перемещениях пользователей мобильных телефонов можно воспользоваться данными телекоммуникационных компаний. По объему информации о пользователях они могут сегодня на равных конкурировать с социальными сетями. При соответствующей обработке эта информация может составить огромный массив знаний, недоступный иным образом. Проанализировав геопространственные данные, можно получить точную и, что немаловажно, оперативную информацию и понять, какие процессы протекают на городской территории в течение дня. Вся информация аккумулируется в запись сведений о звонке — Call detail record (CDR) [6], куда включены сведения о местоположении устройства в каждый момент времени, журнал звонков, в том числе информация о другом абоненте, и данные о сессиях выхода в интернет. В июле 2016 г. в России был принят так называемый Закон Яровой [7], обязывающий операторов сотовой связи хранить метаданные в течение трех лет. Кроме того, с 1 октября 2018 г. операторы обязаны хранить в течение 30 сут как минимум (но не более шести месяцев) текстовые, голосовые, видео- и другие сообщения пользователей [8].

Использование данных мобильного телефона в качестве альтернативного источника информации о распределении населения значительно, по нашему мнению, повысит точность картографирования. С быстрым развитием информационно-коммуникационных технологий данные мобильных телефонов становятся важным источником для исследования территориального распределения населения и перемещения (маятниковая миграция) городских жителей [9]. Несмотря на явные преимущества применения данных с мобильных телефонов в качестве источника информации для исследования распределения и перемещения населения, в России этот метод не так популярен, как у иностранных коллег. При этом если за рубежом данные сотовых операторов используют уже более 15 лет, то в российской практике исследования на их основе появились относительно недавно. Среди исследований в этом направлении можно выделить работы по Москве и Московской области, проводившиеся компанией Habidatum, КБ «Стрелка», специалистами Высшей школы экономики и МГУ [10].

Между тем огромное количество базовых станций мобильных телефонов с соответствующими данными пользователей предоставляют сведения о пространственной неоднородности пользователей. Однако данные, полученные из традиционных источников (перепись населения, статистические сборники), имеют четкие границы, как правило, административно-территориальные. Данные о населении, связанные с базовыми станциями мобильных телефонов, не имеют жесткой границы зоны обслуживания. Агрегирование полигонов Вороного с базовыми станциями поможет определить границу.

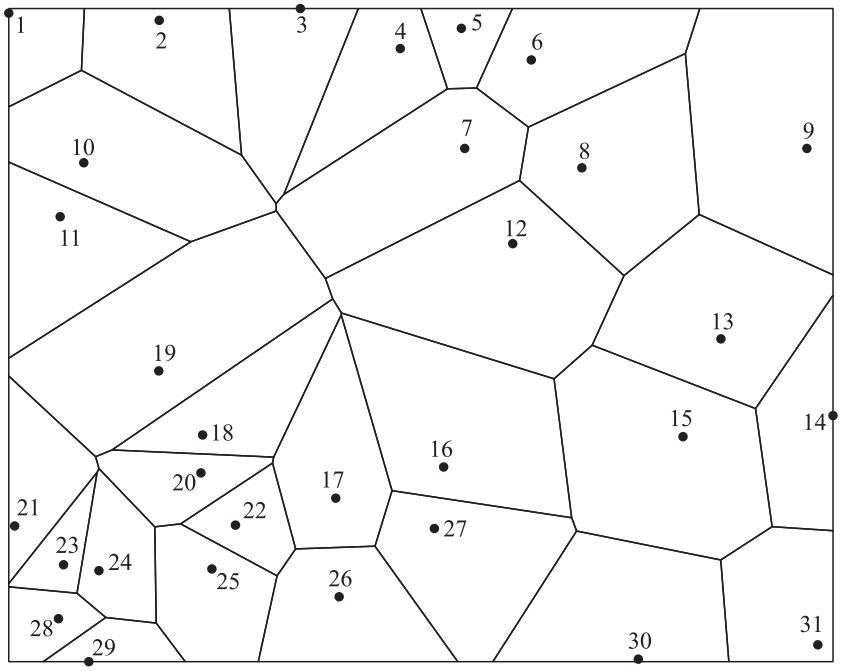
При таких исследованиях основная сложность состоит в получении данных мобильных операторов за определенный период, чтобы отследить динамические эффекты. В России крупнейшие операторы мобильной связи представлены телекоммуникационными компаниями МТС, «Билайн», «Мегафон», «Tele2».

Данные проходят предварительную обработку, чтобы исключить информацию, связанную с конфиденциальностью абонентов. Операторы сотовой связи заявляют, что все данные обезличены и включают в себя только возраст и геолокацию абонента [11].

Точность и трактовка данных мобильных операторов — особый методологический вопрос. Главными уязвимыми сторонами мобильных сведений служит их неперсонализированный характер и связанные с этим проблемы учета мобильных телефонов, зарегистрированных на других лиц (например, на родственников) или людей, не имеющих мобильных телефонов. Кроме того, определенную роль играет недоучет SIM-карт других операторов и выключенных устройств [10].

Основной формат данных — это многопольная таблица, помеченная идентификатором пользователя. В дальнейшем сведения с базовых станций в течение одного месяца (январь–февраль) были разделены на три периода: рабочее время (с понедельника по пятницу) с 7:00 до 19:00, нерабочее — с 7:00 до 7:00, а также выходные дни и праздничные за весь период [12]. Данные за конец января–начало февраля, на наш взгляд, больше всего соответствуют задачам исследования, так как зимние месяцы и отсутствие длительных выходных способствуют более однородному перемещению населения.

Посредством идентификатора пользователя временные и пространственные координаты связаны с базовыми станциями. Полагаясь на географическое местоположение базовых станций, можно проанализировать плотность пользователей в конкретное время в определенной точке. При густом распределении базовых станций в центральных городских районах с плотным населением, особенно в мегаполисах, погрешность местоположения абонента может находиться в пределах нескольких сотен метров. Впоследствии в QGIS (свободная кроссплатформенная геоинформационная система с открытым кодом) наносится пространственное распределение базовых станций и создается реляционная



Построение полигонов Вороного вокруг базовых станций мобильной связи.

Черный пунсон — базовая станция; число — порядковый номер базовой станции; линии — границы полигонов.

таблица различных базовых станций с определенными идентификаторами пользователя. Кроме того, вокруг каждой базовой вышки мы получаем количество пользователей в рабочие и нерабочие дни.

По сравнению с традиционными источниками данных, такими как официальная статистика текущего учета и переписи населения, данные о местонахождении мобильных телефонов обладают очевидным преимуществом в статистической точности и своевременности. Таким образом, данные о местонахождении мобильного телефона способны действительно выявить пространственно-временное распределение городских жителей на макро- и мезоуровнях в административных районах города [13].

В дальнейшем позиции базовой станции обрабатываются в QGIS инструментами пространственного анализа — полигонами Вороного (см. рисунок), представляющими собой разбиение некоторого пространства на области влияния отдельных точек. Полигоны Вороного создаются на основе сети базовых станций мобильной связи, когда количество выходов телефонов в сеть за установленный период (общее время, выходные, рабочие дни) делится на площадь в пределах конкретного полигона. Тем самым мы получаем плотность сигналов на единицу площади в определенный промежуток времени.

ЗАКЛЮЧЕНИЕ

Таким образом, картографирование населения по данным сотовых операторов обладает рядом неоспоримых преимуществ, позволяющих улавливать малейшие изменения в численности и локализации населения, понимать пространственные особенности городов. Исследование наглядно демонстрирует, что развитие геоинформационных систем и технологий дистанционного зондирования в связке с геоданными с пользовательских устройств представляет собой эффективный инструмент для решения географических задач.

Однако при всех своих преимуществах альтернативные исследования изучения мобильности и локализации людей не могут полностью заменить традиционные методы, дающие более глубокое понимание социально-экономических характеристик. Только рациональное сочетание традиционных [14] и новых методов способно повысить качество, точность и оперативность знаний о локализации и мобильности населения.

Исследование выполнено за счет средств государственного задания (№ госрегистрации темы АААА–А17–117041910167–0) и при финансовой поддержке РФФИ в рамках научного проекта № 20–55–44023 Монг_а.

СПИСОК ЛИТЕРАТУРЫ

1. **Черняк Л.** Свежий взгляд на Большие Данные [Электронный ресурс]. — https://www.osp.ru/os/2013/07/13037355 (дата обращения 16.04.2020).
2. **Большие данные** [Электронный ресурс]. — https://ru.wikipedia.org/wiki/ (дата обращения 14.04.2020).
3. **Цветков В.Я., Домницкая Э.В.** Геоданные как основа цифрового моделирования // Современные наукоемкие технологии. — 2008. — № 4. — С. 100–101.
4. **Стартап Geofeedia** позволял полиции США отслеживать протестующих через соцсети [Электронный ресурс]. — https://www.bbc.com/russian/news-37627459 (дата обращения 15.09.2020).
5. **Ныpцов М.В., Ныpцова Т.П.** Большие данные в картографии. Умное картографирование: будущее или технологическое изменение // Изв. вузов. Геодезия и аэрофотосъемка. — 2016. — № 5. — С. 42–45.
6. **Call Detail Record** [Электронный ресурс]: Википедия. Свободная энциклопедия. — URL: https://ru.wikipedia.org/wiki/Call_Detail_Record (дата обращения 28.01.2020).
7. **Закон** (Пакет) Яровой [Электронный ресурс]. — https://ru.wikipedia.org / (дата обращения 10.09.2020).
8. **Федеральный** закон от 6 июля 2016 г. № 374-ФЗ «О внесении изменений в Федеральный закон “О противодействии терроризму” и отдельные законодательные акты Российской Федерации в части установления дополнительных мер противодействия терроризму и обеспечения общественной безопасности» [Электронный ресурс]. — http://www.consultant.ru/cons/cgi/online.cgi?req=doc&base=LAW&n=201078&fld=134&dst=100132&rnd=214990.3492213126493249&#06428611458093945 (дата обращения 10.09.2020).
9. **Бабкин Р.А.** Пространственная динамика Московской агломерации // Географические исследования Сибири и сопредельных территорий. — Иркутск: Изд-во Ин-та географии СО РАН, 2019. — С. 22–25.
10. **Махрова А.Г., Бабкин Р.А., Казаков Э.Э.** Динамика дневного и ночного населения как индикатор структурно-функциональных изменений территории города в зоне влияния Московского центрального кольца с использованием данных операторов сотовой связи // Контуры глобальных трансформаций: политика, экономика, право. — 2020. — Т. 13, № 1. — С. 159–179.
11. **Portela J.N., Alencar M.S.** Cellular Coverage Map as a Voronoi Diagram // Journ. of communication and information systems. — 2008. — N 23 (1). — P. 22–31.
12. **Wenlai Wang, Tao Pei, Jie Chen, Ci Song, Xi Wang, Hua Shu, Ting Ma, Yunyan Du.** Population distributions of age groups and their influencing factors based on mobile phone location data: A case study of Beijing, China // Sustainability. — 2019. — N 11 (24). — P. 1–19.
13. **Gariazzo C., Pelliccioni A.** A multi-city urban population mobility study using mobile phone traffic data // Applied Spatial Analysis and Policy. — 2019. — N 12. — P. 753–771.
14. **Воробьев А.Н.** Картографирование плотности населения редкозаселенного региона (на примере Иркутской области) // Геодезия и картография. — 2019. — № 4. — С. 32–38.

Поступила в редакцию 20.09.2020
После доработки 05.10.2020
Принята к публикации 09.10.2020