КАРТЫ И ГИС В ГЕОГРАФИИ
MAPS AND GIS IN GEOGRAPHY

УДК 911.3, 519.688

P. O. Syomin
Perm State University, Perm, Russia
E-mail: ntsp@ya.ru

MAPPING RUSSIAN SMALL AND MEDIUM-SIZED BUSINESSES USING TAX SERVICE OPEN DATA

This article presents a flexible and reproducible workflow for creating highly detailed maps of small and medium-sized businesses (SMBs) in Russia. Currently, there is a lack of data sources for mapping the Russian economy, and existing official datasets often provide location information in human-readable addresses rather than machine-readable geographical coordinates. The proposed set of tools aims to solve this problem and simplify the mapping process. The study utilizes three open datasets published by the Federal Tax Service, including open dumps of SMB registry, as its data source. To transform the source data into tables and normalize and geocode the addresses, a Python CLI application was developed. Despite the large volume of the initial data archives, the application can run on a regular modern personal computer without the need for cloud computing resources. The workflow's usability is demonstrated by mapping a small subset of data, specifically legal companies in Russia.

Keywords: *spatial data, mapping, geocoding, small and medium-sized business, dataset, big data, Federal Tax Service.*

For citation: Syomin P. O. (2024), Mapping Russian small and medium-sized businesses using tax service open data, *Vestnik Pskovskogo gosudarstvennogo universiteta. Seriya: Estestvennye i fiziko-matematicheskie nauki* [Bulletin of the Pskov State University. Series "Natural and physical and mathematical sciences"], vol. 17, no. 2, pp. 45–64. (In Engl., Russ.).

П. О. Сёмин
Пермский государственный национальный исследовательский университет,
г. Пермь, Россия
E-mail: ntsp@ya.ru

КАРТОГРАФИРОВАНИЕ РОССИЙСКИХ СУБЪЕКТОВ МАЛОГО И СРЕДНЕГО ПРЕДПРИНИМАТЕЛЬСТВА ПО ОТКРЫТЫМ ДАННЫМ НАЛОГОВОЙ СЛУЖБЫ

В статье представлена гибкая и воспроизводимая методика картографирования субъектов малого и среднего предпринимательства (МСП) в России. В настоя-

щий момент не хватает качественных наборов данных для визуализации характеристик экономики страны на карте. Существующие официальные источники содержат информацию о местоположении тех или иных объектов в виде человекочитаемых адресов, в то время как ГИС-платформы требуют машиночитаемые пространственные данные с координатами. Представленный в работе набор инструментов — это попытка решить проблему и упростить подготовку данных для картографирования. В качестве источника сведений используются три набора открытых данных Федеральной налоговой службы, в частности, выгрузки реестра МСП. Для их обработки, нормализации адресов и геокодирования было разработано специальное приложение командной строки на языке программирования Python. Несмотря на большой объём исходных данных, приложение работает на среднем современном компьютере и не требует обращения к облачным вычислительным мощностям. Работоспособность инструментов продемонстрирована путём картографирования небольшой части исходного набора данных — сведений о юридических фирмах.

Ключевые слова: *пространственные данные, картографирование, геокодирование, набор данных, большие данные, Федеральная налоговая служба.*

Для цитирования: *Сёмин П. О.* Картографирование российских субъектов малого и среднего предпринимательства по открытым данным налоговой службы // Вестник Псковского государственного университета. Серия: Естественные и физико-математические науки. 2024. Т. 17. № 2. С. 45–64.

Introduction. Contemporary mapping solutions used for scientific and other purposes usually require sources of well-structured spatial data. “Spatial” refers to data that is provided in an appropriate format with feature coordinates provided in one of the standardized coordinate reference systems. Obtaining datasets suitable for the mapping can be a challenging task that may significantly slow down research or increase its cost. The public demand for high-quality spatial data has led many governments to create geoportals or to facilitate the distribution of such data through other means [17].

In Russia, official data is occasionally published in machine-readable formats with coordinates, but generally, open spatial data is not widely available. This lack of availability may pose a challenge for mapping various aspects of society, including the economy, demography, and culture. Datasets often include a geographical reference in the form of a human-readable address. However, GIS software requires machine-readable geographical coordinates. To convert addresses into coordinates, a process called geocoding is used. Several geocoders have relatively high accuracy for Russian addresses[1]. However, most of these tools are only available through paid online APIs or have legal terms that prohibit the storage and use of geocoding results. Therefore, researchers attempting to map official data often have to pay for a geocoding API and write a program to convert addresses to coordinates. It would be much better to have a user-friendly tool to assist in mapping valuable open datasets.

[1]Examples are Yandex or Google geocoders, DaData geocoding service, and OpenStreetMap Nominatim tool.

The data from the small and medium-sized businesses (SMB) registry, among various official datasets published by Russian state bodies, is of high value for researchers in economic geography. The state SMB registry was created at the end of 2015 in accordance with Art. 4.1 of Federal Law of July 24, 2007 No 209-FZ "On the development of small and medium-sized businesses in the Russian Federation". The SMB registry is operated by the Federal Tax Service (FTS) of Russia and includes information about all Russian SMBs that meet the inclusion criteria. These criteria include limits on revenue and the number of employees, and all eligible organizations and individual entrepreneurs are automatically included in the registry without any action required on their part.

The SMB registry seems to be the most comprehensive openly available collection of Russian commercial companies. Other sources of such data, e. g. business intelligence systems or the web portal of the Unified State Registry of Legal Entities, are either paid or have usage restrictions. The full dumps of the SMB registry are published as open data free of charge, making them a valuable data source for detailed high-resolution mapping of Russian business entities. However, the distributed data only contain human-readable location information and do not include geographical coordinates. Another problem is that these dumps have a large volume, reaching hundreds of gigabytes when archived and several terabytes when unpacked. Additionally, their structure is complex.

Objective. The purpose of this work is to present a flexible and reproducible workflow for mapping Russian SMB using FTS open data. Reproducibility in this context refers to the development and publication of an open-source application for processing and geocoding FTS data, which can be used by any researcher. The paper provides a limited example of SMB mapping, focusing only on legal companies. However, the workflow's flexibility allows for the mapping of any SMB data.

Literature review. A common source of data for economic geography studies in Russia is the resources provided by the Federal State Statistics Service (Rosstat) [4; 11; 14–16]. These resources include official statistics, censuses, and statistical surveys available on the official website of Rosstat, as well as the database of municipality indicators and the unified interdepartmental information and statistical system. These data sources are used in various studies. Rosstat is a crucial source of data in economic geography research, and its data is used explicitly or implicitly in almost all studies. This data is not suitable for cartography due to the lack of coordinates. However, it does provide information about location in terms of regions, municipalities, and cities. This information can be combined with spatial data from other sources, such as the popular GADM or Natural Earth databases, to create a map.

Economic geographers may also use administrative data from the Federal Tax Service of Russia as an alternative source of information. It is published on the official website of this authority or its satellite web resources. Geographers use, in particular, information on the income of individuals [11], on tax revenues by regions or types of economic activity [13; 15], on the average number of employees of organizations [12], on the number of SMBs and measures to support them [1]. FTS of Russia is not a specialized statistical body, but it collects a vast amount of information about various aspects of the Russian economy, including business entities, property, income, and expenses. The geographical location in FTS data is described with address elements rather than coordinates, similar to Rosstat. Additionally, it is worth noting that FTS of Russia operates the Federal Address Information System, which provides open address data widely used by government and business.

Researchers in economic geography have traditionally relied on aggregated data. However, modern studies often use information disaggregated down to the level of individual firms. This disaggregated data can be used directly [3; 12] or to calculate derived metrics that are not present in Rosstat statistics [1; 7; 14]. In city-scale studies [2; 6; 9], information about specific firms holds additional importance. Similarly, in the analysis of narrow sectors of the economy that are not separately represented in aggregated statistics, obtaining such detailed data is of key importance [8; 10]. Getting such detailed data is difficult. Researchers typically use commercial or open geoinformation services, as well as business intelligence systems, to obtain data for analysis [5; 8]. However, these methods often present two challenges: either the data is provided for a fee, or significant effort is required to format it for analysis. Business intelligence systems, such as Spark or Kontur.Focus, accumulate information about various aspects of a company's operations. However, they are paid services. Public mapping platforms, such as 2GIS and Yandex.Maps, generally prohibit the downloading of data. Open platforms, like OpenStreetMap or the SMBs registry, are free, but the data dumps from them are large and complex. Converting the gathered data into a ready-to-map format requires a lot of time and effort.

Data. The mapping procedure uses a primary data source together with two additional data sources and two lookup tables.

The primary data source is the SMB registry open data dumps[2], which are published every month starting from August 2016. Each dump contains the full registry for the date of publication, resulting in a significant amount of duplicated data. Technically, a dump is a ZIP archive containing thousands of XML files, with each file including information about 900 organizations or individuals. The XML file's data attributes contain, in particular, the taxpayer number, business name, and registration address (region, district, city, settlement), as well as activity codes based on the All-Russia Classifier of Economic Activity Kinds.

Two FTS open datasets serve as additional data sources. The first of them is called "Information about revenue and expenditure of organizations by their accounting (financial) reporting"[3], and the second is named "Information about the average list count of employees of organizations"[4]. Both sources have been available since approximately 2019 and are updated at least once a year, although updates may be irregular. Technically, these datasets are collections of ZIP archives containing XML files. Each archive contains data for a specific date, and each XML file lists 900 organizations. Information about individuals is not included in these datasets due to personal data issues.

[2] Unified registry of small and medium-sized businesses. [Electronic resource]: URL: https://www.nalog.gov.ru/opendata/7707329152-rsmp/ (access date: 29.02.2024).

[3]Information about revenue and expenditure of organizations by their accounting (financial) reporting. [Electronic resource]: URL: https://www.nalog.gov.ru/opendata/7707329152-revexp/ (access date: 29.02.2024).

[4]Information about the average list count of employees of organizations. [Electronic resource]: URL: https://www.nalog.gov.ru/opendata/7707329152-sshr2019/ (access date: 29.02.2024).

The lookup tables used for the mapping are the “Settlements of Russia: population and geographic coordinates” dataset provided by the “Infrastructure of scientific data” project[5] and the “Cities of Russia” dataset published by the commercial company HFLabs[6]. Both datasets contain addresses and geographical coordinates of Russian cities and settlements. Additionally, the author added a small supplement to the ‘Cities of Russia’ dataset. This supplement includes information about missing cities that was manually extracted from the Federal Address Information System[7].

A Python command-line (CLI) application was developed to process the source data. Technically, it is based on the Python standard library and several third-party packages, including the fast XML processing library called *lxml*, the iconic tabular data processing tool named *Pandas*, the popular big data framework *Apache Spark*, and the easy-to-use CLI framework *Typer*. This application was intended to be published as an open-source tool for other researchers, so its source code with brief documentation is available at GitHub public repository[8]. Readers who are familiar with Python programming language may go to the repository and look at the details of the code. This section describes the top-level structure of the application, and the data processing flow for the mapping is nearly identical.

Step 1 is data download. The source ZIP archives were downloaded from the FTS website and stored locally.

Step 2 is data extraction and filtering. The data was extracted from ZIP archives and stored in CSV tables so that one ZIP archive was transformed into 1 CSV table. A lot of unused data attributes were dropped, so the resulting size of the CSV tables is much smaller than the size of the original archives. In addition, filtering by activity code took place in this step to select data only about legal companies. According to the classifier, the main activity code equal to 69.10 (activity in the area of law) was used as a filtering criterion. The application itself allows to filter by other codes or groups of codes or to disable filtering at all, thus providing the opportunity to generate various slices of companies (e.g. agricultural, forestry, health services, heavy industries, etc) with various structural resolution (i. e. up to groups or particular low-level codes of classifier). Also, in this step, the additional source datasets are extracted in the same fashion.

Step 3 is data aggregation: due to the specifics of data publication, a lot of information in archives and corresponding CSV tables is duplicated, and thus duplicates have to be dropped. This step dropped the duplicates and thus reduced the volume of data even further. Also, the additional data on revenue, expenditure, and employees was deduplicated and filtered to leave only the rows with organizations present in the main data. Taxpayer identification number (TIN) was used to find the necessary companies because it acts as a unique persistent company identifier.

---

[5] Settlements of Russia: population and geographic coordinates. [Electronic resource]: URL: https://data.rcsi.science/data-catalog/datasets/160/ (access date: 29.02.2024).

[6] Cities of Russia [Electronic resource]: URL: https://github.com/hflabs/city (access date: 29.02.2024).

[7] Federal Address Information System. [Electronic resource]: URL: https://fias.ru (access date: 29.02.2024).

[8] PavelSyomin/russian-smb-companies: Dataset on small and medium business companies registered in Russia based on Federal Tax Service open data. [Electronic resource]: URL: https://github.com/PavelSyomin/russian-smb-companies (access date: 29.02.2024).

Step 4 is geocoding. The addresses were normalized and converted to geographical coordinates using the look-up tables, and ≈ 98 % of all addresses were successfully matched. Because the address in the source data is detailed down to cities or settlements, the geographical coordinates refer to cities or settlements, and this is the lower bound of spatial resolution of the map. No action is performed with additional datasets in this step.

Optional *step 5* combines SMB registry data with data on revenue, expenditure, and employees and transforms it to produce a yearly panel table. Here, it was not performed, because the geocoded dataset was sufficient for mapping.

After this preprocessing, the tabular dataset was filtered by year (2021 was chosen). The maps were plotted with a script written in R programming language with additional packages, including, in particular, ggplot2 and sf. The source code for the paper is available in a separate repository[9].

Results. The spatial resolution of the dataset is up to cities or settlements, but regions and their subdivisions are also included, so aggregation and mapping are also possible. Figure 1 shows the count of legal companies by region. Some spatial patterns are seen from this visualization, for instance, the high concentration of legal firms in the capital of Russia and the surrounding region, the "Urals-Volga-Krasnodar" belt, and small values in the Central economic region.

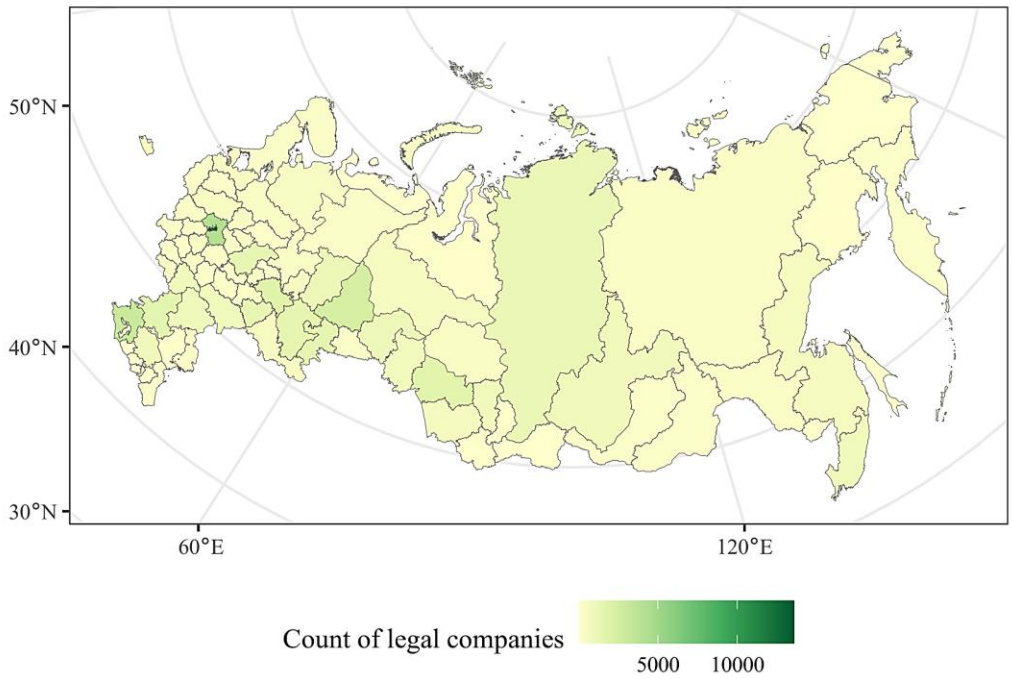

Fig. 1. Count of legal companies by region (constituent entity) of Russia

[9][Electronic resource]: URL: https://github.com/PavelSyomin/ru-smb-companies-papers/tree/main/legal-companies-mapping (access date: 29.02.2024).

Figure 2 displays the mapping of legal companies by cities and rural settlements. This high-resolution map provides a quick overview of legal business in Russia. The spatial distribution resembles the region-scale map, but city-level data allows us to notice other tendencies, e. g. the concentration of businesses in regional centres. Also, the high concentration in Moscow and its surroundings is clear, and a similar pattern for St. Petersburg is visible. A map may be also drawn for a particular region (Sverdlovsk oblast in Figure 3).

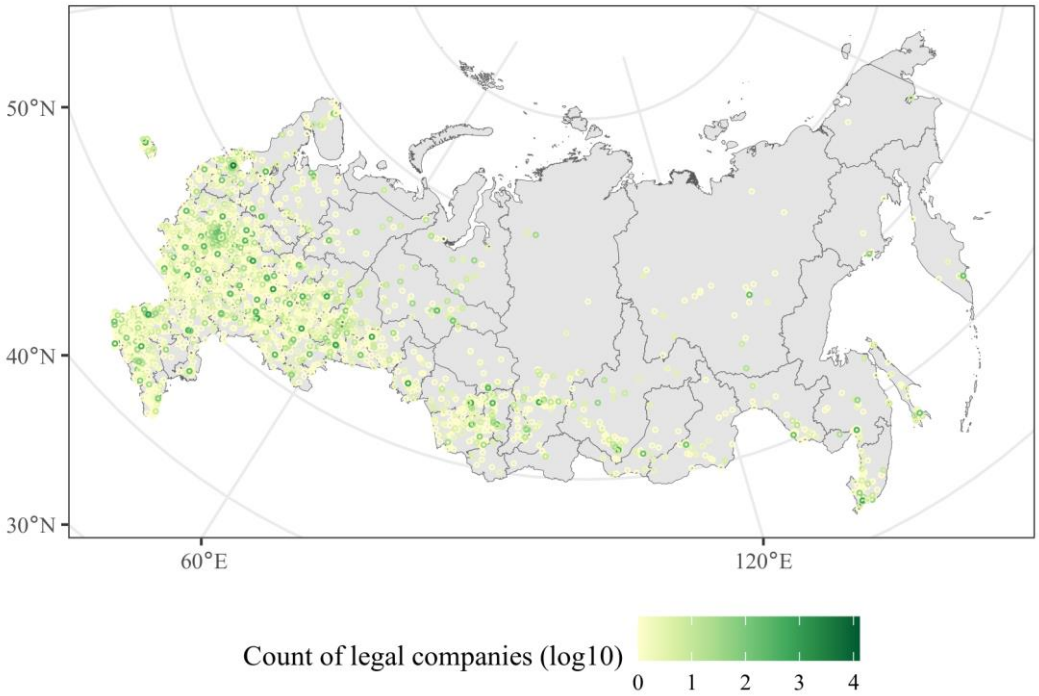

Fig. 2. Count of legal companies in Russian urban and rural settlements

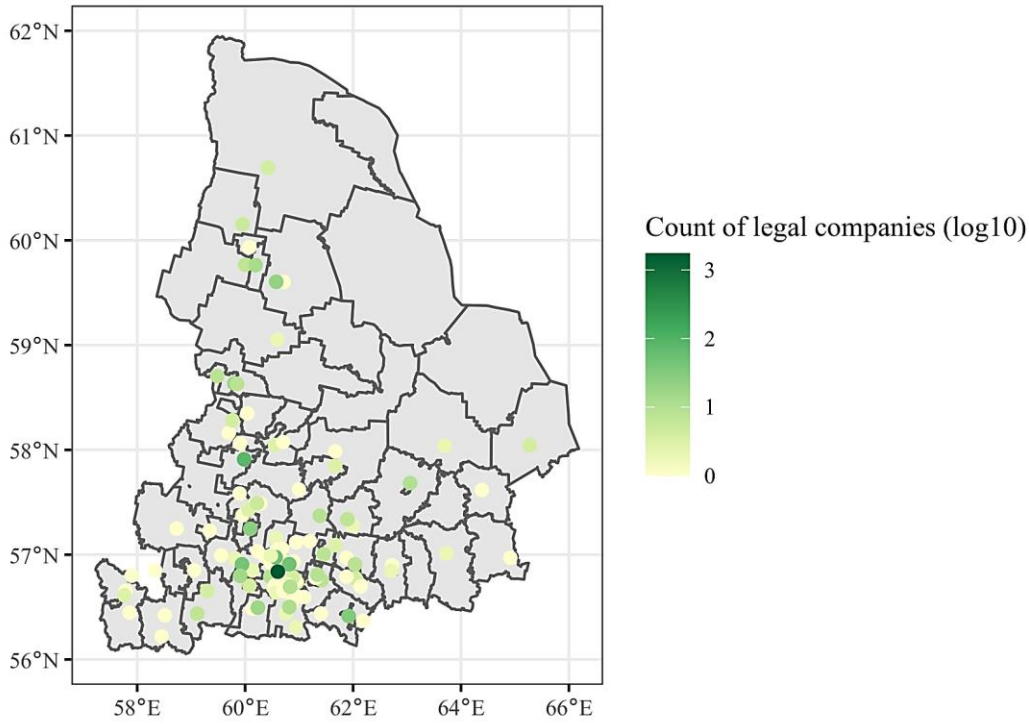

Fig. 3. Count of legal companies in urban and rural settlements of Sverdlovsk oblast (the Middle Urals)

The presence of additional data about financial metrics and count of employees allows us to draw more maps. In Figure 4, the spatial distribution of legal companies' profit is displayed, and in Figure 5, a similar distribution concerning the number of employees is shown. Both maps include only companies with non-zero profit or number of employees. The maps look like the counts map (see Figure 2), but exact spatial patterns vary.

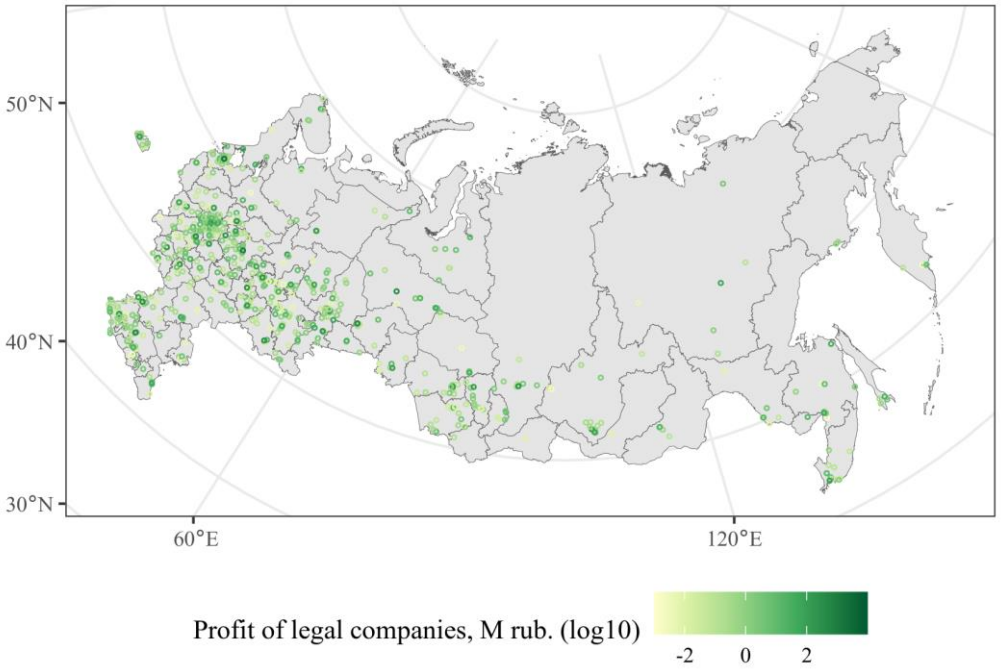

Fig. 4. Profit of legal companies by urban and rural settlements in Russia

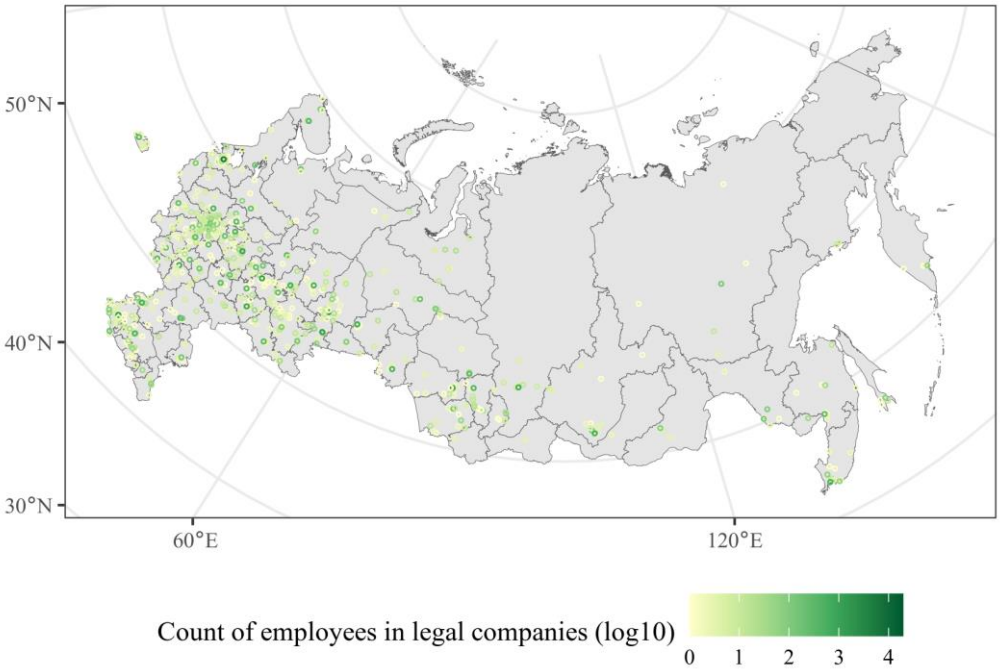

Fig. 5. Number of employees in legal companies by urban and rural settlements in Russia

Conclusion. In this paper, a flexible solution for high-resolution mapping of small and medium-sized companies in Russia is proposed. It is based on open data provided by Federal Tax Service which is preprocessed and geocoded with a Python CLI application to make a ready-to-mapping tabular spatial dataset. The example plots of legal companies and their business metrics (profit and number of employees) are provided to prove the usability of the suggested method. The code of the application used to make the dataset and the code of scripts used to draw maps are made open source and published in public repositories on GitHub. Despite the large amount of source data (up to several terabytes if unpacked), the tools used for the data processing and mapping are optimized to work on a regular modern PC and do not require to purchase cloud computing resources. Also, the geocoding is performed with simple look-up tables instead of commercial APIs, thus the entire workflow is free of charge and offline.

There are several drawbacks and limitations of the described approach. First of all, even though the tools are designed to be as simple as possible, they still require some basic understanding of computer programming, because a researcher has to launch a Python CLI application and thus must be able to install Python and work with a command line. Next, although many optimization efforts were made, the generation of the dataset (mainly, the extraction of data) still takes a considerable time (up to one or two days if no filtering is used). Third, the accuracy of geocoding is high but does not equal 100 %. Finally, the source data is limited to small and medium-sized companies only, thus any insights based on it must be done with caution. Nevertheless, the advised method allows making the most comprehensive mapping of Russian businesses based entirely on open free of charge data and algorithms. It may be especially useful for studying the SMB sector as is and for analysis of narrow economic areas that do not fall within the regular statistical reporting.

Введение. Современные картографические инструменты, используемые в науке и практике, обычно предполагают использование источников хорошо структурированных пространственных данных. Под «пространственными» в данном случае понимаются такие данные, которые представлены в одном из распространённых и широко используемых форматов и содержат географические координаты объектов в одной из стандартных систем координат (например, WGS84). Получение таких наборов данных может быть трудной задачей, которая существенно замедлит процесс исследования или увеличит его стоимость. Запрос со стороны академического сообщества и бизнес-кругов на качественные пространственные данные настолько важен, что понемногу стимулирует правительства разных стран создавать геопорталы или упрощать распространение геоданных иными способами [17].

В России официальная информация время от времени публикуется в машиночитаемом виде с координатами, но как правило, открытые геоданные доступны весьма плохо. Относительно слабая доступность таких данных может создавать проблемы при картографировании различных общественных процессов: экономических, географических, культурных. Даже если набор государственных данных содержит географическую привязку в виде «обычного», человекочитаемого адреса, его не по-

лучится использовать напрямую для создания карты, так как ГИС-приложениям необходимы машиночитаемые координаты. Для преобразования адресов в координаты служит процедура, известная как геокодирование (geocoding). Для российских адресов обычно используется несколько популярных геокодеров: Geocoding API от Яндекса или Google, сервис геокодирования компании DaData, инструмент OpenStreetMap Nominatim. Они обеспечивают сравнительно хорошее качество, но обычно предоставляются за плату, а в бесплатном варианте сложны в настройке или запрещают сохранять получаемые данные для последующего использования. Соответственно, актуален вопрос об удобном инструменте, который позволил бы облегчить картографирование ценных наборов открытых государственных данных.

Одним из наборов открытых государственных данных в России, который потенциально имеет высокую ценность для исследователей в сфере экономической географии, может являться реестр субъектов малого и среднего предпринимательства (МСП). Указанный реестр ведётся с конца 2015 г. на основании ст. 4.1 Федерального закона от 24.07.2007 №209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации». Оператор реестра — Федеральная налоговая служба (ФНС России). Реестр автоматически включает в себя сведения обо всех коммерческих юридических лицах и индивидуальных предпринимателях, которые соответствуют установленным критериям. Основных критериев два: среднесписочная численность работников не превышает 250 чел., а доход за год — 2 млрд руб.

Реестр МСП представляет собой наиболее полный общедоступный источник дезагрегированных сведений о российских коммерческих организациях. Другие варианты, такие как системы бизнес-аналитики (например, Спарк-Интерфакс или Контур.Фокус) или веб-портал Единого государственного реестра юридических лиц, доступны за плату или ограничивают бесплатное использование, что уменьшает их полезность. Данные официальной статистики бесплатны, но предоставляются в сводном (агрегированном) виде, т. е. без детализации до уровня отдельных фирм. Напротив, полные выгрузки реестра МСП ежемесячно размещаются в разделе открытых данных ФНС России, доступны бесплатно и детализированы до конкретных организаций и предпринимателей. Проблема в том, что координат в них нет — только адреса. Кроме того, выгрузки имеют большой объём (суммарно несколько сотен гигабайт в сжатом виде и более двух терабайт в несжатом) и сложную структуру.

Цель статьи. Данная работа направлена на то, чтобы представить гибкую и воспроизводимую методику картографирования российских малых и средних фирм на основе открытых данных ФНС России. Воспроизводимость достигается за счёт разработки и публикации открытого приложения для обработки и геокодирования данных налоговой службы, которое может быть использовано исследователями для своих задач. В статье приводится пример картографирования российских юридических фирм на основе описанной методики. Аналогичный подход может использоваться для картографирования юридических лиц и индивидуальных предпринимателей в любой сфере деятельности.

Обзор литературы. Типичным источником данных для исследований по экономической географии России являются ресурсы Федеральной службы государственной статистики (Росстат) [4; 11; 14–16]. К ним относятся официальная статистика,

переписи населения и статистические обследования, доступные на официальном сайте этого государственного органа, а также база данных показателей муниципальных образований (БДМО) и единая межведомственная информационно-статистическая система (ЕМИСС). Росстат является важнейшим провайдером информации для исследований в области экономической географии, и его данные явно или неявно используются почти во всех научных работах. Эти данные не пригодны для картографирования напрямую из-за отсутствия координат. Тем не менее, они содержат информацию о местоположении с точностью до регионов, муниципалитетов и городов, которую для создания карты можно объединить с пространственными данными из других источников, таких как GADM или Natural Earth.

Экономико-географы используют административные данные ФНС России в качестве альтернативного источника информации. Эти данные публикуется на официальном сайте указанного органа власти или связанных с ним веб-ресурсах. В исследованиях применяются, в частности, сведения о доходах физических лиц [11], о налоговых поступлениях по регионам или видам экономической деятельности [13; 15], о среднесписочной численности работников организаций [12], о числе субъектов малого и среднего предпринимательства и мерах их поддержки [1]. ФНС России не является специализированным статистическим органом, но собирает огромный объём информации о различных аспектах российской экономики, в т. ч. о хозяйствующих субъектах, имуществе, доходах и расходах. Географическое положение в данных ФНС, как и у Росстата, описывается с помощью адресных элементов. Стоит отметить, что ФНС России поддерживает Федеральную информационную адресную систему (ФИАС), предоставляющую официальные открытые адресные данные по России, которые широко используются другими государственными органами, а также бизнесом.

В современных исследованиях географы часто оперируют информацией, детализированной до уровня отдельных фирм. Эти дезагрегированные данные можно использовать напрямую [3; 12] или для расчета производных показателей, которых нет в статистике Росстата [1; 7; 14]. В исследованиях на уровне городов [2; 6; 9] информация о конкретных фирмах имеет особую важность. Такая же детализированность крайне ценна при анализе узких секторов экономики, которые отдельно не представлены в агрегированной статистике [8; 10]. Для получения подобных данных исследователи обычно используют коммерческие или открытые геоинформационные сервисы, а также системы бизнес-аналитики [5; 8]. Однако эти источники характеризуются двумя недостатками: они либо платные, либо сложны для использования. Так, системы бизнес-аналитики, такие как Спарк или Контур.Фокус, являются коммерческими, хотя и содержат огромный набор сведений о различных аспектах деятельности компаний. Общедоступные картографические сервисы наподобие 2ГИС и Яндекс.Карт позволяют выгружать данные только за плату, а автоматизированный сбор сведений, доступных бесплатно (так называемый парсинг, или скрейпинг), запрещён их пользовательскими соглашениями. Открытые платформы, такие как OpenStreetMap, бесплатны, но выгрузки (дампы) их данных имеют большой объём и сложную структуру — преобразование в удобный формат требует много времени и сил.

Данные. Картографирование по методике, предложенной в этой статье, предполагает использование одного основного источника данных вместе с двумя дополнительными, а также двух справочных таблиц.

Основным источником данных являются выгрузки (дампы) реестра МСП[10], которые публикуются каждый месяц, начиная с августа 2016 г. Каждая выгрузка содержит полную информацию реестра на дату публикации, что приводит к существенному дублированию. Технически выгрузка представляет собой ZIP-архив, содержащий тысячи XML-файлов, каждый из которых содержит информацию о 900 юридических лицах или индивидуальных предпринимателях. Атрибуты XML-файла включают, в частности, номер налогоплательщика (ИНН), наименование юридического лица или имя физического лица, адрес регистрации (область, район, город, населённый пункт), а также коды видов деятельности (основного и дополнительных) по Общероссийскому классификатору видов экономической деятельности (ОКВЭД) в редакции 2014 г.

Дополнительными источниками информации выступают два других набора открытых данных ФНС России. Первый из них называется «Сведения о суммах доходов и расходов по данным бухгалтерской (финансовой) отчётности организации»[11], а второй — «Сведения о среднесписочной численности работников организации»[12]. Оба доступны приблизительно с 2019 г. и обновляются не реже одного раза в год, хотя обновления могут быть нерегулярными. Технически эти наборы данных тоже представляют собой коллекции ZIP-архивов, содержащих файлы XML. Каждый архив содержит данные за определённую дату, а в каждом XML-файле представлены сведения для 900 юридических лиц. Информация об индивидуальных предпринимателях не включена в эти наборы из-за необходимости защиты персональных данных.

Справочными таблицами, используемыми для картографирования, являются набор данных «Населённые пункты России: число жителей и географические координаты», предоставленный проектом «Инфраструктура научно-исследовательских данных»[13] и набор данных «Города России», опубликованный коммерческой компанией HFLabs[14]. Оба содержат адреса и географические координаты российских городов и населённых пунктов. Поскольку информация о некоторых городах в них отсут-

[10]Единый реестр субъектов малого и среднего предпринимательства. [Электронный ресурс]: URL: https://www.nalog.gov.ru/opendata/7707329152-rsmp/ (дата обращения: 29.02.2024).

[11]Сведения о суммах доходов и расходов по данным бухгалтерской (финансовой) отчетности организации. [Электронный ресурс]: URL: https://www.nalog.gov.ru/opendata/7707329152-revexp/ (дата обращения: 29.02.2024).

[12]Сведения о среднесписочной численности работников организации. [Электронный ресурс]: URL: https://www.nalog.gov.ru/opendata/7707329152-sshr2019/ (дата обращения: 29.02.2024).

[13] Населённые пункты России: численность населения и географические координаты. [Электронный ресурс]: URL: (дата обращения: 29.02.2024) https://data.rcsi.science/data-catalog/datasets/160/.

[14]Города России. [Электронный ресурс]: URL: https://github.com/hflabs/city (дата обращения: 29.02.2024).

ствовала, автор статьи добавил её вручную на основе сведений Федеральной информационной адресной системы[15].

Для обработки исходных данных было разработано приложение командной строки (CLI) на языке программирования Python. Технически оно основано на стандартной библиотеке этого языка и нескольких сторонних компонентах, включая быструю библиотеку обработки XML *lxml*, широко известное средство для работы с табличными данными *Pandas*, популярную платформу больших данных *Apache Spark* и простой в использовании фреймворк консольных приложений *Typer*. Исходный код приложения с краткой документацией размещён в общедоступном репозитории на платформе GitHub[16]. Далее в статье описывается принципиальный (т. е. без погружения в детали) алгоритм обработки данных, реализованный в приложении. «Входом» этого алгоритма являются исходные наборы данных ФНС России (справочные таблицы де-факто являются частью приложения). На «выходе» получаются таблицы с геокодированными сведениями, пригодные для использования в любом ГИС-приложении, поддерживающем CSV-файлы в качестве формата данных для векторных слоёв.

Шаг 1 — загрузка данных. Исходные ZIP-архивы загружаются с сайта ФНС России и сохраняются локально.

Шаг 2 — извлечение и фильтрация данных. Данные реестра МСП извлекаются из ZIP-архивов и сохраняются в таблицах в формате CSV, так что один ZIP-архив преобразуется в одну CSV-таблицу. Многие неиспользуемые атрибуты отбрасываются, поэтому размер CSV-таблиц получается намного меньше размера исходных архивов. Кроме того, на этом этапе опционально осуществляется фильтрация по коду основного вида деятельности: это ускоряет процесс и дополнительно уменьшает объём CSV-таблиц. Например, для подготовки демонстрационного примера о юридических фирмах был выбран код основного вида деятельности 69.10 (деятельность в области права по ОКВЭД). Аналогичным образом преобразуются два вспомогательных набора данных, только без фильтрации по ОКВЭД, потому что кодов вида деятельности в них нет.

Шаг 3 — агрегация данных: она позволяет удалить дублирующиеся записи и собрать сведения из множества CSV-таблиц в одну. Дублирование происходит из-за того, что каждая выгрузка реестра МСП — это сведения реестра целиком, а не только лишь обновления ранее размещённой информации. Кроме того, в ходе агрегирования вспомогательные наборы данных фильтруются так, чтобы в них остались сведения только о тех фирмах, которые были отобраны по коду ОКВЭД на шаге 2. В качестве уникального идентификатора компании при этом используется ИНН.

Шаг 4 — геокодирование. Адрес нормализуются и преобразуются в географические координаты с использованием справочных таблиц. Полнота геокодирования

[15]Федеральная информационная адресная система. [Электронный ресурс]: URL: https://fias.ru (дата обращения: 29.02.2024).

[16] PavelSyomin/russian-smb-companies: Dataset on small and medium business companies registered in Russia based on Federal Tax Service open data. [Электронный ресурс]: URL: https://github.com/PavelSyomin/russian-smb-companies (дата обращения: 29.02.2024).

— около 98%. Поскольку в исходных данных адрес детализирован до городов или населенных пунктов, географические координаты также относятся к городам или населенным пунктам.

Опциональный *шаг 5* — объединение и трансформация результатов, полученных на шагах 3–4, в панельную таблицу. В панельной таблице сведения о каждой фирме за каждый год представлены одной строкой. Таблицы, которые производятся на шагах 3 и 4, построены по-другому: в них для каждой строки указаны даты актуальности сведений, что позволяет добиться большей компактности, но менее удобно для анализа.

Для подготовки демонстрационного примера набор табличных данных, полученный с помощью приложения, был отфильтрован по году (выбран 2021 г.). Карты создавались с помощью скрипта, написанного на языке программирования R, с дополнительными пакетами, включая, в частности, ggplot2 и sf. Исходный код статьи доступен в отдельном репозитории[17].

Результаты. Хотя данные детализированы до конкретных фирм, они также содержат информацию о местонахождении с указанием как региона (субъекта федерации), так и населённого пункта, поэтому их можно агрегировать. На рисунке 1 показано количество юридических фирм по регионам. Из этой визуализации видны некоторые пространственные закономерности, например: с одной стороны, концентрация юридических фирм в Москве и Московской области, а также в поясе, протянувшемся от Урала через Поволжье к Краснодару, а с другой стороны, сравнительно небольшое их число, например, в субъектах, образующих Центральный экономический район, что может быть связано с близостью столицы, оттягивающей на себя большинство компаний.

Рис. 1. Число юридических фирм по регионам (субъектам федерации) в России в 2021 г.

[17] [Электронный ресурс]: URL: https://github.com/PavelSyomin/ru-smb-companies-papers/tree/main/legal-companies-mapping (дата обращения: 29.02.2024).

На рисунке 2 представлена карта числа юридических фирм по городам и сельским населённым пунктам. По сути это выполненный с довольно высокой детализацией (до населённых пунктов) обзор юридического бизнеса в России. Карта похожа на региональную (см. рис. 1), однако данные на уровне городов позволяют заметить и другие тенденции, например, концентрацию бизнеса в региональных центрах. Кроме того, очевидно высокое сосредоточение в Москве и её окрестностях. Аналогичная картина просматривается в С.-Петербурге. Подобную карту можно построить не только для страны в целом, но и для отдельного региона (Свердловская область на рис. 3).

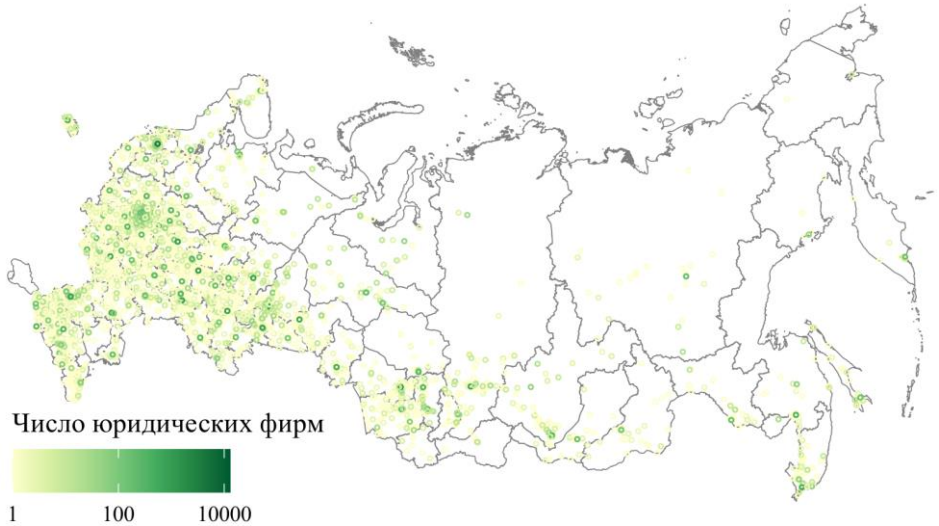

Рис. 2. Число юридических фирм в городских и сельских населённых пунктах России в 2021 г.

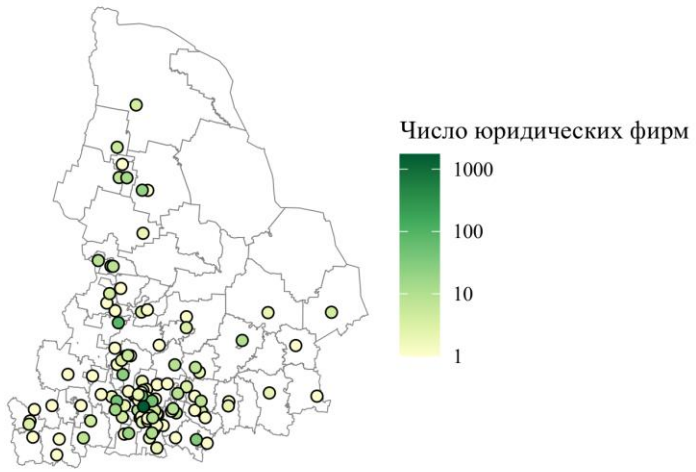

Рис. 3. Число юридических фирм в городских и сельских населённых пунктах Свердловской области в 2021 г.

Наличие дополнительных данных о финансовых результатах и числе сотрудников позволяет построить дополнительные карты. На рисунке 4 показано пространственное распределение прибыли юридических компаний, а на рисунке 5 — аналогичное распределение по численности сотрудников. Обе карты включают только компании с ненулевой прибылью или количеством сотрудников. Карты похожи на ту, что изображена на рисунке 2, но отличаются в деталях.

Рис. 4. Суммарная прибыль юридических фирм по городам и сельским населённым пунктам в России в 2021 г.

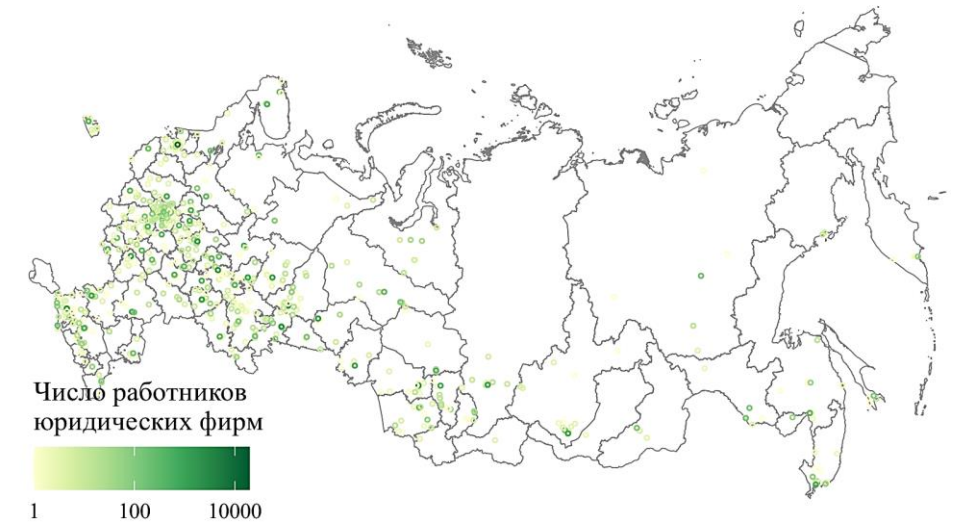

Рис. 5. Суммарное число работников юридических фирм по городам и сельским населённым пунктам в России в 2021 г.

Заключение. В данной статье предлагается гибкое решение для картографирования субъектов малого и среднего предпринимательства в России. Оно основано на использовании открытых данных, публикуемых Федеральной налоговой службой, которые предварительно обрабатываются и геокодируются с помощью специально разработанного приложения с открытым исходным кодом на языке программирования Python. В качестве примера приведены карты юридических фирм и их бизнес-показателей (прибыль и количество сотрудников), подтверждающие работоспособность предлагаемого метода. Код приложения, используемого для подготовки набора данных, и код скриптов, используемых для создания карт, опубликованы в общедоступных репозиториях на GitHub. Несмотря на большой объём исходных данных (до нескольких терабайт в распакованном виде), инструменты оптимизированы для работы на обычном современном компьютере средней конфигурации и не требуют приобретения облачных вычислительных ресурсов. Кроме того, геокодирование выполняется с помощью простых справочных таблиц вместо коммерческих API, поэтому весь процесс осуществляется бесплатно и в автономном режиме.

Описанный подход имеет ряд недостатков и ограничений. Прежде всего, несмотря на то, что инструменты сделаны максимально простыми, они всё равно требуют некоторых базовых знаний в области программирования, поскольку исследователь должен запустить приложение на Python и, следовательно, должен уметь устанавливать Python и работать с командной строкой. Далее, генерация набора данных, несмотря на усилия по оптимизации, по-прежнему занимает значительное время (до одного-двух дней, если не используется фильтрация). В-третьих, точность геокодирования хотя и высока, но не равна 100 %. Наконец, исходные данные ограничены только субъектами малого и среднего предпринимательства, поэтому любые выводы, основанные на них, следует делать с осторожностью. Тем не менее, предложенный метод позволяет осуществить наиболее полное картографирование российского бизнеса, основанное на открытых данных и алгоритмах. Это может быть особенно полезно для разведочного анализа, для изучения малого и среднего предпринимательства как такового и для анализа узких секторов экономики, которые не попадают в обычную статистическую отчётность.

Литература

1. *Гуменюк И. С.* К вопросу о динамике экономической активности и ее влиянии на бюджетную устойчивость муниципальных образований Калининградской области // Вестник БФУ им. И. Канта: Естественные и медицинские науки. 2022. № 1. С. 44–56.

2. *Кожевников С. А.* Модернизация экономики малых городов российского Севера на основе активизации межмуниципальных хозяйственных связей // Север и рынок: формирование экономического порядка. 2023. Т. 26. № 3. С. 150–164.

3. *Коломак Е. А., Шерубнёва А. И.* Оценка влияния агломерационных факторов на экономическую активность (микроэкономический анализ) // Экономика региона. 2023. Т. 19. № 3. С. 766–781.

4. *Кузнецова О. В.* Структура экономики российских регионов и уровень их социально-экономического развития // Научные труды: Институт народохозяйственного прогнозирования РАН. 2018. С. 473–493.

5. *Кузьминов И. Ф., Лобанова П. А.* Использование текст-майнинга в экономико-географическом отраслевом анализе целлюлозно-бумажной промышленности Европейской России // Региональные исследования. 2021. № 1. С. 18–33.

6. *Лачининский С. С.* Пространственная структура и особенности развития поселений Санкт-Петербургской агломерации // Балтийский регион. 2021. Т. 13. № 1. С. 48–69.

7. *Макушин М. А., Бобровский Р. О., Демидова К. В.* [и др.]. Социально-экономическое развитие территорий в зоне влияния БАМ: советские планы и российские реалии // Географический вестник = Geographical bulletin. 2023. Социально-экономическое развитие территорий в зоне влияния БАМ. № 2 (65). С. 12–25.

8. *Моисеева Е. Н., Скугаревский Д. А.* Рынок юридических услуг в России: что говорит статистика (Серия «Аналитические обзоры по проблемам правоприменения»). Рынок юридических услуг в России. ИПП ЕУ СПб, 2016.

9. *Никоноров С. М., Кривичев А. И., Максимов Ю. И.* Управление социально-экономической политикой в моногородах Республики Коми // Экономика устойчивого развития. 2021. № 4 (48). С. 123–129.

10. *Панкратов А. А.* Анализ современного состояния российской ИТ-отрасли: ключевые проблемы и тенденции // ИнтерКарто. ИнтерГИС. 2023. Т. 29. Анализ современного состояния российской ИТ-отрасли. № 1. С. 201–216.

11. *Петров Ю. В.* Пространственное сочетание сельской и городской местности на юге Тюменской области: проблемы, возможные решения // Географическая среда и живые системы. 2021. Пространственное сочетание сельской и городской местности на юге Тюменской области. № 3. С. 54–75.

12. *Ростислав К. В.* Влияет ли географическое сосредоточение на прибыльность российских предприятий? // Региональные исследования. 2021. № 1 (71). С. 4–17.

13. *Ростислав К. В.* Экономико-географическое положение как фактор различий в производительности между регионами России // Региональные исследования. 2020. № 3. С. 79–91.

14. *Саранча М. А.* Методика оценки уровня и масштабов развития малого предпринимательства в Приволжском федеральном округе (на примере деятельности

гостиниц и ресторанов) // Вестник ассоциации вузов туризма и сервиса. 2014. Т. 8. С. 27–32.

15. *Сафронов С. Г.* Трансформация третичной сферы экономики в регионах России в постсоветский период // Известия Российской академии наук. Серия географическая. 2021. Т. 85. № 4. С. 485–499.

16. *Фёдоров Г. М.* Экономика регионов России на Балтике: уровень и динамика развития, структура, внешнеторговые партнёрства // Балтийский регион. 2022. Т. 14. Экономика регионов России на Балтике. № 4. С. 20–38.

17. *Jozefowicz S., Stone M., Aravopoulou E.* Geospatial data in the UK // The Bottom Line. 2020. Vol. 33. No. 1. P. 27–41.

Сведения об авторе

Сёмин Павел Олегович — аспирант кафедры социально-экономической географии Пермского государственного национального исследовательского университета, г. Пермь, Россия; разработчик в ООО «АгроСофтвер», г. Москва, Россия.

E-mail: ntsp@ya.ru

ORCID: 0000-0002-4015-9206

About the author

Pavel Syomin, PhD Student, Department of Social and Economic Geography, Perm State University, Perm, Russia; Software Developer, AgroSoftware LLC, Moscow, Russia.

E-mail: ntsp@ya.ru

ORCID: 0000-0002-4015-9206

Поступила в редакцию 18.04.2024 г.
Поступила после доработки 30.05.2024 г.
Статья принята к публикации 03.06.2024 г.

Received 18.04.2024.
Received in revised form 30.05.2024.
Accepted 03.06.2024.