УДК 911.3, 519.688

П.О. Сёмин

КОНСОЛЬНОЕ ПРИЛОЖЕНИЕ ДЛЯ СОЗДАНИЯ ГЕОПРИВЯЗАННОГО НАБОРА ДАННЫХ О СУБЪЕКТАХ МАЛОГО И СРЕДНЕГО ПРЕДПРИНИМАТЕЛЬСТВА В РОССИИ НА БАЗЕ ОТКРЫТЫХ ДАННЫХ ФЕДЕРАЛЬНОЙ НАЛОГОВОЙ СЛУЖБЫ

Аннотация. Представлено консольное приложение на языке Python, которое создает полный геопривязанный набор данных о субъектах малого и среднего предпринимательства в России. Источник сведений – открытые данные Федеральной налоговой службы. Системные требования – 12 Гб оперативной памяти, 300 Гб свободного места на диске. Создаваемый набор данных содержит: наименование, вид, категорию, ИНН, ОГРН, адрес (населенный пункт, его координаты, ОКТМО), код ОКВЭД, доходы, расходы, среднесписочную численность работников за каждый год с 2016 по 2023 гг.

Ключевые слова: открытые данные, малое и среднее предпринимательство, ФНС России, набор данных, big data.

В современных экономико-географических исследованиях нередко возникает потребность в дезагрегированных данных. К ним, в частности, относятся сведения о конкретных организациях: наименование, местонахождение, прибыль, число работников, вид деятельности. Такая информация важна при построении математических моделей [6], для оценки уровня развития территорий [1], при анализе узких сфер деятельности, по которым нет отдельной статистики [4; 5], для исследований на локальном уровне (отдельные города, административные районы, муниципальные образования) [2; 3]. Для получения дезагрегированных данных исследователи обычно обращаются к системам бизнес-аналитики («Спарк», «Контур.Фокус»), онлайн-картам (коммерческим, таким как 2ГИС или Яндекс.Карты, или открытым, например, OpenStreetMap [3]. У такого подхода есть существенные недостатки: если данные покупаются, то надо платить, заключать отдельные договоры, соблюдать ограничения, а если данные собираются с общедоступных карт, то их обработка требует существенных усилий и не всегда законна.

В статье представлен инструмент, позволяющий самостоятельно создать набор геопривязанных данных о субъектах малого и среднего предпринимательства (МСП) в России. Получаемый набор содержит такую информацию, как наименование и местонахождение фирмы, ее регистрационные номера (ИНН, ОГРН), вид и категорию (физическое или юридическое лицо; микропредприятие, малое или среднее предприятие), код основного вида деятельности по ОКВЭД, доходы, расходы и среднесписочная численность работников за каждый год, начиная с 2019 г. Местонахождение дано с точностью до населенного пункта (городского или сельского), при этом указаны его географические координаты и код ОКТМО. Сведения доступны с середины

2016 г., что позволяет отслеживать временную динамику. Источник информации – открытые данные Федеральной налоговой службы (ФНС России). Использование открытых данных налоговой службы в экономико-географических исследованиях уже практиковалось рядом специалистов [1; 3; 6]. Обрабатывать открытые данные ФНС России сложно из-за большого объема и запутанного формата. Предложенный инструмент упрощает эту задачу.

Приложение использует три набора открытых данных ФНС России: один основной и два вспомогательных. Основной – «Реестр субъектов малого и среднего предпринимательства»[19], вспомогательные – «Сведения о доходах и расходах организаций» [20] и «Сведения о среднесписочной численности работников организаций»[21]. Каждый из них – это коллекция zip-архивов xml-файлов, где один архив содержит данные на определенную дату. Наборы обновляются с разной периодичностью, но не реже раза в год. Каждое обновление – это публикация текущей версии данных полностью, а не только изменений, поэтому сведения в архивах частично дублируются.

Для геопривязки приложение использует две справочные таблицы населенных пунктов России: набор «Населенные пункты России»[22] от проекта «Инфраструктура научно-исследовательских данных» и набор «Города России»[23], бесплатно опубликованный коммерческой компанией «DaData».

Приложение представляет собой консольную программу на языке программирования Python. Оно основано на нескольких популярных сторонних компонентах: Pandas (средство обработки табличных данных), lxml (библиотека для чтения xml-файлов), Apache Spark (инструмент для обработки больших объёмов данных), Typer (фреймворк для создания интерфейсов командной строки). Приложение работает на компьютере средней конфигурации: достаточно 12 Гб оперативной памяти и 300 Гб свободного места на жестком диске.

Алгоритм обработки данных состоит из пяти шагов: загрузка данных, извлечение данных из архивов (с опциональной фильтрацией по виду деятельности), агрегирование данных для удаления повторяющихся записей, геопривязка, создание панельного представления (по годам). Соответственно, приложение включает в себя пять ключевых компонентов: загрузчик (Downloader), распаковщик (Extractor), агрегатор (Aggregator), геопривязчик (Georeferencer), генератор панельного представления (Panelizer). Командный интерфейс, в свою очередь, содержит пять основных команд: download, extract, aggregate, georeference, panelize. У каждой команды есть опции, описание которых имеется в документации. Поскольку предполагается, что основной сценарий использования приложения – это создание набора данных «под ключ»,

19 Реестр субъектов малого и среднего предпринимательства. URL: https://www.nalog.gov.ru/opendata/7707329152-rsmp.

20Сведения о доходах и расходах организаций. URL: https://www.nalog.gov.ru/opendata/7707329152-revexp.

21 Сведения о среднесписочной численности работников организаций. URL: https://www.nalog.gov.ru/opendata/7707329152-sshr2019.

22Населенные пункты России. URL: https://data.rcsi.science/data-catalog/datasets/160.

23Города России. URL: https://github.com/hflabs/city.

то интерфейс включает комбинированную команду process, которая берёт исходные данные, по очереди запускает все этапы обработки и производит панельный набор данных. Приложение размещено в открытом репозитории по адресу: https://github.com/PavelSyomin/russian-smb-companies. Там же доступна документация с примерами использования.

Для проверки и демонстрации работы приложения был подготовлен набор данных о фирмах, ведущих деятельность в сфере сельского, лесного хозяйства, охоты, рыболовства и рыбоводства (группа «А» ОКВЭД). Объем исходных данных – чуть больше 200 Гб в сжатом виде и около 2,5 Тб в несжатом, размер итогового панельного представления – 630 Мб, т.е. в сотни раз меньше. Время обработки – около 15 ч. Набор данных содержит сведения о 417 467 фирмах. Картограмма (рис. 1) показывает все организации в сфере сельского хозяйства в 2021 г., сведения о которых имеются в наборе данных.

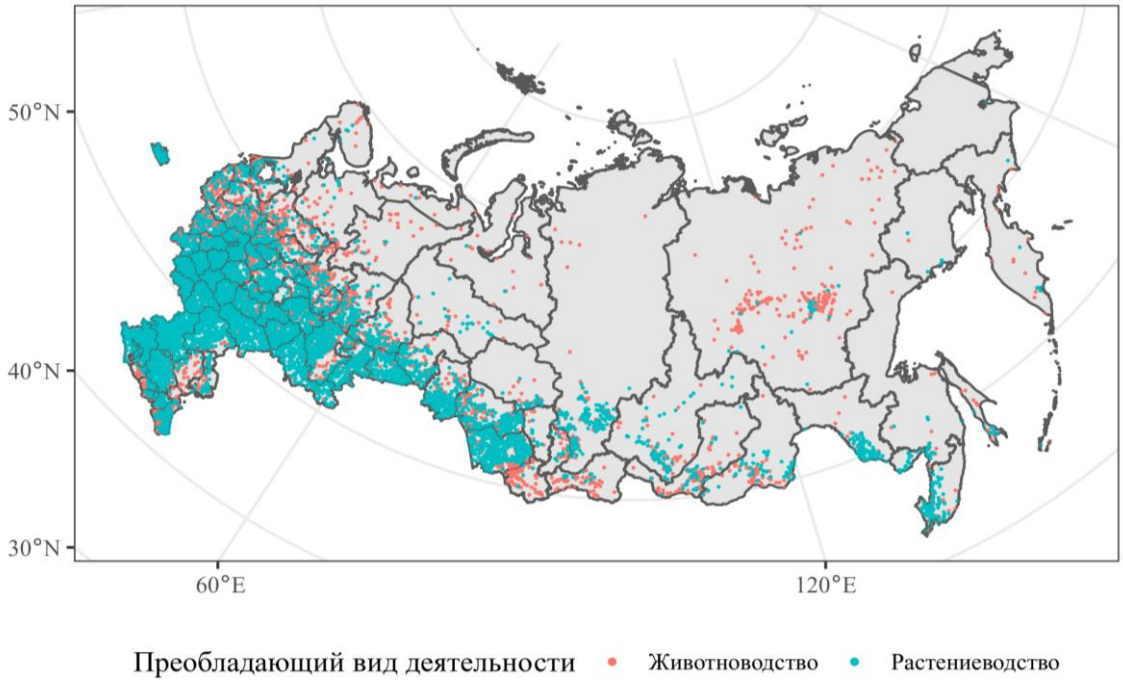

Рис. 1. Населенные пункты регистрации российских малых и средних организаций, ведущих деятельность в сфере сельского хозяйства (растениеводства и животноводства) в 2021 г.

Если в одном и том же населенном пункте находится несколько организаций разной отрасли, то цвет показывает, каких организаций в этом месте больше. В таблице 1 представлен пример одной записи из набора данных, который позволяет понять состав сведений.

Таблица 1

Пример одной записи из набора данных, созданного с помощью приложения

Признак	Значение
id	248357
tin	6658130962
reg_number	1026602328268
kind	1

Признак	Значение
category	1
first_name	–
last_name	–
patronymic	–
org_name	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ “КЛИФ - А”
org_short_name	ООО “КЛИФ - А”
activity_code_main	02.20
region	Свердловская область
area	–
settlement	Екатеринбург
settlement_type	Г
oktmo	65701000001
lat	56.8385216
lon	60.6054911
address_raw	СВЕРДЛОВСКАЯ / ОБЛАСТЬ / / / ЕКАТЕРИНБУРГ / ГОРОД / /
year	2021
confidence	1
revenue	758000
expenditure	2031000
employees_count	3

Таким образом, приложение помогает исследователю собрать и подготовить исходные данные для дальнейшего анализа. Оно может использоваться как частичная альтернатива коммерческим сервисам, предоставляющим дезагрегированные данные об организациях. Получаемый набор данных подходит для изучения малого и среднего предпринимательства как такового, а также анализа тех отраслей экономики и видов экономической деятельности, где малые и средние предприятия преобладают.

Библиографический список

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.

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

3. Лачининский С.С. Пространственная структура и особенности развития поселений Санкт-Петербургской агломерации // Балтийский регион. 2021. № 1 (13).

4. Моисеева Е.Н., Скугаревский Д.А. Рынок юридических услуг в России: что говорит статистика // Аналитические обзоры по проблемам правоприменения. 2016. № 1.

5. Панкратов А.А. Анализ современного состояния российской ИТ-отрасли: ключевые проблемы и тенденции // ИнтерКарто. ИнтерГИС. 2023. № 1 (29).

6. Ростислав К.В. Влияет ли географическое сосредоточение на прибыльность российских предприятий? // Региональные исследования. 2021. № 1 (71).

P.O. Syomin

A PYTHON CLI APPLICATION TO GENERATE A GEO-REFERENCED DATASET OF SMALL AND MEDIUM-SIZED BUSINESSES IN RUSSIA BASED ON FEDERAL TAX SERVICE OPEN DATA

Abstract. A command-line Python tool generates a complete geo-referenced dataset of small and medium-sized businesses in Russia. Open data of Federal Tax Service is used as a source. System requirements are 12 Gb of RAM and 300 Gb of free storage. The created dataset includes the following features for each company for each year from 2016 to 2023: name, kind, category, taxpayer number, registration number, address (settlement, geographic coordinates, municipal code), activity code, revenue, expenditure, count of employees.

Key words: open data, small and medium-sized businesses, FTS of Russia, dataset, big data.

Сведения об авторе

Сёмин Павел Олегович аспирант, Пермский государственный национальный исследовательский университет, бэкенд-разработчик, ООО «АгроСофтвер», ntsp@ya.ru.

УДК 911.3

Н.М. Скобеев

ТЕНДЕНЦИИ В ИЗМЕНЕНИИ ЗЕМЛЕПОЛЬЗОВАНИЯ И СПЕЦИФИКА ИХ УЧЕТА НА ПРИМЕРЕ ТУЛЬСКОЙ ОБЛАСТИ

Аннотация. В статье проведен анализ внутрирегиональной динамики землепользования в начале XXI века. Показано, в каких аспектах существующая система учета земель отображает лучше или хуже реальное состояние земельных ресурсов и их использования. Проведено сравнение и верификация данных из двух источников: сельскохозяйственной переписи Росстата и региональных докладов о состоянии и использовании земель Росреестра.

Ключевые слова: землепользование, сельскохозяйственная перепись, категории земель, угодья, сельскохозяйственные организации, фермерские хозяйства, личные подсобные хозяйства, поляризация, концентрация.

Различия в характере освоения, динамике сельскохозяйственного производства и землепользования внутри регионов зачастую не только не уступают межрегиональным, но порой и превосходят их [4]. Наиболее активно трансформационные процессы в землепользовании происходят в земледельчески освоенных регионах Центральной России в пределах южнотаежной лесной и лесостепной природных зон [10; 2; 5].