КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 6 С. 1713–1734
DOI: 10.20537/2076-7633-2023-15-6-1713-1734

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.2, 339.37

Прогнозирование розничной торговли на высокочастотных обезличенных данных

В. М. Тимирьянова[1,a], И. А. Лакман[1], М. М. Ларькин[2]

[1]Уфимский университет науки и технологий,
Россия, 450076, Республика Башкортостан, г. Уфа, ул. Заки Валиди, д. 32
[2]АО «Энергетические системы и коммуникации»,
Россия, 115114, г. Москва, Кожевнический проезд, д. 3

E-mail: [a] 79174073127@mail.ru

Получено 04.04.2023, после доработки — 06.07.2023.
Принято к публикации 18.08.2023.

Развитие технологий определяет появление данных с высокой детализацией во времени и пространстве, что расширяет возможности анализа, позволяя рассматривать потребительские решения и конкурентное поведение предприятий во всем их многообразии, с учетом контекста территории и особенностей временных периодов. Несмотря на перспективность таких исследований, в настоящее время в научной литературе они представлены ограниченно, что определяется их особенностями. С целью их раскрытия в статье обращается внимание на ключевые проблемы, возникающие при работе с обезличенными высокочастотными данными, аккумулируемыми фискальными операторами, и направления их решения, проводится спектр тестов, направленный на выявление возможности моделирования изменений потребления во времени и пространстве. Особенности нового вида данных рассмотрены на примере реальных обезличенных данных, полученных от оператора фискальных данных «Первый ОФД» (АО «Энергетические системы и коммуникации»). Показано, что одновременно со спектром свойственных высокочастотным данным проблем существуют недостатки, связанные с процессом формирования данных на стороне продавцов, требующие более широкого применения инструментов интеллектуального анализа данных. На рассматриваемых данных проведена серия статистических тестов, включая тест на наличие ложной регрессии, ненаблюдаемых эффектов в остатках модели, последовательной корреляции и кросс-секционной зависимости остатков панельной модели, авторегрессии первого порядка в случайных эффектах, сериальной корреляции на первых разностях панельных данных и др. Наличие пространственной автокорреляции данных тестировалось с помощью модифицированных тестов множителей Лагранжа. Проведенные тесты показали наличие последовательной корреляции и пространственной зависимости данных, обуславливающих целесообразность применения методов панельного и пространственного анализа применительно к высокочастотным данным, аккумулируемым фискальными операторами. Построенные модели позволили обосновать пространственную связь роста продаж и ее зависимость от дня недели. Ограничением для повышения предсказательной возможности построенных моделей и последующего их усложнения, за счет включения объясняющих факторов, стало отсутствие в открытом доступе статистики, сгруппированной в необходимой детализации во времени и пространстве, что определяет актуальность формирования баз высокочастотных географически структурированных данных.

Ключевые слова: фискальные данные, обезличенные высокочастотные данные, оператор фискальных данных, пространственная регрессия на панельных данных

COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 6 P. 1713–1734
DOI: 10.20537/2076-7633-2023-15-6-1713-1734

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.2, 339.37

Retail forecasting on high-frequency depersonalized data

V. M. Timiryanova[1,a], I. A. Lakman[1], M. M. Larkin[2]

[1]Ufa University of Science and Technology,
32 Zaki Validi st., Ufa, Republic of Bashkortostan, 450076, Russia
[2]JSC "Energy Systems and Communications",
3 Kozhevnichesky proezd, Moscow, 115114, Russia

E-mail: [a] 79174073127@mail.ru

Received 04.04.2023, after completion — 06.07.2023.
Accepted for publication 18.08.2023.

Technological development determines the emergence of highly detailed data in time and space, which expands the possibilities of analysis, allowing us to consider consumer decisions and the competitive behavior of enterprises in all their diversity, taking into account the context of the territory and the characteristics of time periods. Despite the promise of such studies, they are currently limited in the scientific literature. This is due to the range of problems, the solution of which is considered in this paper. The article draws attention to the complexity of the analysis of depersonalized high-frequency data and the possibility of modeling consumption changes in time and space based on them. The features of the new type of data are considered on the example of real depersonalized data received from the fiscal data operator "First OFD" (JSC "Energy Systems and Communications"). It is shown that along with the spectrum of problems inherent in high-frequency data, there are disadvantages associated with the process of generating data on the side of the sellers, which requires a wider use of data mining tools. A series of statistical tests were carried out on the data under consideration, including a Unit-Root Test, test for unobserved individual effects, test for serial correlation and for cross-sectional dependence in panels, etc. The presence of spatial autocorrelation of the data was tested using modified tests of Lagrange multipliers. The tests carried out showed the presence of a consistent correlation and spatial dependence of the data, which determine the expediency of applying the methods of panel and spatial analysis in relation to high-frequency data accumulated by fiscal operators. The constructed models made it possible to substantiate the spatial relationship of sales growth and its dependence on the day of the week. The limitation for increasing the predictive ability of the constructed models and their subsequent complication, due to the inclusion of explanatory factors, was the lack of open access statistics grouped in the required detail in time and space, which determines the relevance of the formation of high-frequency geographically structured data bases.

Keywords: cash registers data, retail scanner data, depersonalized high-frequency data, fiscal data operator, spatial regression on panel data

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 6, pp. 1713–1734 (Russian).

Введение

В 2003 году Федеральным законом № 54-ФЗ «О применении контрольно-кассовой техники при осуществлении наличных денежных расчетов и (или) расчетов с использованием электронных средств платежа» были обозначены правила, а в 2016 г. была сформирована инфраструктура и были выданы первые разрешения на сбор и обработку фискальных данных. В 2021 году абсолютно все работающие на территории России розничные предприятия, а также ИП без наемных работников перешли на кассовые аппараты, предусматривающие передачу данных операторам фискальных данных (далее ОФД). Фактически это ознаменовало возможность полного учета розничных продаж товаров населению, а следовательно, формирования базы данных, характеризующей потребление в России.

С научной точки зрения данное событие является неотъемлемой частью наблюдаемой революции данных в социально-экономических исследованиях и в области государственного управления. Оно способствует переходу на новый уровень эмпирических исследований, позволяющих оценивать существующие политики [Kolsrud, Landais, Spinnewijn, 2017], и получать исчерпывающие данные об обороте розничной торговли и о потребительских решениях во всем их многообразии.

Принимая во внимание ценность таких данных, многие страны запустили аналогичные программы сбора фискальных данных [Chacaltana, Leung, Lee, 2018; Casey, Castro, 2015; Андрианова, Рябинина, 2018]. Интерес к ним проявили и негосударственные структуры (например, Nielsen Retail Scanner, Sandbox UNECE). Однако, несмотря на множественность предположений о возможных направлениях использования фискальных данных во благо развития общества [Kokh, Kovaleva, Ivanova, 2021; Dubois, Griffith, O'Connell, 2022; Muth et al., 2020; Ткачёв и др., 2020; OECD, 2019; Жабин, Турков, Волков, 2017], на текущий момент в открытом доступе очень мало научных исследований, опирающихся на данные, собираемые таким способом. Как правило, собранные фискальные данные остаются в распоряжении государства и чаще всего используются для альтернативной оценки индекса цен [Leclair et al., 2019; Калинин, Волин, 2022] и наблюдения за предприятиями [Жабин, Турков, Волков, 2017]. В значительно большем количестве можно встретить исследования, выполненные на лонгитюдных данных о транзакциях по кредитным, дебетовым и текущим счетам людей, полученных из таких источников, как банки [Carvalho et al., 2021], приложения для ведения учета (например, FinTech [Baker et al., 2020], Check.me [Gelman et al., 2014]), в том числе объединяемые в базы данных (например, Fable Data [Gathergood et al., 2021]). Эти исследования позволили расширить представления о потребительском поведении, балансах домохозяйств и их реакциях на колебания доходов или изменения государственной политики и значительно усилили понимание неоднородности в доходах и предпочтениях людей, что стало особенно актуально для принятия быстрых решений в условиях введения ограничительных мер и структурной перестройки экономик, вызванной коронавирусной инфекцией [Baker, Kueng, 2021]. Ключевой особенностью таких данных является огромный размер выборки наблюдений, что позволяет достичь достоверности результатов моделирования и увидеть все разнообразие персональных решений индивидов, в том числе ненаблюдаемых при анализе агрегированных данных.

Актуальность исследования микроданных, собираемых посредством фискальных аппаратов, в России достаточно высока, однако в научной литературе очень редко встречаются как результаты моделирования, так и описательная информация о данном виде данных, проблемах, возникающих при их анализе, и методах решения этих проблем. Фискальные данные, как первичная информация о продажах товаров на розничном рынке, структурирована во времени и пространстве. Но те редкие исследования фискальных данных, которые встречаются в научной литературе, либо оперируют агрегированными дневными данными, либо не учитывают

пространственные эффекты, концентрируя внимание на эффектах по периодам [Guha, Ng, 2019; Lovics et al., 2019; Cotti et al., 2020; Waldenström, Angelov, 2021; Gathergood et al., 2021]. Цель проведенного исследования состояла в выявлении ключевых проблем прогнозирования розничной торговли на высокочастотных обезличенных данных, включая оценку возможности выделения эффектов во времени и пространстве. Исследование проводилось на обезличенных данных, полученных от оператора фискальных данных «Первый ОФД» (АО «Энергетические системы и коммуникации»).

Проблемы высокочастотных данных и их предобработка

Первичная информация, собираемая автоматизированным способом, не лишена недостатков, а «цифровая природа больших данных определяет существенные особенности связанных с ними ошибок наблюдения» [Оксенойт, 2018]. Применительно к данным, получаемым посредством фискальных аппаратов, проблема в том, что «миллиарды чеков, в каждом из которых написаны те или иные наименования товаров без какого-либо стандарта, с ошибками и сокращениями, — крайне сложный материал для анализа», требующий активного внедрения технологий машинного обучения и других современных подходов [CNEWS, 2019]. Учитывая это, при их анализе нельзя исключать случайный некорректный ввод данных, механические ошибки и умышленное искажение данных со стороны продавца.

Многообразие товаров, в совокупности с определенной свободой в порядке ввода данных, определяют еще одну проблему фискальных данных, а именно различное отражение товаров и их единиц измерения (литры, килограммы, штуки и т. п.) [Muth et al., 2020]. Одинаковые по типу товары должны быть приведены к единой размерной сетке, что в реальности осложняется не только различным вариантом, но и местом отражения единицы измерения товара (как правило, внутри текста столбца наименования), что требует дополнительной разработки алгоритмов, в том числе предусматривающих глубокий анализ текстов. Отдельно выделяется проблема измерения продуктов, реализуемых наборами, которые рекомендуется исключать из анализа [Muth et al., 2020].

Исследования показывают, что данные часто содержат выбросы, которые рекомендуется отбрасывать [Aladangady et al., 2019; Lovics et al., 2019; Deshaies-Moreault, Harper, Yung, 2018]. Например, К. Дешайез-Моролт и его соавторы исключили из анализа продукты, объем продаж которых за месяц составил менее $ 10 [Deshaies-Moreault, Harper, Yung, 2018]. В практическом руководстве по обработке данных фискальных аппаратов, разработанном Евростатом, предложена система фильтров, среди которых пороги цен, предусматривающие исключение продаж с использованием очень больших скидок или купонов на бесплатное получение товара, и алгоритмы выделения товаров с низкими объемами продаж [Eurostat, 2017]. Проблема нулевых продаж по купонам и акциям отмечается также в работе М. К. Мус и соавторов [Muth et al., 2020]. На стороне покупателей возникает другая проблема, связанная с появлением в исследуемых данных необычных наборов при экстремально высоких расходах по картам, не соответствующим розничному потреблению. Такие чеки рассматриваются как выбросы и исключаются из анализа [Clark et al., 2021]. Например, Х. Чен и соавторы исключили 1-й и 99-й перцентили выборки [Chen, Qian, Wen, 2020].

Для розничной торговли характерны циклические колебания продаж (месяц, день недели, час) [Timiryanova et al., 2023]. При панельном анализе в рамках года на изменение оценок по периодам могут оказывать влияние экстремальные продажи в праздники, не привязанные к конкретным дням недели. Учитывая это, в работе В. М. Карвало и его соавторов предобработка данных предполагает корректировку данных в праздничные дни (7–8 января) [Carvalho et al., 2021]. Отмечая те же проблемы, другие ученые сглаживают объемы транзакций во времени [Gathergood

et al., 2021]. В свою очередь, неоднородность данных [Buda et al., 2022], их левосторонняя асимметрия, вызванная преобладанием в чеках товаров по более низкой цене, определяет необходимость их логарифмирования [Bounie, Camara, Galbraith, 2020; Carvalho, Peralta, Pereira dos Santosa, 2020].

В исследованиях, в которых проводится анализ первичных данных о продажах товаров, обращается внимание не только на сами чеки/транзакции, но и на выборку розничных предприятий. Например, в исследовании А. Аладангаду и его соавторов отбрасываются данные предприятий, у которых более 40 % объема транзакций сконцентрировано в один день в месяце или отмечается менее 4 дней с движениями по счету, а также высокое процентное изменение продаж и т. д. [Aladangady et al., 2019]. Проблема изменения выборки в результате закрытия/открытия предприятий решается либо путем отбрасывания данных, по которым нет сведений за весь анализируемый период [Andersen et al., 2020; Cotti et al., 2020], либо путем формирования скользящей выборки продавцов [Aladangady et al., 2019]. В каждой географической зоне может наблюдаться разная доля предприятий, деятельность которых учтена в данных. Соответственно, нельзя сопоставить абсолютные значения товарооборота разных территорий, так как они отражают значения только доли продавцов, а эта доля не одинакова для каждой рассматриваемой территории. Решением этой проблемы является расчет относительных показателей [Lee, Lee, 2022; Powell, Leider, Léger, 2020; Alfaro, Park, 2020]. Дополнительным способом приведения данных к сопоставимому набору являются их стандартизация и нормализация [Carvalho et al., 2021; Gathergood et al., 2021]. Учитывая отсутствие данных об изменении цен в требуемой детализации по периодам и видам товаров, как правило, высокочастотные данные рассматриваются в номинальном выражении, без поправок на уровень инфляции [Carvalho et al., 2021].

Отмеченные проблемы высокочастотных данных в полной мере характерны и для данных, собираемых фискальными операторами в России. Высокий интерес к ним определяет развитие инструментов их обработки, позволяющее успешно использовать их для последующего моделирования.

Моделирование изменения продаж во времени и пространстве на высокочастотных данных

Высокочастотные данные в розничной торговле, представляющие собой данные о реализации товаров и услуг, имеющие высокую детализацию по времени (час, минута, секунда) и объектам наблюдения (транзакция, товар или услуга), могут быть получены из различных источников.

Опубликованных исследований, основанных на анализе фискальных данных, аккумулируемых государственными организациями, не много, и они часто опираются на агрегированные данные. Например, на фискальных данных французских розничных предприятий, агрегированных по дням недели, исследовалась перспектива построения индекса цен по 17 группам товаров [Leclair et al., 2019]. В России агрегированные по дням недели данные об объеме продаж и ценах в разрезе социально значимых товаров, опубликованные Федеральной налоговой службой, использовались в анализе пространственной динамики цен [Timiryanova et al., 2023]. Агрегированные на уровне месяца данные фискальных аппаратов (online cash registers, OCR), в целях оценки влияния введения обязательных онлайн-касс на деятельность предприятий Венгрии, сопоставлялись со сведениями о налогах и декларируемых ими оборотах [Lovics et al., 2019]. Данные регистраторов (tax registers), полученные из Налогового агентства Швеции, использовались для анализа изменения экономической активности в условиях пандемии [Waldenström, Angelov, 2021]. В исследовании также не рассматривались высокочастотные данные, так как в Швеции данные фискальных аппаратов передаются ежемесячно или ежеквартально.

Высокочастотные данные аккумулируются не только у государства. Значительное количество исследований выполняется на базе Nielsen (Nielsen retail scanner dataset, NRSD), содержащей сгруппированные по универсальным продуктовым кодам данные, генерируемые торговыми терминалами (universal product code (UPC)-level store scanner data). Так, в разрезе 10 товарных категорий на данных 55 метрополий США были изучены цикличные колебания продаж товаров и проанализирована их региональная вариация [Guha, Ng, 2019]. Сравнение продаж в отдельных районах и городах США проводилось отдельно для напитков [Powell, Leider, Léger, 2020; Rojas, Wang, 2021] и сигарет [Cotti et al., 2020]. Данные базы Fable Data Limited были использованы для анализа совокупного потребления и оценки региональной неравномерности восстановления объемов расходов населения в Великобритании [Gathergood et al., 2021].

Проведенный обзор показывает, что на текущий момент опубликовано сравнительно небольшое количество исследований, опирающихся на высокочастотные фискальные данные. Фискальные данные намного реже используются в исследованиях по сравнению с данными транзакций, получаемых от банков и платежных систем. И в случае использования фискальных данных, и в случае использования данных транзакций, при наличии высокочастотных данных, в ряде случаев характеризующих продажи с точностью до минуты, в модели, как правило, включаются агрегированные дневные данные [Bounie, Camara, Galbraith, 2020; Gathergood et al., 2021]. Акцент в этих исследованиях концентрируется на изменениях продаж во времени [Guha, Ng, 2019; Cotti et al., 2020; Chen, Qian, Wen, 2020; Bounie, Camara, Galbraith, 2020] и влиянии событий (получение заработной платы, социальных трансферов, изоляции, природных катаклизмов и т. д.), а также различных факторов на рассматриваемую динамику [Alfaro, Park, 2020; Guha, Ng, 2019; Lovics et al., 2019]. Это определяет выбор моделей и методов анализа. Как правило, применяются панельный анализ [Lovics et al., 2019; Guha, Ng, 2019; Carvalho, Peralta, Pereira dos Santosa, 2020; Cotti et al., 2020; Carvalho et al., 2021; Waldenström, Angelov, 2021] и метод разниц в разнице (difference-in-difference, DiD) [Chen, Qian, Wen, 2020; Rojas, Wang, 2021]. При этом можно отметить активное включение в модели фиктивных переменных [Lovics et al., 2019; Bounie, Camara, Galbraith, 2020; Chen, Qian, Wen, 2020]. Это связано с отсутствием объясняющих переменных в той же детализации. Сложности возникают с автоматической классификацией и последующей группировкой данных в случае их получения из разных источников [Baker, Kueng, 2021]. Как результат, чем выше детализация данных во времени и пространстве, тем меньше факторов включается в модели.

Методы пространственного анализа практически не применяются в исследованиях, проведенных на фискальных данных. Вопросы пространства поднимаются преимущественно с целью привязки данных, полученных из разных источников, к конкретным географическим единицам [Cotti et al., 2020] или для исследования региональной вариации [Guha, Ng, 2019; Gathergood et al., 2021]. В то же время пространственный аспект активно изучается на высокочастотных данных о транзакциях, аккумулируемых платежными системами. Интерес к данному направлению исследования возрос в период пандемии коронавируса. Например, анализ потребительского поведения осуществлялся на данных об объеме транзакций во Франции, предоставленных национальной системой расчетов Groupement des Cartes Bancaires CB. Данные включали дату и время покупки с точностью до минуты, тип и канал покупки, а также данные о местоположении предприятий на уровне пятизначного номера почтового индекса, которые использовались для анализа мобильности населения путем оценки «расстояния пройденного картой» [Bounie, Camara, Galbraith, 2020]. Модель с фиксированными эффектами, построенная на ежедневных данных о транзакциях в Сеуле, аккумулируемых платежной системой Hyundai Card, позволила обосновать наличие пространственной неоднородности в реакциях населения и показала отсутствие доказательств пространственного замещения локализованного спроса в сторону районов с более низкими рисками COVID-19 [Lee, Lee, 2022]. Исследования, проводимые на данных транзак-

ций, позволили обосновать, что расходы по картам фиксируют закономерности и во времени, и в пространстве, при этом показали, что в последних меньше шума, чем в первых [Carvalho et al., 2021], что определяет целесообразность более активного исследования пространственных зависимостей.

Значимость пространственных зависимостей неоднократно подтверждалась и в анализе данных отдельных торговых сетей [Trivedi, 2011; Verhelst, Van den Poel, 2013]. В этих исследованиях фактически проводилось пространственное моделирование на фискальных данных, но применительно к относительно небольшой группе торговых точек. Таким образом, фискальные данные могут применяться в анализе пространственной зависимости торговых процессов, однако в настоящее время сложно встретить такое исследование, построенное на высокочастотных данных, аккумулируемых фискальными операторами.

Высокочастотные обезличенные данные и их предобработка

Ключевые проблемы прогнозирования розничной торговли на высокочастотных обезличенных данных и возможности выделения эффектов во времени и пространстве были рассмотрены на примере высокочастотных данных, предоставленных АО «Энергетические системы и коммуникации» (далее — Первый ОФД) по соглашению о научно-техническом сотрудничестве в обезличенной форме в соответствии со статьей 4.1 Федерального закона № 54-ФЗ. Рассматривались данные о продаже хлеба и хлебобулочных изделий в г. Уфе в период с 25 июля по 8 августа 2019 г. Выбор группы товаров определялся их социальной значимостью и массовостью потребления. Данные были получены в необработанном виде с целью выявления особенностей исходной первичной информации. Процесс предобработки проходил в несколько этапов.

Первичное выделение данных осуществлялось на основе ключевых слов, характеризующих хлебобулочные изделия в общей массе товаров («хлеб», «булка», «батон» и т. п.). Оно показало, что отбор исключительно на основе ключевых слов недостаточен, так как в выборку попадают товары, фактически не являющиеся хлебом. Например, водка, шоколадные изделия, мясная продукция (табл. 1, строки 2–5). Соответственно, необходимы более жесткие алгоритмы для выделения групп товаров. Для этих целей Первый ОФД разработал свой собственный алгоритм, позволяющий ему выделять различные группы товаров для раздельного представления тенденций изменения их продаж.

Таблица 1. Пример обрабатываемых данных

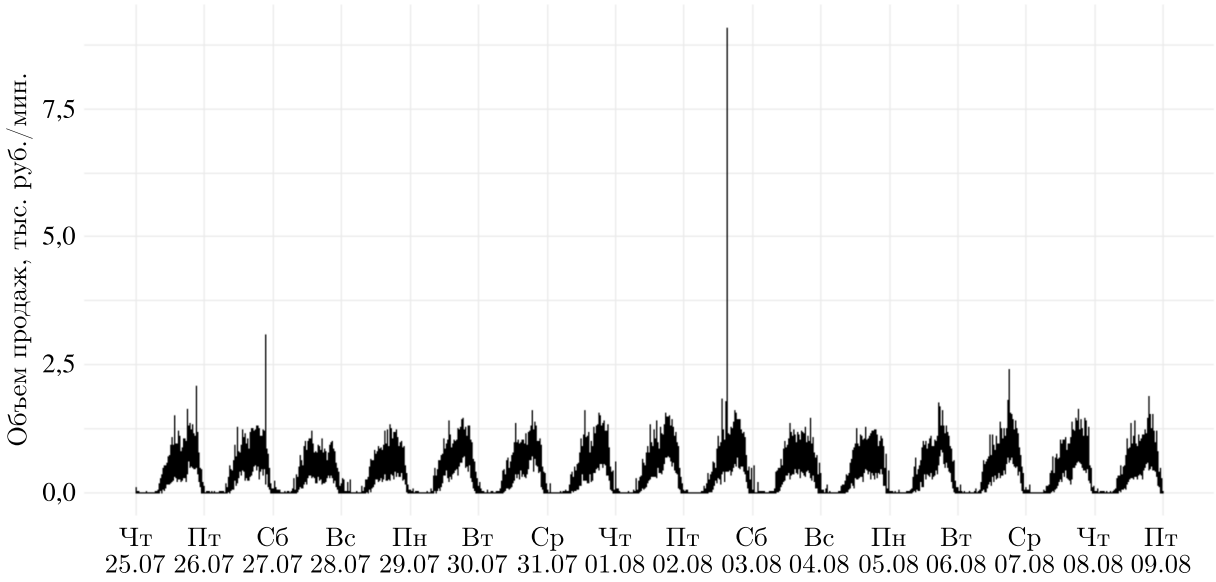
№	Дата и время	Наименование*	Сумма	Количество	Широта	Долгота
1	2019-07-27T12:59	хлеб 40 г	50,00	5,0	54,77	56,02
2	2019-07-29T21:12	вод. хлеб.пол.пш.мяг.40 %	295,99	1,0	54,81	56,06
3	2019-08-01T09:43	бабаевский батон 50гр	31,00	1,0	54,72	55,94
4	2019-08-03T16:54	fitofruit батон прот	21,00	1,0	54,67	55,92
5	2019-08-07T18:51	хлеб мясной празд кг	177,97	0,619	54,72	56,00
6	2019-07-28T13:37	хлеб цельнозерн вес	41,76	0,144	54,67	55,92
7	2019-07-30T20:04	хлеб кукурузный 250гр маг	158,00	7,9	54,76	56,05
8	2019-07-25T15:49	хлеб на хмелю	58	2	54,78	56,03
9	2019-07-25T20:11	хлеб энергия ржи рж-пш.нар.300г	0,00	1,0	54,71	55,99
10	2019-08-01T19:14	багет классический 1с, 100г	0,01	1,0	54,74	56,02
11	2019-08-02T15:04	булка хот-дог 250г	8427	75	54,67	55,92

* Текст представлен без редакторской правки.
Источник: обезличенные данные, предоставленные Первым ОФД.

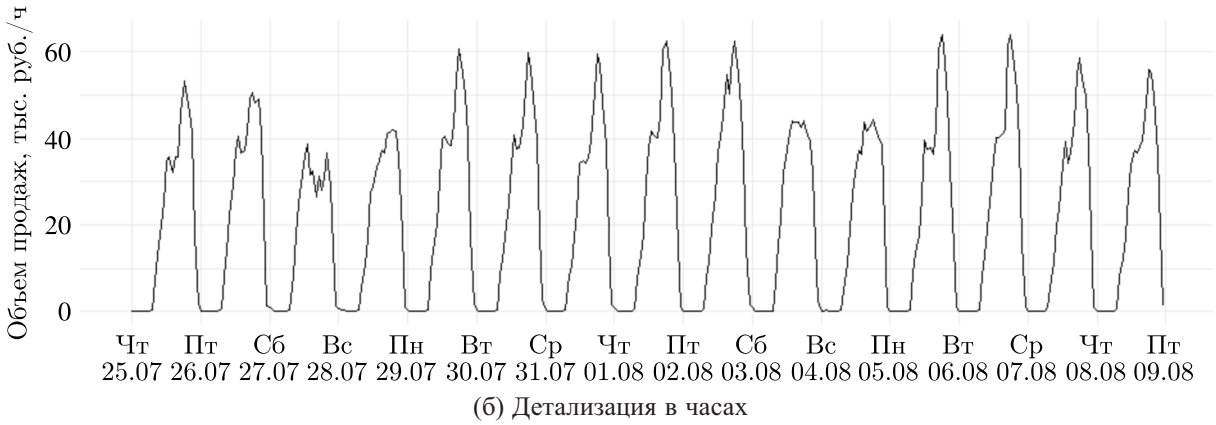
На следующем этапе анализировались сведения о количестве реализованных товаров. Результаты данного этапа согласуются с аналогичными исследованиями [Muth et al., 2020] и указы-

вают на наличие в рассматриваемом наборе данных с различным отражением единиц измерения товаров, присутствие неадекватных данных, нулевых продаж (табл. 1, строки 7, 9, 10). Например, выделяется одновременное отражение веса хлеба в наименовании и дробного значения в столбце «количество», являющееся следствием некорректного ввода данных на стороне продавца (табл. 1, строка 7). Неадекватные данные и нулевые продажи были исключены из анализа.

Далее данные анализировались на наличие выбросов. Их можно увидеть на минутном графике (рис. 1, *а*). На этом этапе данные очищались, опираясь на критерий 3σ, согласно которому число, отклоняющееся от среднего больше чем на три стандартных отклонения, рассматривалось как выброс.



(а) Детализация в минутах



(б) Детализация в часах

Рис. 1. Динамика объемов продаж хлеба, регистрируемая Первым ОФД в г. Уфе в период с 25 июля по 8 августа 2019 г. с лагом в минуту (а) и в час (б). Источник: составлено авторами по обезличенным данным, предоставленным Первым ОФД

В итоге с учетом всех отмеченных выше недостатков были исключены для последующего анализа 5,4 % полученных данных, что в целом соотносится с аналогичными исследованиями высокочастотных данных [Deshaies-Moreault, Harper, Yung, 2018; Lovics et al., 2019; Clark et al., 2021].

Исходные данные были представлены в максимально возможной детализации по времени (табл. 1), но в связи с тем, что объемы рынка хлеба Уфы не настолько большие, чтобы обеспечить ежеминутные продажи хлеба по всей территории города, они были агрегированы с шагом в 1 час (рис. 1). На графиках отчетливо видно внутридневное цикличное изменение с пиками продаж в вечернее время и их падением в ночное время. Также визуализируется снижение продаж в выходные дни.

Данные были получены с максимально возможной детализацией с точки зрения соблюдения требования сохранения конфиденциальности данных по признаку пространства. Как видно на рис. 2, полученные данные не покрывают всю территорию Уфы.

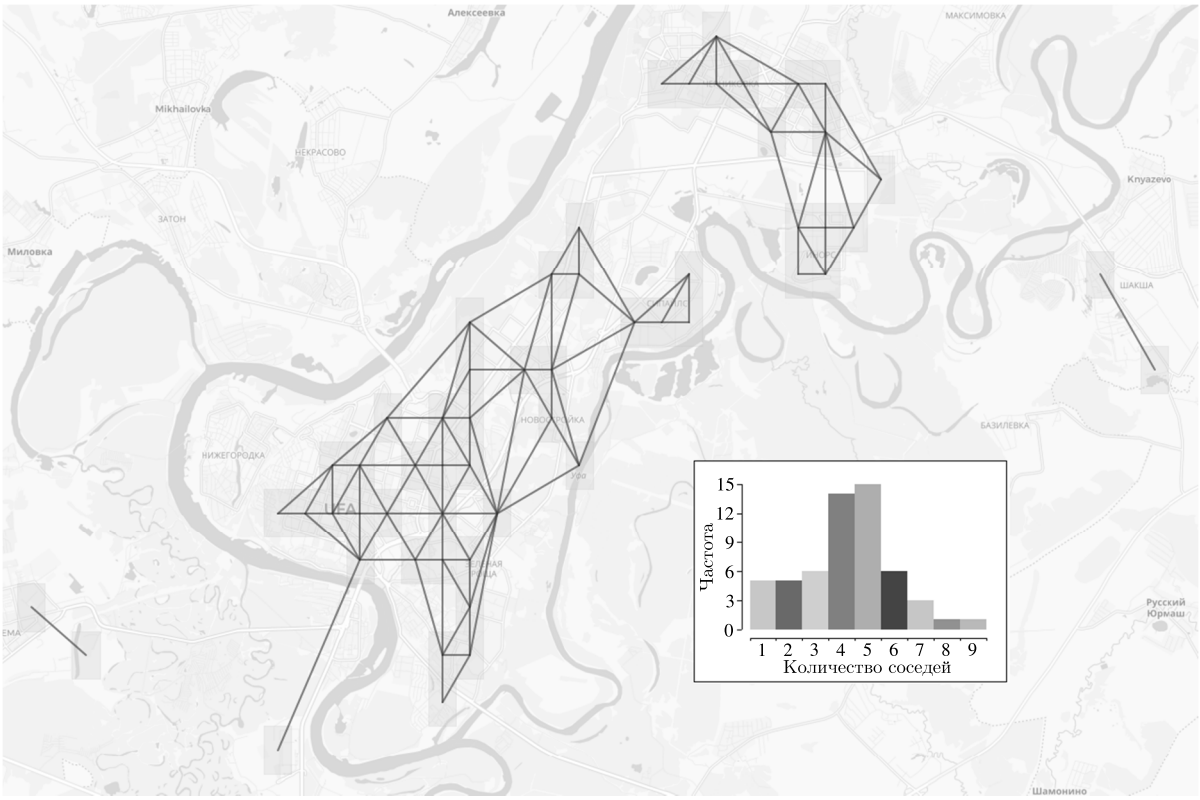


Рис. 2. Карта связанности, отражающая соседство объектов наблюдения, учитываемое весовой матрицей (56 географических зон г. Уфы). Источник: составлено авторами

Отсутствие данных по ряду зон связано с двумя причинами. Во-первых, не во всех зонах г. Уфы в представленной детализации (до 2 знаков после запятой для географических координат) у оператора фискальных данных (далее — ОФД) в анализируемом периоде были зарегистрированные пользователи. Во-вторых, не были получены данные по зонам, где отсутствовала возможность обеспечения конфиденциальности данных. В итоге были выделены 56 географических зон, в которых максимально исключена возможность выделения конкретных точек продаж. Следует также учитывать, что были неизвестны ни количество, ни доля предприятий, выбравших в качестве оператора фискальных данных Первый ОФД в рассматриваемых географических зонах, что не позволяло определять степень охвата рынка. С целью решения проблемы сопоставимости абсолютных значений объемов реализации в разрезе географических зон, они были стандартизированы:

$$y_{it}^{s} = \frac{y_{it} - \mu_i}{\sigma_i}, \tag{1}$$

где y_{it}^{s} — стандартизованный элемент признака, y_{it} — значение показателя в географической зоне i в период t, μ_i — среднее арифметическое для географической зоны i, σ_i — стандартное отклонение для географической зоны i.

Таким образом, без учета объемов продаж в зонах далее фактически рассматривалась наблюдаемая в них динамика продаж.

Методика оценки эффектов пространства и времени

В самом обобщенном виде пространственная модель, построенная на панельных данных, принимает вид

$$Y_{it} = \alpha + \mu_i + \gamma_t + \beta_1 T_{it} + \rho W Y_{it} + \lambda W u_{it} + \varepsilon_{it}, \tag{2}$$

где Y_{it} — стандартизированное в рамках i-й географической зоны значение объема продаж в t-й период времени (результирующая, зависимая переменная); T_{it} — фиктивная панельная переменная, учитывающая день недели (0 — выходной день, 1 — будний день) в t-й период в i-й географической зоне; WY_{it} — пространственно взвешенные значения зависимой переменной (зависимая панельная переменная Y_{it}, умноженная на весовую пространственную матрицу); W_{ij} — весовая матрица, отражающая пространственную связь объектов наблюдения между i-й и j-й географическими зонами (при ее создании учитывалось два критерия отсечения: во-первых, в качестве соседей рассматривались только соседи первого порядка (непосредственное соседство); во-вторых, учитывались только те наблюдения, расстояние до которых было менее $d = 0{,}039°$):

$$W_{ij} = \begin{cases} 0, & \text{если } i = j, \\ \dfrac{1}{d}, & \text{если } d_{ij} \leqslant d, \\ 0, & \text{если } d_{ij} > d, \end{cases} \tag{3}$$

где d_{ij} — расстояние между центрами i-го и j-го объектов наблюдения, Wu_t — матрица пространственной автокорреляции ошибки; ρ — авторегрессионный пространственный коэффициент[1]; λ — авторегрессионный пространственный коэффициент ошибок; μ_i — панельный индивидуальный эффект i-й географической зоны, не зависящий от времени t; γ_t — панельные эффекты по периодам t (панельные эффекты μ_i и γ_t могут быть как фиксированными, так и случайными [Croissant, Millo, 2019]); β_1 — коэффициент при регрессоре, не зависящий ни от периода времени t, ни от географической зоны i, подлежащий оцениванию; α — обобщенная константа, подлежащая оценке; ε_{it} — остатки (ошибки), нормально распределенные случайные величины с нулевым математическим ожиданием и постоянной дисперсией; t — индекс по периодам наблюдения ($t = 1, \ldots, K$), i — индекс по территориям ($i = 1, \ldots, 62$).

Для получения эффективных и несмещенных оценок коэффициентов при регрессорах требуется определиться с корректной спецификацией модели. Для этого необходимо проводить серию статистических тестов, в том числе для оценки возможности выделения эффектов пространства и времени на обезличенных высокочастотных данных. Соответственно, логика подбора наилучшей спецификации модели (2) укладывалась в следующую схему.

На первом этапе тестировали панельную переменную Y_{it} на наличие единичных корней. При этом для проверки гипотезы о наличии возможных обобщенных единичных корней использовали тест Левина – Лина – Чу, а для определения индивидуальных единичных корней — тест Има – Песарана – Шина. При этом нулевой гипотезой в обоих критериях служило предположение о том, что все временные ряды имеют единичный корень (являются независимыми процессами случайного блуждания), но в тесте Левина – Лина – Чу данные рассматриваются как

[1] Здесь под авторегрессией понимается авторегрессия не по временному, а по пространственному лагу.

единый временной ряд без учета его разделения на кросс-секции (в рассматриваемом случае это территории), а в тесте Има – Песарана – Шина рассматривается множество временных рядов, для каждого кросс-секционного наблюдения отдельно. В обоих критериях тестировалось авторегрессионное уравнение, рассматриваемое в трех модификациях: с включением и без включения индивидуальных констант μ_i и с включением в качестве регрессора времени t (по сути, детерминированного линейного тренда). Кроме того, во всех критериях использовались их расширенные варианты, учитывающие авторегрессию тестируемой переменной, требующие определения максимального лага запаздывания. Выбор длины лага в соответствующих тестах осуществлялся автоматически, исходя из минимумов значений модифицированных информационных критериев Акайка, Шварца, Холла [Hall, 1994]. Проведение такого исследования оправдано, так как позволяет ответить на вопрос о том, следует ли брать в уравнении (2) на исходные уровни переменной Y_{it} ее первые или вторые разницы или отклонение от детерминированного тренда.

На втором этапе исследования были проведены тесты на наличие ненаблюдаемых эффектов в остатках модели (тест Вулдриджа), последовательной корреляции (тест Вулдриджа на последовательную корреляцию) и кросс-секционной зависимости остатков панельной модели с фиксированными эффектами, в случае если выбиралась панельная модель с фиксированными эффектами (тест Песарана), авторегрессии первого порядка в случайных эффектах, в случае если выбиралась панельная модель со случайными эффектами (тест Бера – Соса – Эскудеро – Юна), сериальной корреляции (тест Балгати – Ли). Данные тесты проводились для определения спецификации модели и необходимости введения авторегрессионной составляющей в модель (2) для получения робастных оценок коэффициента при рассматриваемом регрессоре.

Третий этап подбора спецификации модели подразумевал определение наличия панельных эффектов в модели (2): при $\gamma_t = 0$, $\alpha = 0$ — модель панельных данных с фиксированным эффектами по объектам исследования; при $\mu_i = 0$, $\alpha = 0$ — модель с фиксированными эффектами по периодам. При $\alpha = 0$ рассматривается двунаправленная модель панельных данных, которая содержит одновременно как индивидуальные по объектам, так и временные эффекты. При этом разнообразие моделей определяется существующими подходами к выделению эффектов времени и пространства. Выделение эффекта времени возможно в рамках обобщенной модели панельных данных, модели с фиксированными эффектами, модели со случайными эффектами. В первом случае предписывается одинаковое изменение всех i объектов выборки во все периоды времени t и не учитываются индивидуальные различия. Во втором и третьем случаях отличие заключается в предположении о случайности и неслучайности индивидуальных отличий объектов наблюдения. В случае если ни одна из перечисленных выше спецификаций панельных моделей согласно проводимым тестам построена быть не может, имеет место семейство несвязных регрессионных уравнений. С целью выявления случайных и индивидуальных эффектов проводился тест Балтаги – Сонга – Коха и Броша – Погана [Croissant, Millo, 2019].

В свою очередь, учесть пространственную зависимость данных можно путем оценки пространственного лага и/или пространственной ошибки. Если в уравнении (2) $\lambda = 0$ — то модель пространственного лага, а если $\rho = 0$ — то модель пространственной ошибки. В первом случае учитывается связь зависимой переменной с ее значениями в соседних территориях, во втором случае определяется пространственная зависимость ошибок. Третий вариант спецификации модели объединяет оба подхода. Спецификация пространственных панельных моделей, помимо описанных выше вариантов, дополнительно расширяется за счет различных методов оценки пространственной автокорреляции ошибок (по методу Б. Балтаги [Baltagi, Song, Koh, 2003] или М. Капура [Kapoor, Kelejian, Prucha, 2007]).

Четвертый этап исследования был посвящен определению наилучшей с точки зрения получения надежных оценок спецификации, учитывающей пространственную связность территорий. Граф связанности, показывающий учитываемое при построении моделей соседство территорий,

представлен на рис. 2. Среди наблюдений пять зон имеют только одного соседа, что связано с объективной причиной удаленности отдельных районов г. Уфы. Следует также отметить, что на севере г. Уфы существует часть города, отделенная промышленной и парковой зоной, наблюдения в которой не связаны с остальными наблюдениями. Полученная весовая матрица, на наш взгляд, адекватно описывает существующие связи, так как исследуется динамика продаж товара повседневного спроса, для приобретения которого население вряд ли решится на преодоление больших расстояний. Следовательно, учитывать соседство более высокого порядка и/или более высокого радиуса охвата не целесообразно. В свою очередь, сокращение радиуса охвата привело бы к появлению изолянтов. Для оценки целесообразности построения пространственных панельных моделей были проведены модифицированные тесты множителей Лагранжа, проверяющие несколько гипотез, разделяя случайные пространственные эффекты, пространственно автокоррелированные остатки и авторегрессионные остатки первого порядка. Наряду с тестами Лагранжа были проведены тесты Балтаги – Сонга – Коха, в основе которых рассматриваются вариации сложных нулевых гипотез о равенстве нулю коэффициента пространственной автокорреляции и об отсутствии при этом случайных эффектов. Предварительное тестирование не только направлено на обоснование возможности построения пространственно-панельных моделей, но и необходимо для определения наиболее подходящей спецификации модели.

Критериями отбора подходящей спецификации модели являются минимальные значения информационных критериев Акайке и Шварца и максимальные значения коэффициента детерминации R^2. Для того чтобы удостовериться в эффективности и несмещенности получаемых оценок, по подобранной спецификации модели (2) исследовали ее остатки. Во-первых, тестировали с использованием теста Вальда равенство нулю среднего значения остатков; во-вторых, анализировали постоянство дисперсии остатков с помощью критерия Уайта и анализировали отсутствие автокорреляции с лагом запаздывания по времени на основе теста Дарбина – Уотсона. Для анализа отсутствия в остатках пространственной связности использовали тест Морана. В настоящем исследовании не проводилось исследование асимптотической состоятельности получаемых оценок по двум причинам. Во-первых, целью построения модели (2) является не перспективное, а каузальное прогнозирование, и для этих целей необходимо, чтобы оценка при исследуемом факторе влияния (день недели) была бы эффективной и несмещенной. Во-вторых, остатки модели, построенной на высокочастотных временных рядах, не имеют нормального распределения за счет большого эксцесса (> 3), но при этом коэффициент асимметрии распределения всегда ноль [Buda et al., 2022].

Как отмечалось ранее, в обзоре существующих подходов к моделированию объема продаж в розничной торговле, для моделей, выстраиваемых на высокочастотных данных, характерно небольшое количество включаемых факторов, что во многом связано с отсутствием других данных в той же детализации. В рамках данного исследования был протестирован только один фактор влияния — день недели.

Анализ проводился в среде R, с применением библиотек sf, spdep, lmtest, plm, splm и др.

Результаты

С целью определения возможности применения инструментов панельного анализа к исследованию продаж были проведены тесты на стационарность панельных данных. Все проведенные тесты показали, что нулевая гипотеза об отсутствии стационарности отвергается в пользу альтернативной (табл. 2). Это означает, что статистические свойства процесса, генерирующего ряд, не изменяются со временем.

Далее были проведены тесты, позволяющие выявить зависимости в данных, проявляющиеся в связи между остатками, относящимися к разным наблюдениям, в разных временных периодах (табл. 3).

Таблица 2. Результаты тестирования гипотезы о нестационарности продаж (на панельных данных)

Наименование теста	Обобщенная модель (без индивидуальных эффектов)	Модель с индивидуальными эффектами	Модель с индивидуальными эффектами и линейным трендом по времени
Однородная альтернатива Левин – Ли – Чу 0–k (Levin – Lin – Chu Unit-Root Test)	Выбор длины лага на основе модифицированного критерия Шварца		
	$Z = -28{,}6$; $p < 0{,}0001$	$Z = -66{,}7$; $p < 0{,}0001$	$Z = -148{,}6$; $p < 0{,}0001$
	Выбор длины лага на основе модифицированного критерия Айкаике		
	$Z = -28{,}6$; $p < 0{,}0001$	$Z = -56{,}6$; $p < 0{,}0001$	$Z = -123{,}6$; $p < 0{,}0001$
	Выбор длины лага на основе модифицированного критерия Холла		
	$Z = -28{,}6$; $p < 0{,}0001$	$Z = -63{,}2$; $p < 0{,}0001$	$Z = -131{,}3$; $p < 0{,}0001$
Неоднородная альтернатива Има – Песарана – Шина (Im – Pesaran – Shin Unit-Root Test)	—	$W_{tbar} = -71{,}7$; $p < 0{,}0001$	$W_{tbar} = -114{,}2$; $p < 0{,}0001$

Источник: рассчитано авторами по обезличенным данным, предоставленным Первым ОФД.

Результаты тестов указывают на существование зависимости значений показателя от их значения в предыдущие периоды времени, что определяет важность включения в модель лаговой переменной. При включении авторегрессии первого порядка (AR(1)) случайные эффекты становятся незначимыми, что исключает из последующего анализа модели со случайными эффектами. Кросс-секционная зависимость указывает на наличие связи между наблюдениями, которая в том числе может возникать из-за пространственных эффектов, в связи с чем дополнительно проводится тестирование на наличие пространственных эффектов (табл. 4).

Проведенные тесты указывают на наличие пространственной зависимости данных (отмечена значимость как пространственного лага, так и ошибки), что обуславливает целесообразность использования пространственных регрессионных моделей.

Таким образом, проведенные тесты позволяют сделать вывод о том, что выстраиваемая модель должна:

- оценивать индивидуальные эффекты и игнорировать случайные эффекты;
- учитывать пространственную ошибку и лаг;
- учитывать наличие авторегрессии первого порядка (AR(1)).

Данным критериям соответствуют модели с фиксированными индивидуальными эффектами с включением авторегрессии первого порядка, учитывающие пространственный лаг и ошибку.

Для сравнения в таблице 5 представлены результаты построения панельных моделей, учитывающих пространственные зависимости, авторегрессию первого порядка и нет.

Как видно, модель, учитывающая авторегрессию первого порядка и пространственные эффекты, имеет более низкие значения критериев Айкаике и Шварца при более высоких значениях логарифма правдоподобия и R^2.

На эффективность и несмещенность получаемых оценок по подобранной спецификации модели указывают результаты тестов Вальда (оценки коэффициентов статистически значимы при $p < 0{,}05$), критерия Уайта (отсутствие гетероскедастичности в остатках модели, $BP = 84{,}822$, $df = 6$, $p < 0{,}0001$) и тест Дарбина – Уотсона (отсутствие автокорреляции в остатках модели, $DW = 2{,}07$; $p = 1$). На отсутствие в остатках пространственной связности указывает тест Морана ($Z = 1{,}77$; $MoranI = 0{,}0097$; $p = 0{,}034$).

Таблица 3. Результаты тестов на наличие ненаблюдаемых эффектов, последовательную корреляцию и кросс-секционную зависимость

Наименование статистического теста	Статистика, степени свободы (df), p-уровень	Вывод о принятии/отклонении нулевой гипотезы
Тест Вулдриджа на ненаблюдаемые эффекты (Wooldridge's test for unobserved individual effects)	$Z = -6{,}3$; $p < 0{,}0001$	Ненаблюдаемые индивидуальные эффекты в остатках значимы
Тест Бера – Соса – Эскудеро – Юна на AR(1)-серийную корреляцию (Bera, Sosa – Escudero and Yoon locally robust test)	$\chi^2 = 192{,}2$; $df = 1$; $p < 0{,}0001$	Наличие авторегрессии первого порядка (AR(1)) в остатках
Совместный тест Балгати и Ли на случайные эффекты и последовательную корреляцию при нормальности и гомоскедастичности идиосинкразических ошибок (Baltagi and Li AR-RE joint test)	$\chi^2 = 220{,}1$; $df = 2$; $p < 0{,}0001$	Наличие последовательной корреляции и/или случайных эффектов
Односторонний тест Бера – Соса – Эскудеро – Юна на ненаблюдаемые эффекты (Bera, Sosa – Escudero and Yoon locally robust test: one-sided)	$Z = -4{,}2$; $p = 1$	Случайные эффекты не значимы при учете в остатках авторегрессии первого порядка (AR(1))
Тест Вулдриджа для последовательной корреляции в моделях с фиксированными эффектами (Wooldridge's test for serial correlation in FE panels)	$F = 4{,}5$; $df_1 = 1$; $df_2 = 20102$; $p = 0{,}03$	Наблюдается последовательная корреляция
Тест Вулдриджа на последовательную корреляцию в ошибках после взятия первых разностей (Wooldridge's first-difference test for serial correlation in panels)	$F = 0{,}09$; $df_1 = 1$; $df_2 = 20046$; $p = 0{,}76$	Отсутствует последовательная корреляция в ошибках после перехода от исходных данных к их первым разностям
Тест Вулдриджа на последовательную корреляцию ошибок исходных данных (Wooldridge's first-difference test for serial correlation in panels)	$F = 159{,}4$; $df_1 = 1$; $df_2 = 20046$; $p < 0{,}0001$	Наблюдается последовательная корреляция в ошибках исходных данных
Тест Песарана на кросс-секционную зависимость (Pesaran CD test for cross-sectional dependence in panels)	$Z = 193{,}35$; $p < 0{,}0001$	Наблюдается кросс-секционная зависимость

Источник: рассчитано авторами по обезличенным данным, предоставленным Первым ОФД.

Согласно данным, представленным в таблице, коэффициент пространственного лага положительный, что указывает на наличие прямой связи. Таким образом, рост продаж наблюдается на фоне роста продаж в соседних территориях. В то же время коэффициент пространственной ошибки отрицательный. Следовательно, существует фактор, не включенный в модель, рост которого негативно сказывается на изменении объема продаж на соседних территориях. Невысокие значения коэффициентов пространственного лага и пространственной ошибки указывают на то, что они значимы, но не являются определяющими, по крайней мере для данного вида продукта в данном промежутке времени.

На текущий момент в открытом доступе находится ограниченное число данных в аналогичной детализации, что оказывает влияние на спектр учитываемых в моделях факторов. Введенная фиктивная переменная (выходной день недели) показала статистическую значимость на уровне менее 0,1. Будний день оказывает положительное влияние на рост продаж хлеба.

Таблица 4. Тесты на наличие случайных, индивидуальных и пространственных эффектов

Наименование теста	Значение	Вывод
Односторонний тест Балтаги – Сонга – Коха (Baltagi, Song and Koh LM-H one-sided joint test) H0: $\sigma_\mu^2 = \rho = 0$ против альтернативы, что $\sigma_\mu^2 \neq 0$ или $\rho \neq 0$	$LM - H = 1629{,}4$; $p < 0{,}0001$	Присутствуют случайные региональные эффекты и/или пространственная автокорреляция
Тест Балтаги – Сонга – Коха на маржинальный эффект (Baltagi, Song and Koh SLM1 marginal test) H0: $\sigma_\mu^2 = 0$ при предположении, что $\rho = 0$, против односторонней альтернативы, что $\sigma_\mu^2 \neq 0$	$LM = 5{,}3$; $p = 1$	Отсутствуют случайные эффекты
Тест Балтаги – Сонга – Коха на маржинальный эффект (Baltagi, Song and Koh LM2 marginal test) H0: $\rho = 0$ при условии отсутствия случайных эффектов ($\sigma_\mu^2 = 0$) против альтернативы, что $\rho \neq 0$	$LM_2 = 40{,}4$; $p < 0{,}0001$	Присутствует пространственная автокорреляция
Тест Балтаги – Сонга – Коха (Baltagi, Song and Koh LM-lambda conditional LM test) H0: $\rho = 0$ при условии возможного существования случайных эффектов (σ_μ^2 может быть равен или не равен нулю) против альтернативы, что $\rho \neq 0$	$LM^* - \lambda = 1{,}95$; $p = 0{,}051$	Присутствует пространственная автокорреляция
Условный тест Балтаги – Сонга – Коха (Baltagi, Song и Koh LM-mu) H0: $\sigma_\mu^2 = 0$ при условии возможного существования пространственных эффектов и односторонней альтернативы, что вариационный компонент больше нуля	$LM^* - \mu = 5{,}0$; $p < 0{,}0001$	Присутствуют региональные эффекты
LM-тест (Бреуша – Пагана) для сбалансированных панелей (Lagrange Multiplier Test (Breusch – Pagan) for balanced panels) для проверки наличия индивидуальных эффектов при поддерживаемой гипотезе об отсутствии серийной корреляции	$\chi^2 = 27{,}97$; $df = 1$; $p < 0{,}0001$	Наличие индивидуальных эффектов в панельной модели
LM-тест на наличие пространственного лага зависимой переменной (LM test for spatial lag dependence)	$LM = 3456{,}1$; $df = 1$; $p < 0{,}0001$	Пространственный лаг значим
LM-тест на пространственную зависимость ошибок (LM test for spatial error dependence)	$LM = 1629{,}4$; $df = 1$; $p < 0{,}0001$	Пространственная ошибка значима
LM-тест на наличие пространственного лага зависимой переменной при допущении пространственной ошибки: робастные оценки (Locally robust LM test for spatial lag dependence sub spatial error)	$LM = 1858{,}3$; $df = 1$; $p < 0{,}0001$	Пространственный лаг значим
LM-тест на пространственную зависимость ошибок при допущении пространственного лага: робастные оценки (Locally robust LM test for spatial error dependence sub spatial lag)	$LM = 31{,}6$; $df = 1$; $p < 0{,}0001$	Пространственная ошибка значима

Источник: рассчитано авторами по обезличенным данным, предоставленным Первым ОФД.

Обсуждение полученных результатов

Проведенный анализ заострил внимание на проблеме развития исследований на фискальных данных. Перспективы использования данных ОФД [Жабин, Турков, Волков, 2017; Самородова, Олейник, 2021], так же как и других высокочастотных данных, оптимистичны. Однако

Таблица 5. Результаты построения пространственно-панельных моделей с фиксированными индивидуальными эффектами

Наименование показателя	Модели с фиксированными индивидуальными эффектами, оценка коэффициента (стандартная ошибка, SE)		
	учитывающая AR(1), без учета пространственных связей	учитывающая пространственный лаг и ошибку, без AR(1)	учитывающая AR(1), пространственный лаг и ошибку
Константа	−0,037*** (0,0098)	−0,03*** (0,01)	−0,01 (0,009)
Авторегрессия первого порядка (Y_{t-1})	0,697*** (0,005)		0,466*** (0,006)
День недели, β_1	0,05*** (0,01)	0,044*** (0,009)	0,0148* (0,0095)
Пространственный лаг, ρ		0,002*** (0,00001)	0,0014*** (0,00001)
Пространственная ошибка, λ		−0,0009*** (0,00005)	−0,0007*** (0,00005)
Логарифм правдоподобия	−21830,1	−22677,1	−19913,8
Критерий Айкаике	43678,2	45362,1	39837,6
Критерий Щварца	43749,4	45393,8	39877,2
R^2	0,49	0,44	0,58

*** $p < 0{,}001$; * $p < 0{,}1$.

Источник: рассчитано авторами по обезличенным данным, предоставленным Первым ОФД.

в России очень мало публикаций, основанных на анализе высокочастотных данных, в открытом доступе, что актуализирует описание особенностей работы с ними.

Основные проблемы, связанные с качеством исходных данных, такие как некорректное отражение наименований и количества товаров, наличие данных с нулевой ценой, проблема регистрации в чеках товаров в виде наборов, выбросы, представляющие собой не относящиеся к розничной торговле высокие продажи товаров, выявленные в ходе предобработки данных, согласуются с проводимыми ранее исследованиями [Deshaies-Moreault, Harper, Yung, 2018; Aladangady et al., 2019; Lovics et al., 2019; Muth et al., 2020; Chen, Qian, Wen, 2020; Clark et al., 2021]. Эти проблемы в целом решаются до начала исследований и часто в исследованиях не упоминаются. Как правило, в руки исследователей попадают уже очищенные данные, сгруппированные по запросу [Cotti et al., 2020; Waldenström, Angelov, 2021; Timiryanova et al., 2023]. В России фискальные данные в дневной детализации по социально значимым группам товаров представлены на сайте Федеральной налоговой службы (ФНС) в разделе «Открытые данные» [ФНС, 2022]. ОФД, так же как и ФНС, значительно продвинулись в решении отмеченных проблем в результате широкого использования инструментов интеллектуального анализа данных, в том числе методов искусственного интеллекта и алгоритмов машинного обучения для задач классификации. В частности, команда по работе с большими данными Первого ОФД запустила платформу Продажи.рф. Фактически ими размещаются уже очищенные обезличенные данные, однако понимание особенностей формирования данных может быть важно для последующего их анализа.

Объективно учитывая, что с 2021 года абсолютно все предприятия и индивидуальные предприниматели перешли на передачу данных ОФД, в руках государства сосредоточен исчерпывающий объем данных о розничных продажах. Однако если речь идет о данных, полученных только от одного или группы ОФД, следует обратить внимание на стабильность выборки. Каждое отдельное ОФД не обладает данными генеральной совокупности. Клиенты могут переходить от

одного ОФД к другому, а следовательно, их база не только не полная, но и изменяющаяся. Таким образом, если для проведения трендового анализа можно просто ограничить выборку функционирующими на протяжении всего рассматриваемого периода предприятиями, то для пространственного анализа существует объективная проблема различной доли рынка ОФД в различных географических зонах. С целью решения данной проблемы ОФД объединяют свои данные и разрабатывают программное обеспечение, позволяющее получить сводную аналитику (например, распределенная система анализа розничных продаж с открытым исходным кодом Юпана). Другим вариантом решения проблемы являются переход на относительные показатели и стандартизация данных в пределах каждой географической зоны [Lee, Lee, 2022; Alfaro, Park, 2020; Gathergood et al., 2021; Carvalho et al., 2021; Powell, Leider, Léger, 2020], что и было реализовано в представленном исследовании.

Важным ограничением использования неагрегированных фискальных данных в целях моделирования является отсутствие данных в той же детализации. Сложно найти факторы, по которым представлены сведения в той же детализации по времени (минуты, час) и пространству, поэтому в модели включается небольшое количество объясняющих переменных [Bounie, Camara, Galbraith, 2020; Chen, Qian, Wen, 2020]. Кроме того, существуют различные проблемы автоматической классификации и последующей группировки данных в случае их получения из разных источников [Baker, Kueng, 2021]. Проблема отсутствия экзогенных переменных в той же детализации определила низкую предсказательную способность построенной в данной работе модели. В ней учтена только одна фиктивная переменная, характеризующая день недели. Таким образом, в перспективе следует заострить внимание на формировании баз высокочастотных данных, характеризующих различные сферы жизнедеятельности человека, а также на разработке классификаторов, справочников, развитии алгоритмов, обеспечивающих наложение данных из различных источников.

Проведенные тесты и построенные модели позволили обосновать, что новый тип данных может учитывать изменения не только во времени, но и пространстве, расширяя возможности для прогнозирования явлений в географической плоскости. Проведенные тесты однозначно указывали на наличие пространственных зависимостей, не только последовательную, но и пространственную корреляцию в данных. Включаемые коэффициенты, учитывающие как пространственный лаг, так и автокорреляцию ошибки, были значимы. Таким образом, спектр инструментов, используемых для анализа высокочастотных данных, может быть расширен методами пространственно-панельного анализа.

Раскрытие новых возможностей анализа очень важно, так как благодаря беспрецедентной детализации, в том числе по периодам, новый тип данных дает возможность глубже понимать поведенческие паттерны, быстрее реагировать на изменения и прогнозировать не только развитие розничной торговли, но и реакции населения на внешние и внутренние шоки, довольно часто возникающие в современных условиях (пандемия, резкий рост курсов валют и т. п.). Примером такого анализа является оцениваемый «Платформой ОФД» коронавирусный индекс (Covindex), отражающий изменение картины потребления семи товаров, ажиотажный спрос на которые сделал их «символами» пандемии коронавируса: антисептики, одноразовые перчатки, медицинские маски, кусковое мыло, туалетная бумага, лимон и гречка [РИА Новости, 2020]. Благодаря получению данных практически в реальном времени, инструмент позволяет принимать решения в более короткий период времени.

Заключение

Представленная работа предусматривает анализ данных чеков, включающих наименование, цену, количество, стоимость, место и время реализации товаров в разрезе 56 зон г. Уфы, полученных в обезличенном виде от ОФД, на предмет их качества и возможности применения

в исследовательской работе, в том числе для оценки эффектов во времени и пространстве. Работа концентрирует внимание на особенностях нового вида данных, возможностях и проблемах их использования в научных целях и, как следствие, не предусматривает разработки каких-либо управленческих рекомендаций предприятиям или муниципальным органам управления анализируемой территории.

В ходе проведенного исследования выявлены целый спектр проблем, характеризующих новый вид данных, и пути их решения. Одновременно на высокочастотных данных протестированы гипотезы о нестационарности продаж, на последовательную корреляцию и кросс-секционную зависимость, на наличие случайных и пространственных эффектов. Проведенный анализ показал, что методы пространственно-панельного анализа на высокочастотных данных позволяют увидеть закономерности изменения продаж во времени и пространстве, что раскрывает новые возможности их использования в моделировании развития торговли. Проведенное исследование является одной из пилотных работ, раскрывающих на конкретном примере для последующего научного обсуждения проблемы и возможности использования обезличенных высокочастотных данных, собираемых фискальными аппаратами в России.

Список литературы (References)

Андрианова И. Д., Рябинина Е. В. Налоговый контроль в период цифровой трансформации в России и зарубежных странах // Ключевые проблемы социально-гуманитарных наук в современной России: сборник научных трудов по материалам Международной научно-практической конференции / под общ. ред. Е. П. Ткачевой. — 2018. — С. 99–103.

Andrianova I. D., Ryabinina E. V. Nalogovyi kontrol' v period tsifrovoi transformatsii v Rossii i zarubezhnykh stranakh [Tax control during the period of digital transformation in Russia and foreign countries] // Klyuchevye problemy sotsial'no-gumanitarnykh nauk v sovremennoi Rossii: sbornik nauchnykh trudov po materialam Mezhdunarodnoi nauchno-prakticheskoi konferentsii / pod obshch. red. E. P. Tkachevoi [Key problems of social sciences and humanities in modern Russia: a collection of scientific papers based on the materials of the International Scientific and Practical Conference / ed. E. P. Tkacheva]. — 2018. — P. 99–103 (in Russian).

Жабин Д. В., Турков М. М., Волков Д. В. Потенциал использования информации оператора фискальных данных // Социальная политика и социология. — 2017. — Т. 16, № 5 (124). — С. 25–33. — DOI: 10.17922/2071-3665-2017-16-5-25-33

Zhabin D. V., Turkov M. M., Volkov D. V. Potentsial ispol'zovaniya informatsii operatora fiskal'nykh dannykh [Potential of using the information of the fiscal data operator] // Sotsial'naya politika i sotsiologiya [Social policy and sociology]. — 2017. — Vol. 16, No. 5 (124). — P. 25–33 (in Russian). — DOI: 10.17922/2071-3665-2017-16-5-25-33

Калинин А. М., Волин И. А. Информационные источники для расчета индекса потребительских цен: большие данные сети Интернет и систем ФНС России // Вопросы статистики. — 2022. — № 29 (1). — С. 44–51. — DOI: 10.34023/2313-6383-2022-29-1-44-51

Kalinin A. M., Volin I. A. Informatsionnye istochniki dlya rascheta indeksa potrebitel'skikh tsen: bol'shie dannye seti Internet i sistem FNS Rossii [Data sources for CPI: big data of the internet and the systems of the federal tax service of Russia] // Voprosy statistiki. — 2022. — Vol. 29 (1). — P. 44–51 (in Russian). — DOI: 10.34023/2313-6383-2022-29-1-44-51

Оксенойт Г. К. Цифровая повестка, большие данные и официальная статистика // Вопросы статистики. — 2018. — № 25 (1). — С. 3–16.

Oksenoyt G. K. Tsifrovaya povestka, bol'shie dannye i ofitsial'naya statistika [Digital agenda, big data and official statistics] // Voprosy statistiki. — 2018. — Vol. 25 (1). — P. 3–16 (in Russian).

РИА Новости. РИА Новости и «Платформа ОФД» запускают Covindex. — 2020. — [Электронный ресурс]. — https://ria.ru/20200430/1570773164.html (дата обращения: 5.10.2022).

RIA News. RIA Novosti i "Platforma OFD" zapuskayut Covindex [RIA News and OFD Platform launch Covindex]. — 2020. — [Electronic resource]. — https://ria.ru/20200430/1570773164.html (accessed: 5.10.2022, in Russian).

Самородова Е. М., Олейник Д. А. К вопросу о цифровизации экономической жизни общества (на примере деятельности Федеральной налоговой службы) // Вестник ОрелГИЭТ. — 2021. — № 4 (58). — С. 82–91. — DOI: 10.36683/2076-5347-2021-4-58-82-91

Samorodova E. M., Oleinik D. A. K voprosu o tsifrovizatsii ekonomicheskoi zhizni obshchestva (na primere deyatel'nosti Federal'noi nalogovoi sluzhby) [To the problem of digitalization of economic life of the society (on the example of the Federal tax service activities)] // Vestnik OrelGIET. — 2021. — No. 4 (58). — C. 82–91 (in Russian). — DOI: 10.36683/2076-5347-2021-4-58-82-91

Ткачёв И., Скобелев В., Старостина Ю., Агеева О. Контрольно-кассовая инфляция // Газета РБК. — 2020. — № 102 (3269). — https://www.rbc.ru/newspaper/2020/11/30/5fc0c9fe9a79472cdddcbd2f

Tkachev I., Skobelev V., Starostina Yu., Ageeva O. Kontrol'no-kassovaya inflyatsiya [Control-cash inflation] // Gazeta RBK [RBC Newspaper]. — 2020. — No. 102 (3269) (in Russian).

ФНС. Сведения о ценах и объемах реализации продуктовых товаров и горюче-смазочных материалов в субъектах Российской Федерации по данным контрольно-кассовой техники. — [Электронный ресурс]. — https://www.nalog.gov.ru/opendata/7707329152-fnsprice/ (дата обращения: 5.10.2022).

FNS. Svedeniya o tsenakh i ob"emakh realizatsii produktovykh tovarov i goryuche-smazochnykh materialov v sub"ektakh Rossiiskoi Federatsii po dannym kontrol'no-kassovoi tekhniki [Information on prices and sales volumes of food products and fuels and lubricants in the constituent entities of the Russian Federation according to the data of cash registers]. — [Electronic resource]. — https://www.nalog.gov.ru/opendata/7707329152-fnsprice/ (accessed: 5.10.2022, in Russian).

ЦБ РФ. Использование больших данных в финансовом секторе и риски финансовой стабильности: доклад для общественных консультаций. — М.: Центральный банк Российской Федерации, 2021. — 31 с.

Ispol'zovanie bol'shikh dannykh v finansovom sektore i riski finansovoi sta-bil'nosti: doklad dlya obshchestvennykh konsul'tatsii [Use of big data in the financial sector and risks to financial stability: public consultation report]. — Moscow: Central Bank of the Russian Federation, 2021. — 31 p. (in Russian).

Aladangady A., Aron-Dine Sh., Dunn W., Feiveson L., Lengermann P., Sahm C. From transactions data to economic statistics: constructing real-time, high-frequency, geographic measures of consumer spending // Finance and Economics Discussion Series 2019-057. — Washington: Board of Governors of the Federal Reserve System, 2019. — 37 p. — DOI: 10.17016/FEDS.2019.057

Alfaro I., Park H. Firm uncertainty and household spending // SSRN Working Paper. — 2020. — No. 3669359. — DOI: 10.2139/ssrn.3669359

Andersen A. L., Hansen E. T., Johannesen N., Sheridan A. Consumer responses to the COVID-19 crisis: evidence from bank account transaction data // CEPR Discussion Paper. — 2020. — No. DP14809. — DOI: 10.2139/ssrn.3609814

Baker S. R., Farrokhnia R. A., Meyer S., Pagel M., Yannelis C. Income, liquidity, and the consumption response to the 2020 economic stimulus payments // NBER Working Paper. — 2020. — No. 27097. — DOI: 10.3386/w27097

Baker S. R., Kueng L. Household financial transaction data // NBER Working Paper. — 2021. — No. 29027. — DOI: 10.3386/w29027

Baltagi B. H., Song S. H., Koh W. Testing panel data regression models with spatial error correlation // Journal of Econometrics. — 2003. — Vol. 117. — P. 123–150. — DOI: 10.1016/s0304-4076(03)00120-9

Bounie D., Camara Y., Galbraith J. W. Consumers' mobility, expenditure and online-offline substitution response to COVID-19: evidence from French transaction data // SSRN Working Paper. — 2020. — No. 3588373. — DOI: 10.2139/ssrn.3588373

Buda G., Hansen S., Rodrigo T., Carvalho V. M., Ortiz Á., Mora J. V. R. National accounts in a world of naturally occurring data: a proof of concept for consumption // Cambridge working papers in economics. — 2022. — 73 p. — https://doi.org/10.17863/CAM.93381

Carvalho B., Peralta S., Pereira dos Santosa J. What and how did people buy during the Great Lockdown? Evidence from electronic payments // CEPR (Centre for Economic Policy Research) working paper. — 2020. — Vol. 28. — P. 119–158.

Carvalho V. M., Garcia J. R., Hansen S., Ortiz Á., Rodrigo T., Rodríguez Mora J. V., Ruiz P. Tracking the COVID-19 crisis with high-resolution transaction data // Royal Society Open Science. — 2021. — No. 8. — 210218. — DOI: 10.1098/rsos.210218

Casey P., Castro P. Electronic fiscal devices (EFDs) an empirical study of their impact on taxpayer compliance and administrative efficiency // IMF Working Papers. — 2015. — No. 073. — DOI: 10.5089/9781475521023.001

Chacaltana J., Leung V., Lee M. New technologies and the transition to formality: The trend towards e-formality // Employment: Working Paper. International Labour Organization. — 2018. — No. 247.

Chen H., Qian W., Wen W. The impact of the COVID-19 pandemic on consumption: learning from high frequency transaction data // SSRN Working Paper. — 2020. — No. 3568574. — DOI: 10.2139/ssrn.3568574

Clark S. D., Shute B., Jenneson V., Rains T., Birkin M., Morris M. A. Dietary patterns derived from UK supermarket transaction data with nutrient and socioeconomic profiles // Nutrients. — 2021. — No. 13/1481. — DOI: 10.3390/nu13051481

CNEWS. ОФД объединяются для работы с большими данными: интервью директора по развитию бизнеса «Первого ОФД». — 2019. — [Электронный ресурс]. — https://www.cnews.ru/articles/2019-03-29_ofd_obedinyayutsya_dlya_raboty_s_bolshimi_dannymi (дата обращения: 2.10.2021).
CNEWS. OFD ob"edinyayutsya dlya raboty s bol'shimi dannymi: interv'yu direktora po razvitiyu biznesa "Pervogo OFD" [OFD unite to work with big data: interview of the business development director of the First OFD]. — 2019. — [Electronic resource]. — https://www.cnews.ru/articles/2019-03-29_ofd_obedinyayutsya_dlya_raboty_s_bolshimi_dannymi (accessed: 2.10.2021, in Russian).

Cotti Ch. D., Courtemanche Ch. J., Maclean J. C., Nesson E. T., Pesko M. F., Tefft N. The effects of e-cigarette taxes on e-cigarette prices and tobacco product sales: evidence from retail panel data // NBER working paper series. — 2020. — No. 26724. — DOI: 10.3386/w26724

Croissant Y., Millo G. Panel data econometrics with R. — EU, USA: John Wiley & Sons Ltd, 2019. — 301 p.

De Beer B., Tissot B. Official statistics in the wake of the Covid-19 pandemic: a central banking perspective // Theoretical Economics Letters. — 2021. — Vol. 11. — P. 695–723. — DOI: 10.4236/tel.2021.114047

Deshaies-Moreault C., Harper B., Yung W. Analysis of scanner data for the consumer price index at Statistics Canada. Chapter 17. The unit problem and other current topics in business survey methodology / ed. B. Lorenc, M. Bavdaz, G. Haraldsen, D. Nedyalkova, P. Smith, L.-Ch. Zhang, T. Zimmermann. — UK, Newcastle upon Tyne: Cambridge Scholars Publishing, 2018. — P. 237–252.

Dubois P., Griffith R., O'Connell M. The use of scanner data for economics research // CEPR Discussion Paper. — 2022. — No. DP16954.

Eurostat. Practical guide for processing supermarket scanner data. — EU: Eurostat, 2017. — 37 p.

Gathergood J., Gunzinger F., Guttman-Kenney B., Quispe-Torreblanca E., Stewart N. Levelling down and the COVID-19 lockdowns: uneven regional recovery in UK consumer spending // Covid Economics. — 2021. — No. 67. — P. 24–52. — DOI: 10.2139/ssrn.3798679

Gelman M., Kariv S., Shapiro M. D., Silverman D., Tadelis S. Harnessing naturally occurring data to measure the response of spending to income // Science. — 2014. — Vol. 345 (6193). — P. 212–215. — DOI: 10.1126/science.1247727

Guha R., Ng S. A. A machine learning analysis of seasonal and cyclical sales in weekly scanner data // NBER Working Paper. — 2019. — No. 25899. — DOI: 10.3386/w25899

Hall A. Testing for a unit root in time series with pretest data-based model selection // Journal of Business & Economic Statistics. — 1994. — No. 12 (4). — P. 461–470.

Kapoor M., Kelejian H. H., Prucha I. R. Panel data model with spatially correlated error components // Journal of Econometrics. — 2007. — Vol. 140, No. 1. — P. 97–130. — DOI: 10.1016/J.JECONOM.2006.09.004

Kokh L. V., Kovaleva Ju. V., Ivanova O. P. Big data in public administration // Proceedings of International Scientific and Practical Conference "Russia 2020 — a new reality: economy and society" (ISPCR 2020). — 2021. — P. 250–254. — DOI: 10.2991/aebmr.k.210222.049

Kolsrud J., Landais C., Spinnewijn J. Studying consumption patterns using registry data: lessons from Swedish administrative data // CEPR Discussion Papers. — 2017. — DP12402.

Leclair M., Léonard I., Rateau G., Sillard P., Varlet G., Vernédal P. Scanner data: advances in methodology and new challenges for computing consumer price indices // Economie et Statistique / Economics and Statistics. — 2019. — No. 509. — P. 13–29. — DOI: 10.24187/ecostat.2019.509.1981

Lee K. O., Lee H. Public responses to COVID-19 case disclosure and their spatial implications // Journal of Regional Science. — 2022. — Vol. 62, No. 3. — P. 732–756. — DOI: 10.1111/jors.12571

Lovics G., Szőke K., Tóth C. G., Ván B. The effect of the introduction of online cash registers on reported turnover in Hungary // MNB (Magyar Nemzeti Bank) Occasional papers. — 2019. — No. 137. — 24 p.

Millo G., Piras G. Splm: spatial panel data models in R // Journal of Statistical Software. — 2012. — Vol. 47, No. i01. — P. 1–38. — DOI: 10.18637/jss.v047.i01

Muth M. K., Okrent A., Zhen C., Karns Sh. Using scanner data for food policy research. — Elsivier Academic Press, 2020. — 342 p. — DOI: 10.1016/C2017-0-01027-3

Mutl J., Pfaffermayr M. The Hausman test in a Cliff and Ord panel model // Econometrics Journal. — 2011. — Vol. 14. — P. 48–76. — DOI: 10.1111/j.1368-423X.2010.00325.x

OECD. Implementing online cash registers: benefits, considerations and guidance. — Paris: OECD, 2019. — 73 p.

Powell L. M., Leider J., Léger P. T. The impact of a sweetened beverage tax on beverage volume sold in Cook County, Illinois, and its border area // Annals of Internal Medicine. — 2020. — No. 172, No. 6. — P. 390–397. — DOI: 10.7326/m19-2961

Rojas C., Wang E. Do taxes on soda and sugary drinks work? Scanner data evidence from Berkeley and Washington state // Economic Inquiry. — 2021. — Vol. 59, No. 1. — P. 95–118. — DOI: 10.1111/ecin.12957

Timiryanova V., Lakman I., Prudnikov V., Krasnoselskaya D. Spatial dependence of average prices for product categories and its change over time: evidence from daily data // Forecasting. — 2023. — Vol. 5, No. 1. — P. 102–126. — DOI: 10.3390/forecast5010004

Trivedi M. Regional and categorical patterns in consumer behavior: revealing trends // Journal of Retailing. — 2011. — No. 87. — P. 18–30. — DOI: 10.1016/J.JRETAI.2010.11.002

Verhelst B., Van den Poel D. Deep habits in consumption: a spatial panel analysis using scanner data // Empirical Economics. — 2013. — Vol. 47, No. 3. — P. 959–976. — DOI: 10.1007/s00181-013-0776-4

Waldenström D., Angelov N. The impact of COVID-19 on economic activity: evidence from administrative tax registers // CEPR Discussion Paper. — 2021. — No. DP16332.