NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Dmitrii Tereshchenko

COMPETITION AMONG RUSSIAN GROCERY STORES: DATABASE ON ST. PETERSBURG, 2017–2021

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: ECONOMICS
WP BRP 258/EC/2022

Competition among Russian Grocery Stores: Database on St. Petersburg, 2017–2021*

Dmitrii Tereshchenko†

Abstract

The Russian grocery retail industry has developed dynamically over the past two decades. The accompanying changes in competition and consumer behaviour make it an interesting subject for analysis, but full-fledged empirical studies in this area have not yet been conducted. This is largely due to the difficulty of accessing data on retail, which are often subject to commercial secrecy. However, in recent years the situation has changed and some data have become available to researchers. In this paper, we describe the data we obtained for the study of the grocery retailing industry in St. Petersburg. Particularly, we describe three blocks of data, including (i) store location data, (ii) socioeconomic characteristics of local markets, and (iii) sales data. For each data block, we outline the stages of data collection and processing, and provide basic descriptive statistics and graphs. In addition, we discuss the potential uses of the collected data in further research.

Keywords: retail industries, grocery sector, market studies
JEL classification: L0, L81

*This work is part of a larger research project on the Russian grocery retail industry, but I remain solely responsible for any errors, inaccuracies, or interpretations in this working paper. Sergey Kokovin (HSE University) is the main driving force behind our project and without his communicability and energy it would hardly have happened. The idea for this working paper originally came from Fedor Iskhakov (Australian National University), so I thank him for that as well as for his comments on intermediate versions of the paper. I am also grateful to Evgeniy M. Ozhegov (HSE University) and Alina Ozhegova (Norwegian School of Economics) for their help and advice during the data processing. These data were used by HSE University students during their term papers' work in the 2021/22 academic year. So I thank Otabek Mamatkulov and Darina Veryaskina, Ruslan Roschupkin and Maria Suslova, and Tatyana Yakushina for useful information discovered during their work and discussions of their term papers. The data described in this paper were obtained on terms of paid access from the Geointellect company. We are all very grateful to Geointellect for providing the data, as well as for the consultations that revealed its' specifics.

†HSE University in St. Petersburg, Department of Economics, email: dtereshchenko@hse.ru

1 Introduction

The grocery retail market in Russia has been developing dynamically since the late 1990s. The so-called "supermarkets revolution" has changed the landscape of Russian retailing dramatically, provoking changes in both consumer behaviour and the nature of competition. This process in Russia is late in comparison with the U.S. and Western Europe and has not yet been fully explored. Some papers describe the situation in the Russian retail sector in its early stages (see Robinson, 1998; Radaev, 2006, for example). However, there are no full-fledged empirical studies that include the estimation of structural models in this area yet. This is largely due to the difficulty of obtaining the data necessary to conduct this kind of research.

For many years, microdata on store activity was largely available only to the retailers themselves, as well as to the Federal Tax Service. Even now, only aggregated financial and accounting data is publicly available, disclosed by retailers at the company level only, not at the individual store level. However, there have been significant changes in recent years that have made some of the data more accessible. First of all, interactive map services have developed these days, allowing access to data on the locations of both stores and consumers.[1] In addition, since 2016, almost all firms in Russia are required to have so-called online cash registers, the data from which are transmitted to the Operators of Fiscal Data, which in turn transmits this data to the Federal Tax Service. The Operators of Fiscal Data themselves are private organizations and can sell their data in an aggregated form, which makes it possible to obtain information on demand and revenues.

This paper describes the new data available for research, as well as the possibilities of using it.

For our research, we received geocoded data on consumers and grocery retail-

[1] For example, Google Maps and Open Street Maps are known worldwide. In Russia, there are analogues such as Yandex Maps and 2GIS.

ers in St. Petersburg from one of the leading geoanalytics platforms in Russia, which put together, among other things, the data described above. As part of our agreement, we had access to the data on a paid basis and could periodically consult with the firm's analysts.

Note that this paper is part of an extensive and long-running research project on the Russian grocery retailing industry. Preliminary results from this project, as well as our future plans, are described in Gaivoronskaia et al. (2021).

The remainder of the paper is organized as follows. In section 2, we provide a brief overview of Russian grocery retail industry, with a focus on the context of the city of St. Petersburg. In section 3, we describe our data sources and summarize the procedure for collecting and processing the main blocks of data. In section 4, we explain our data cleaning and quality control procedure. In section 5, we provide descriptive statistics and graphs for our datasets. Section 6 contains proposals for future research opportunities involving the data described in the paper. Section 7 concludes.

2 Industry context

In our previous paper, we reviewed the main stages of development and the actual situation of the grocery retail industry in Russia (Gaivoronskaia et al., 2021). As a quick reminder, here we provide an overview of the current state of the industry. Next, we talk more specifically about the grocery retail industry in St. Petersburg.

2.1 Overview of Russian grocery retail industry

The modern stage in the development of the grocery retail industry in Russia, or the so-called "supermarket revolution", dates back to the late 1990s when many chain retailers emerged, later becoming industry leaders. Back then, chain retailers such as Magnit, Pyaterochka, and Dixy began their activities as a hard discounters

or cash and carry stores after the 1998 crisis. As they evolved, they changed their store format first to soft discounters and then to convenience stores, encouraged by rising household incomes until 2008. Another industry leader, Lenta, also began in the format of cash and carry in the 1990s, later became a hypermarket chain, and is now gradually switching to a format of supermarkets and convenience stores. It was this period, from about 1998 to 2008, that saw the most rapid development of modern store formats and consumption styles, somewhat similar to those in the U.S. and Europe. These trends were also complemented by the steady development of premium grocery chains, such as Azbuka Vkusa. Another major chain, Perekrestok, has been steadily developing its supermarket format since 1996.

However, the period after the global financial crisis and up to 2020 is characterized by economic stagnation and a gradual decline in households' real incomes. During this period, new hard discounter chains emerged and began to fill the low-price segment of the market. For example, Svetofor chain has been one of the most booming chains, entering the Top-10 grocery retailer chains in Russia. These trends have only been exacerbated by the COVID-19 pandemic and economic consequences associated with it.[2] In addition to falling incomes, another consequence of the pandemic has been mandatory and voluntary mobility restrictions and the subsequent development of e-grocery. As a result, specialized delivery services, online hypermarkets and online stores have emerged and expanded on the market. In addition, traditional retailers began to develop their own delivery services. Nevertheless, the role of the online sector in grocery retailing is still not too big. For example, the share of e-grocery in the total turnover of food sales in 2020 does not exceed 1%. It is worth noting, however, that hypermarkets have been significantly affected by the shifts caused by the pandemic, and many retailers are now either closing such stores or reformatting them.

[2]Note also that the Russian invasion of Ukraine in 2022 and its consequences were a huge shock to the entire Russian economy, which also hit the grocery retail industry, but that period is beyond the scope of this study.

The continued expansion of chains is accompanied by industry consolidation through periodic mergers and acquisitions. So, the current industry leaders are largely represented by companies managing a multi-brand portfolio. For example, X5 Group owns such brands as Pyaterochka (convenience stores)[3], Perekrestok (supermakets), and Karusel (hypermarkets). Another large company DKBR owns Dixy, the chain of convenience stores, along with Krasnoe & Beloe and Bristol, which are chains of liquor stores.[4] Additionally, in the face of declining solvency of households, as well as anti-COVID-19 restrictions, many large retailers have opened their own delivery services.[5] However, the level of concentration in the industry at the moment is not very high and noticeably lags behind similar indicators in the USA and Western European countries. In fact, the share of the ten largest chains is only 37.4% of the total turnover in grocery retailing.[6] Despite this, the grocery retail sector has been subjected to regular and noticeable control by the government. In particular, so-called "Law on Commerce", was adopted in 2009 and can be considered as the main legal act regulating grocery retailing industry. Among other things, this law regulates the relationship between retailers and suppliers, as well as limits the allowed market share of a chain retailer to 25 percent within the boundaries of a municipal district or region.

To summarize, we can say that the modern landscape of Russian retail is a combination of different formats from low to premium price segment. Other industry leaders worth mentioning are such foreign retailers as Auchan and METRO C&C, as well as Russian retailers O'KEY and VkusVill. Once again, we encourage

[3]Note that the official website describes the format of Pyaterochka as a proximity store, but we think they fit under the broad definition of convenience stores.

[4]This information about the two largest Russian retail groups is relevant for the period for which we have data. Let us note, however, that the industry is developing dynamically. For example, in 2021, X5 Group announced its plans to close the Karusel hypermarket chain and began to gradually reduce its' retail space. In the same year 2021 Magnit company purchased Dixy chain from DKBR.

[5]Besides, to compete with Svetofor and other emerging hard discounter chains, major market players such as X5 Group and Magnit have already established and begun to develop their own hard discounter brands to complement their core businesses.

[6]https://infoline.spb.ru/news/?news=207558.

all those interested to read the detailed industry overview in Gaivoronskaia et al. (2021). However, note that the level and nature of competition in the industry varies significantly by region. So we turn to describing the grocery retail industry specifically in St. Petersburg.

2.2 Grocery retailing in St. Petersburg

St. Petersburg is the second largest city in Russia in terms of size and economic development after Moscow. The population of St. Petersburg is more than 5 million people, and its area is more than 1400 square kilometres.

St. Petersburg is the birthplace of many large grocery retailers. The first stores such chains as Pyaterochka, Lenta, and O'KEY, among others, were opened in St. Petersburg. The grocery retail market in St. Petersburg is considered to be the most consolidated and competitive market in Russia. It is estimated that the Top-10 retailers in the city account for around 80% of the industry's turnover.[7] Recall that in Russia as a whole this figure does not exceed 40%.

Despite the above, the distribution of market power between the largest chain retailers in St. Petersburg is somewhat similar to the nationwide landscape. X5 Group is the market leader. Other large federal-level chains, such as Magnit and Dixy, are also in the Top-10, although not in the very top positions. A specific feature of the market is the stronger positions of large chains headquartered in St. Petersburg, such as Lenta, O'KEY among others.[8]

Such foreign chains as Auchan and METRO AG are represented in the Top-10 of large grocery retailers in St. Petersburg, which is similar to their position on the Russian market as a whole. They are complemented by the Finnish chain

[7] https://spb.fas.gov.ru/news/11258.

[8] Worthy of mention is the Intertorg company, which owned such brands as Narodnaja Sem'ja, SPAR, and Ideja and was one of the Top-5 retailers in St. Petersburg. Another important examples are the Prodovol'stvennaja birzha company which owned Lime and Polushka chains and the Novaja roznica company which owned the Estnyj chain. Both companies were among top retailers either in terms of turnover or in terms of the number of stores. All the companies closed around 2019.

Table 1: Market shares of major retail chains in St. Petersburg

Retailer / Chain	2017	2018	2019	2020
X5 Group	23.86	27.03	27.37	31.75
Lenta	13.36	13.72	13.3	17.55
O'KEY	11.67	10.96	9.94	10.12
Dixy	6.7	7.09	6.95	7.42
Magnit	4.69	5.14	5.76	7.24
Semishagoff	1.15	1.61	2.1	3.07
METRO AG	2.53	2.08	2.42	2.21
PRISMA	2.11	1.88	1.55	1.49
Vernyj	0.86	1.35	1.3	1.47
Auchan	4.89	1.83	1.43	1.15
Land	1.05	1.02	0.91	0.9
Intertorg	10.28	10.36	10	
Polushka	2.79	2.79		

Sources: Annual reports of the Office of the Federal Antimonopoly Service in St. Petersburg. Values for 2020 are preliminary estimates. https://spb.fas.gov.ru/news/10628, https://spb.fas.gov.ru/news/11258.

PRISMA.[9]

Note that one can observe significant spatial differences between local markets within St. Petersburg. For example, the premium segment stores are more represented in the city center, while the periphery is dominated by low and medium price segments.

It is also worth mentioning that as one of the most developed cities in Russia, St. Petersburg has a wider spread of e-grocery. In terms of online food trade, St. Petersburg lags far behind Moscow, but is strongly ahead of any other city in Russia. Thus, it can be stated that there is more intense competition in the grocery retail industry in St. Petersburg compared to most of Russia, both in the offline and online segments.

Another feature of the St. Petersburg market is the smaller share of hard dis-

[9] Although note that in 2022 PRISMA closed down its operations in Russia, and Perekrestok chain stores were opened on its facilities.

counters, which is associated with higher household incomes relative to the national average, as well as high rents. For example, one of the leaders in the growth of shopping space in Russia, the Svetofor chain is not yet widely represented in St. Petersburg, as well as other stores of a similar format.

St. Petersburg is recognized as a city of federal importance, which means that its status is equal to that of a region. Therefore, in accordance with the "Law on Commerce", it is treated as a single market with a limit of 25% of the share of each chain retailer. At the same time, St. Petersburg consists of 111 municipalities, each of which is also considered a single market in accordance with this law. Note that the law does not apply to a specific chain, but to the owner company. For example, the market share of X5 Group company in St. Petersburg exceeded 25% in 2018, 2019 and 2020.[10] As a result, the company has long been the subject of intense interest from the Federal Antimonopoly Service.

So, we believe that the grocery retailing industry in St. Petersburg is an interesting object for analysis.

3 Data sources and overview of the datasets

3.1 Towards the complete universe of data on the Russian grocery retail

From the perspective of the research project as a whole, we are interested in a wide range of sources and types of data about the Russian grocery retail industry.

To begin with, we are always able to use open data from the Federal Statistics Service (aka Rosstat) and make some aggregated conclusions based on it. Rosstat provides information on a wide range of socioeconomic indicators on its portal[11] and in various statistical yearbooks. In particular, more specific data on retailing

[10] https://spb.fas.gov.ru/news/11258.
[11] https://eng.rosstat.gov.ru/.

are published in the statistical yearbook "Commerce in Russia"[12], which has been published on an biannual basis since 2001. Rosstat generally provides data at national and regional level. Some data are available at the municipal level, but in a less user-friendly form.

For more detailed information, we can refer to consumer panel data. For example, the most widely used panel called "Russian Longitudinal Monitoring Survey" (RLMS) is publicly available. We plan to expand the general information from RLMS, which is representative only at the level of the whole country, by purchasing data from consumer panels devoted specifically to the grocery retailing industry in St. Petersburg. The best known providers of such data in Russia are NielsenIQ[13], GFK[14], Romir[15]. Besides the well-known problems with this kind of data (see Einav et al., 2010), one has to realise that these companies are not oriented towards cooperation in academic research, so it is not easy to negotiate with them and their data may not meet the requirements of sufficient detail. This is why we are also considering the possibility of collecting consumer data through self-administered surveys.

In addition, price information for many of the largest retail chains is available on their official data. Such prices can be parsed on a regular basis, tracking the dynamics of prices and their dispersion between different retailers. So far, we have price data on key retailers in St. Petersburg for several product categories on a month-by-month basis for 2021.[16]

However, the key to our project is geocoded data about retailers and consumers in St. Petersburg. In the remainder of the paper we describe just this part of the

[12] https://rosstat.gov.ru/folder/210/document/13233.
[13] https://nielseniq.com/global/ru/.
[14] https://www.gfk.com/ru/home.
[15] https://romir.ru/eng.
[16] Potentially, this data could be used to study differences in prices between cities, as retailers often have different versions of their websites for different cities and regions. Unfortunately, it is not possible to analyse price dispersion between locations in a large city such as St. Petersburg, as offline prices may differ from online prices as well as from other offline stores in the same chain.

data.

3.2 Overview of geocoded data on grocery retail industry in St. Petersburg

We obtained geocoded data about the grocery retail industry in St. Petersburg from Geointellect[17], one of the leading geoanalytics platforms in Russia (hereafter we will refer to them by name or as a "geoanalytics platform"). As part of our agreement, we had access to the data on a paid basis and were able to periodically consult with the firm's analysts. During consultations, we tried to describe to the firm's employees the ideal dataset that we would like to work with, and they, in turn, explained to us what real possibilities they have and which of their data are closest to our expectations. In the occasional negotiation, we agreed on the next piece of data that was sent to us in *.txt, *.csv, *.xlsx, and other formats. During the data cleaning and preprocessing, which will be described in detail in the next section, we converted the data to long panel format, if possible, and saved in .csv format for further analysis. In addition, we were provided with multipolygons with local market boundaries in *.shp and *.gdb formats.

In the current section, we give a summary of the data we collected, dividing it according to the purpose for which we obtained it. Given our goals and the capabilities of the geoanalytics platform, three main blocks of data were collected: (i) store location data, (ii) socioeconomic characteristics of local markets, and (iii) sales data.

3.2.1 Local market definition

The spatial polygons we have can be used to define local markets in several ways.

First, we can consider a municipality as a local market, because it is this level that is taken into account by the government when carrying out antimonopoly reg-

[17] https://geointellect.com.

Table 2: Descriptive statistics of areas (in sq km) for different definitions of local markets

	Municipalities	Postcodes
n	111	247
mean	13.06	4.00
std	17.82	5.84
min	1.07	0.00
max	106.27	39.70
sum	1450.11	988.45

ulation in accordance with the "Law on Commerce".[18] In our preliminary judgment, the market definition in this law cannot adequately reflect the nature of spatial competition in a big city. The descriptive statistics in the Table 2 show that the markets defined by the boundaries of municipalities are too large to be considered truly local. There are 111 municipalities in St. Petersburg with an average area of about 13 square kilometres. Note that this number is consistent with the official number of municipalities, so we have complete information on it. Also, the total area of the municipalities is roughly the same as that of the city.

The second, and more important, is that we can define local markets at the postal code level. This approach is preferable because it allows us to consider smaller markets compared to municipalities. We have information about 247 postal codes in St. Petersburg with an average area of about 4 square kilometres (see Table 2). Such an area can already be considered correct for determining the local market. For example, in Yang (2020), the average local market area is 1.8 square miles, which corresponds to about 4.7 square kilometres. Note that not all of St. Petersburg's territory necessarily belongs to any postal code. So the total area in all postal codes is less than the area of the entire city.

[18] Federal Law *N* 381-FZ "On the fundamentals of state regulation of commerce in the Russian Federation" or so-called "Law on Commerce" (adopted on December 28, 2009) can be considered as the main legal act related to the subject area in question. The adoption of this law has been repeatedly criticised by various researchers (Avdasheva and Shastitko, 2011; Avdasheva et al., 2015; Radaev, 2018), since the law was adopted without any expert evaluation.

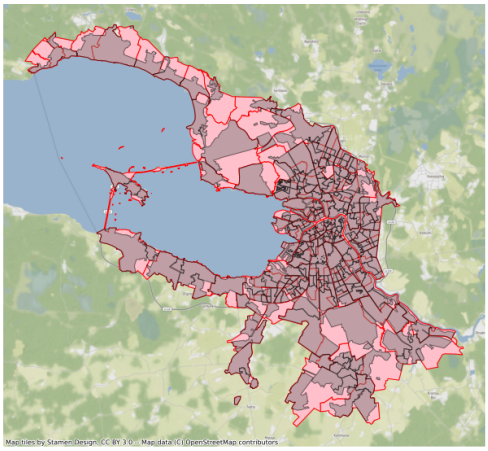

Figure 1: Municipalities (pink with red boundaries) and postcode zones (grey with black boundaries) comparison in St. Petersburg

3.2.2 Store location data

The key block of data we obtained is data on the location of grocery stores in St. Petersburg. For each store, we have access to information about its name, as well as the identifier of the firm that owns it. This allows us to identify to identify the chain affiliation of the store. Also, we know the address (mainly street name and house number) and geographic coordinates (latitude and longitude) of each store. In addition, each store is assigned a 6-digit postal code, which allows us to match the store to a particular local market. Figure 2 shows store locations in St. Petersburg in August 2020.

It is also very important that the data on the stores' locations are available to us on a dynamic basis. We currently have information on eight time slices for St. Petersburg, which include data for January and June 2017, January and July 2018, April and September 2019, and May and August 2020. Unfortunately, the gaps between the slices are not always equal to each other, but we have to work with what is available to us. This data allows us to identify store openings and

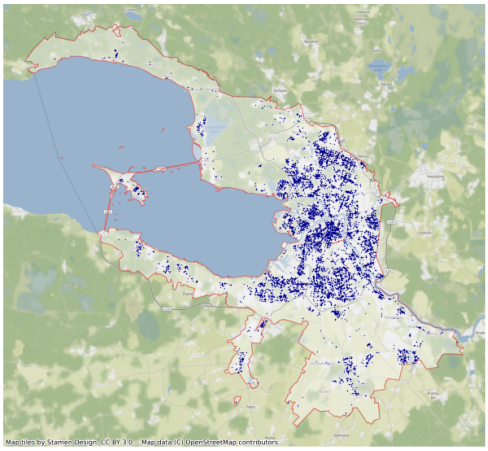

Figure 2: Stores locations within boundaries of St. Petersburg, August 2020

closings, the entries and exits of chain retailers from local markets. For example, if the store was present in the July 2018 slice and was already absent in April 2019, we consider that the store closed on the second date. Similarly, if the store was absent in May 2020 and appeared in August 2020, we consider it to have opened in August 2020.[19]

3.2.3 Socioeconomic characteristics of local markets

We are also able to match defined local markets with various socioeconomic characteristics of local markets. We have data on population, smartphone user density, prices of commercial real estate sales and rents, residential income, residential property rental prices, number of searches for car services, number of public transport stops. Most of these characteristics are related to demand and consumers, although commercial property rental data can be linked to retail costs. All charac-

[19] We also have similar data on other types of retail stores, such as clothes stores, electronics stores, and department stores. Although they are not the focus of our main interest, we can use them to identify retail clusters, which may be important. In the remainder of this paper, we will not describe in detail how to work with this data, since it is generally similar to working with data on grocery stores.

teristics are aggregated at postcode zone level.[20]

More specifically, the following variables are available to us:

- *pop* – population in the postcode zone, the number of individuals, annual data for 2017–2021.
- *flats* – the number of households in the postcode zone, annual data for 2017–2021.
- *devices* – the average number of unique smartphones per hex grid cell in the postcode zone for July 2017, October 2017, April 2019, October 2019.
- *price_sale* – average monthly sale price per square metre of commercial real estate, in rubles, for July 2018, October 2019, December 2019, February 2020, October 2020, December 2020.
- *price_rental* – average monthly rent per square metre of commercial real estate, in rubles, for July 2018, October 2019, December 2019, February 2020, October 2020, December 2020.
- *price_rental_res* – average monthly rent per square metre of residential real estate, in rubles, for June 2018 and March 2020. There are missing values in the data due to the fact that partially missing data on rents in some postcode zones for this period of time.
- *incomes* – average monthly income per family of two workers, in rubles, for June 2018 and March 2020. It is calculated on the basis of monthly rental data as the product of the average monthly rental price per square metre and square of the apartment, divided by the share of income that people are willing to

[20]This level of aggregation is chosen primarily because of the conventional definition of a local market. However, some of this data is also available at a more disaggregated level. For example, the number of households at the building level is public information and Geointellect collects and uses it, as well as calculating estimates of average income at the building level. This data could potentially be used to redefine the local market by abandoning a formal approach, in the spirit of Pennerstorfer and Yontcheva (2021).

spend on renting. These calculations were based on data from Domofond, a website of real estate ads for sale and rent.[21] There are missing values in the data due to the fact that partially missing data on rents in some postcode zones for this period of time.

- *income_model* – average monthly income per family of two workers, in rubles, calculated using the model built by the geodata provider in 2017 for St. Petersburg. The model was built on the basis of data on residential real estate rental ads, the cadastral value of 1 square metre of real estate, distances to the city center, the density of residential development and distances to subway stations.

- *auto* – the number of queries on automobile services for October 2020. Shows where on the map users of Yandex services search for organizations or services related to cars. It is based on the geopositioning data of Yandex services. The category is determined from the query. Requests with a search radius of less than 3 kilometres are taken into account. The values of this indicator are not integers, as the original data were in quite large hexes, which were cut into postal codes.

- *stops* – number of public transport stops for February and July 2018.

3.2.4 Sales data

We also have access to sales data. This data was originally obtained by the operators of fiscal data (hereafter in the paper – OFD).[22] A total of 19 OFDs operate in Russia as of March 2021. The data provided to the geoanalytics platform by three

[21] https://www.domofond.ru.

[22] The operator of fiscal data is a legal entity established specifically to receive, process, store and transmit fiscal data to the Federal Tax Service. Their activity is regulated by the Federal Law of 22.05.2003 N 54-FZ "On the application of control and cash registers for cash payments and (or) settlements using electronic means of payment".

of the OFDs covers 60% of the market and is available for December 2019.[23]

The unit of observation in this dataset is a combination of local market, store format and product category. The local market in this case is defined by the postal code zone, which ensures their compatibility with socioeconomic characteristics.[24] However, OFD data are also available to us at a more disaggregated level, as will be discussed in the next section. The store format variable can take three values, including (i) supermarkets, (ii) hypermarkets, and (iii) discounters and convenience stores. Note that the operator of fiscal data divides these formats by the criterion of the number of cash registers in a store. In this classification, convenience stores can have 1-2 cash registers, discounters – 3-6 cash registers, supermarkets – 7-13 cash registers, hypermarkets – 14 cash registers and more. The product category variable can take 15 values relating to the primary categories. These categories are Alcohol, Bakery products, Cakes, Cat and dog food, Dairy products, Dietary and health food, Fish, Fruits and vegetables, Instant food, Meat, Poultry meat, Soft drinks, Tobacco, Other foods, Non-food. This variable can also take 3 composite categories including Fresh (Dairy products + Meat), Ultra fresh (Bakery products + Cakes + Fish + Instant food + Poultry meat), and All category, which summarizes data for all categories.

For each unit of observation, i.e., for each combination of local market, store format, and product category, the data contain two key measures of retail store sales. These measures are the average number of checks per month (hereafter *avg_traffic*) and the average amount per check in a given month (hereafter *avg_check*).

[23] Unfortunately, this data does not include a considerable block of data on chain retailers, because some major chain retailers work with their own OFD, which does not disclose this information.

[24] Note that this data is originally provided at the level of hexagonal cells with a diameter of 100 metres, which allows for a more detailed analysis. Data aggregation schemes at the hexagonal cell level as well as at the postcode zone level are described in Section 4.

4 Data collecting and cleaning procedures

In this section, we describe the key steps for collecting and cleaning data. Note that during processing, the original files remained unchanged, while new *.csv files were created for further analysis. All operations were performed using a set of scripts, which ensures the reproducibility of the entire process. More specifically, we used Jupyter Notebooks with Python 3.10.

4.1 Spatial polygons for local market definition

The original shp-files provided to us used projected coordinate reference system (CRS) with EPSG code 3857 based on "World Geodetic System 1984" (WGS 84) datum and "Popular Visualisation Pseudo-Mercator" coordinate operation. In such form, these files were used to merge with retail data, as well as to draw maps. The store location coordinates were converted by changing the EPSG code from 4326 to 3857. However, to calculate the areas of local markets, the CRS was converted to "Albers Equal Area" projection for Russia.[25] It was also used to calculate distances.

4.2 Store location data

4.2.1 Data preprocessing

We received the data on the location of stores from the geoanalytics platform, in the form of 8 files with the *.txt extension in a colon-delimited format. Each file is a cross-section, in which a store is the unit of observation. The data were originally obtained by the geoanalytics platform from a large Russian mapping company, so essentially we can think of them as data that were parsed from an interactive map at different points in time (more specifically, we have slices for the following

[25] https://spatialreference.org/ref/sr-org/albers-equal-area-russia/.

periods: January and June 2017, January and July 2018, April and September 2019, and May and August 2020). As a result, the files we received contained both economic and geographic information that was important to us, and technical variables created by employees of both the mapping company and the geoanalytics platform. Also note that these maps (aka slices) are not originally designed for dynamic analysis, so the information between them is not necessarily consistent (e.g., the interactive maps are regularly updated, changing the composition, names, and ways of forming the various variables). This created a number of problems when processing the data. Our solutions to these problems are described below.

In the first stage, we combined data from different slices into a single panel. Technically, one could say that it was just concatenation of several cross-sections. The main tasks of this stage were data cleaning, as well as working with variables, including the creation of new variables. First of all, before merging the slices, "technical" variables that do not carry any meaningful information for us were removed from each of them, as well as variables for which all observations were missed (note that there were no critically important variables among them). We also identified variables whose names differed from slice to slice. All such variables were renamed to merge cross-sections correctly. Also, some variables were renamed to have a more meaningful name, as well as for better compatibility with other blocks of data. The variable containing the Russian-language store names was trasliterated.

The crucial task required to form the panel was the creation of a unique store ID variable. According to the information received from the geoanalytics platform, a unique store ID can be created using two variables: the firm ID and the branch ID. Both variables are encoded as numbers. Using these variables does allow creating a store ID, but the problem will be that such an identifier will vary from period to period, along with the branch ID indicator. To overcome this problem, we created a unique store identifier by combining the firm ID with address variables,

including city, street, and house. Potential inaccuracies in applying this method can occur when more than one store of a retail chain is located at the same address. However, an additional check showed that within each slice this method gives the result identical to the combination of firm ID and branch ID, which indicates the correctness of the chosen method. In addition, this identifier is time invariant and hence comparable across slices.

Since we are interested in analyzing the competition between chain retailers, we need to distinguish chain stores somehow. To solve this problem, we created a dummy variable that takes a value of 1 if the name of the store contains the Russian equivalent of the word "chain", and 0 otherwise. This allowed us to identify both major chain retailers and minor local chains. In addition, we created a dummy variable for major chains, which takes a value of 1 for chains owned by companies whose market share in St. Petersburg exceeded 5% according to data from the Federal Antimonopoly Service in at least one year (recall Table 1).[26] Using these two variables, we created a categorical variable that takes on different values for major chain stores, minor chain stores, and non-chain stores.

Two interrelated problems we had to solve were dealing with the store format variable and removing duplicates from the dataset. The point is that the geoanalytics platform keeps data in the form of "layers", where a separate layer is defined for each store format. The formats, in turn, are selected from a list created by the mapping company for its interactive maps. We got access to layers called "Stores, Supermarkets" and "Hypermarkets" related to the FMCG category. The problem is that the same store can be present in two layers at once, e.g. be listed as both a supermarket and a hypermarket. In order to remove the duplicates while preserving the format information, we created categorical variable taking three possible values: "Stores, Supermarkets", "Hypermarkets", and "Both". The latter category

[26] Another possibility that we can apply to identify large players on the market is the technical definition of chains with the largest number of stores, although such a measure may not reflect the real distribution of power on the market.

account for stores with ”double” format. The duplicates were then removed based on the store ID and period variables. Note that for the majority of stores (95-100% of the total, depending on the slice) there are no ”double format” problems.[27]

A separate problem was the detection of objects present in the panel, but not directly related to the grocery retail industry. One can think of such observations as inaccuracies arising from the subjectivity of assigning a format to a outlet on the interactive map. For example, we have quite a few observations concerning bakeries, as well as fruit and vegetable stalls. Other cases are liquor stores, as well as dollar stores. Presumably, these groups are not strategic competitors of large grocery chains, so we will exclude them from the sample in the future. In addition, their presence in the slices is not stable from period to period due to changes in the layers to which these objects belong. For example, the largest chain of liquor stores in Russia, Krasnoe & Beloe, appears in our data only from April 2019, although in fact this chain came to St. Petersburg somewhat earlier, in 2017. The thing is that until April 2019, stores of this chain were not marked on the interactive map as ”Stores, Supermarkets”. Similar problems apply to bakeries and dollar stores. So, in the case of bakeries, we create a dummy variable indicating that the outlet belongs to this group. The same is for fruit and vegetable stalls. In the case of liquor stores and dollar stores, we will exclude major chains explicitly, if necessary, without additional variables.

In addition, we explicitly exclude certain objects from the analysis. These include the Soyuzpechat newspaper kiosk chain as well as Russian Post offices. Both have many outlets and may sell food, but this is not their core business.

Another problem, bringing confusion to the analysis of the dynamics of the industry, was the presence of a network of payment terminals in some slices. These

[27] With all of the above in mind, note that the store format variable should not be fully trusted. In reality, there is an obvious difference between a supermarket and a hypermarket, both in size and in other parameters. However, the employees of the mapping company or store owners may be interested in getting the store into more layers or search queries, which encourages them to choose more formats in the interactive map settings.

objects obviously do not belong to the grocery retail industry and were therefore removed from the data.

Given the actions described in this subsection, we have formed a panel data, in which the unit of observation is defined by the store ID and the period. The panel contains data for both St. Petersburg and the Leningrad region. The panel is unbalanced because there are observations for each store only for those periods when it was present in the slice.

4.2.2 Completing the data on the store location

The second step in the data processing of store locations was to verify their geographical affiliation, including (i) clarifying information on postal codes to assign stores to local markets, and (ii) identifying stores located in St. Petersburg.

The postal code is a key variable for linking stores to specific local markets. The values of this variable were obtained using geospatial techniques, namely, each value was assigned on the basis of getting the geographic coordinates of the store within the boundaries of the postal code zone. What is important is that we use these same postal code zones to define local markets.[28]

The problem with this variable is the presence of missing values related to the way it is calculated, which is described above. As we noted in section 3, spatial polygons used to define local markets do not cover the whole territory of St. Petersburg (recall Figure 1). If a store did not fall within any of the polygons, its postal code was assigned zero value. It creates problems when assigning stores to local markets. Another issue could be getting the store exactly on the common boundary of two neighboring polygons. As a result, we have from 121 to 201 stores with missing postal codes, depending on the period, which is less than 2% of the total number of stores.

[28]This variable was originally created by the geoanalytics platform staff using the boundaries of postal code zones dating back to around 2019. This can lead to a loss of relevance, since postal codes can change over time. However, in terms of attribution to local markets defined by the same polygons, the use of this particular variable is justified.

To overcome this problem, we used a geospatial technique. For each store with a missing postal code, we assigned a value of this code from its nearest neighbor, i.e., the store located at the minimum straight line distance.[29] So, we got a postal code variable with no missing values.

Recall that the originally received files contained information about stores located both in St. Petersburg and the Leningrad region. Since in our project we are interested in the competition of retailers in a big city, it is important for us to distinguish the first group from the second. In order to do this, we applied roughly the same scheme, i. e. geospatial analysis. Specifically, we used the geographical coordinates of the stores and the polygon with the geographical boundaries of St. Petersburg. For this purpose, we took the polygons of municipalities and merged them into one. This method can be considered reasonable, because we have already shown above that municipalities cover the whole territory of the city, unlike postal code zones (recall Table 2). Thus we have filtered the data, leaving only observations related to St. Petersburg.[30]

As a result of the procedures described in this subsection we received a panel of stores located in St. Petersburg, each with a specific postal code value, which will allow us to further associate each store with a specific local market.

[29] Note that we also had another postal code variable that was originally downloaded from the interactive maps. This variable seems like a natural candidate to fill in the missing values. However, we did not use this variable for several reasons. First, we don't know exactly what methodology is used to determine postal code values in interactive maps. It can suffer from human and technical errors that arise during the generation of interactive maps, which were described above. In addition, Geointellect experts raised doubts about the quality of this data. Second, it also has missing values. Third, a comparison of the available values of the two postal code variables showed significant differences, which means that using them together would be inconsistent. Once again, we are not primarily interested in the actual postal code values, but rather in the possibility of using them to define local markets. Therefore, the use of geospatial techniques to recover missing values seems to us preferable.

[30] At first glance, the other solutions may seem to be more straightforward. To begin with, we have several variables at our disposal to potentially identify stores located in St. Peterburg. The most obvious way to do this is to use the city variable, which takes the names of settlements, including St. Petersburg and small towns in the Leningrad region, as its values. Another straightforward way to distinguish stores located in St. Petersburg is to determine the city affiliation based on the six-digit postal code. Addresses in St. Petersburg expected to have postal codes beginning with the digits "19". Addresses in the Leningrad region, in turn, should have postal codes beginning with the digits "18". Ideally, using the city variable as well as each of the two postal code variables should produce similar results. In practice, however, we were unable to obtain consistent results, and misidentifications were found in all three variables.

4.2.3 Creating a stores' "coding guide"

The next step after the initial preparation and processing of data was the formation of a "coding guide" containing key information about all stores. This guide is a table in which a store is the unit of observation, with each store occurring only once. To analyze the dynamics of the industry, this guide stores the period of the first and last appearance of the store in the data. It also contains basic economic and spatial information about stores. The main challenge in forming this guide is to select such information in the most consistent way possible.

Note that for each store, we can have up to eight observations in the original data (by the number of periods or slices). Roughly speaking, we just need to remove the duplicates from the previously created panel of stores. The problem is that the original data are not quite suitable for dynamics analysis, because it was not created for this purpose at all. In particular, the values of several key variables are revised and adjusted from period to period. Among these variables are potentially the store name, mailing address, postal code, and geographic coordinates. Note that changes in these variables likely do not reflect objective economic processes, but rather are caused by human and technical errors and occasional data adjustments.[31]

We used two different strategies to obtain data that were consistent in the dynamics. First, we simply took the last available value. This approach is sensible, assuming that all adjustments in the interactive maps are positive, i.e., they improve the data. This approach may be more consistent if the adjustments in the interactive maps are not necessarily positive.[32] With all this in mind, for each of

[31] For example, minimal changes in latitude and longitude do not mean that the store has moved, say, 10 metres south, but rather are the result of data adjustment on the interactive map. At the same time, the store's name can change both in real life or virtually on the interactive map, but both of these changes do not affect the dynamics in the industry.

[32] Note that we do not know the exact origin of these adjustments. Presumably, the changes in the interactive maps can be the result of the actions of the mapping company's employees, or the initiative of the owners or managers of a particular store.

the questionable variables, we save both the last value and the mode. Next, in this subsection, we present a quick comparison of these indicators.

After removing ”duplicates” from the dataset with 76080 store-period observations, we have 20452 stores in St. Petersburg that existed in at least one of the periods available to us. The good news is that the mailing address mode for each store is the same as the last value, so it doesn't matter what you leave in the guide. Since one variable is enough, we leave only the last value in the guide. For the other variables, however, there are differences between the mode and the last value. For example, we found 1475 cases of mismatch for store names and in 68 cases it affected the value of the chain attribution variable, which was derived from the name. This mismatch should not cause panic, because when analyzing the competition of large grocery chains, we care more about the firm's identifier than the specific letters in the store name.[33] Mode and last postal code values are different for only 5 stores, so choosing from these two metrics shouldn't affect the results much.

However, we found almost 4000 observations for which the last longitude and latitude do not coincide with the mode. That's quite a lot, so it was important for us to understand how big the differences were. To do this, we calculated the difference between the mode and the last coordinate value for each store. A descriptive analysis of the derived variable showed that for most stores the difference between the mode and the last coordinate value is so small that it can be neglected.[34]

4.2.4 Forming the final dataset

The last step in preparing data on store locations was creating a balanced panel of stores in a long format. Recall that after initially concatenating the data from

[33] However, we have adjusted the chain attribution variable. Now the store will be referred to the chain if the Russian analog of the word ”chain” is either in the mode or in the last value of the name variable.

[34] Only for 294 stores this difference, measured in degrees, appears in the third decimal place, and for the rest only in the fourth or more.

8 slices, we obtained a dataset of 76080 observations with information on 20452 stores. It makes sense that not all stores exist in all 8 periods, because some stores exit the market while new stores enter.

The problem is some dynamic inconsistency in our data. More specifically, we have stores that ”disappear” from the data and then reappear. We have 1587 such cases in total. We assume that these ”disappearances” are not likely related to the real industry processes. Since the gaps between periods are quite long, up to several months, it is unlikely that a store actually left the market, say, by closing for repairs, and then returned to the market. So this is most likely a consequence of technical errors in the interactive maps, and these gaps need to be corrected.

To achieve this, we expanded the original panel so that it contained $20452 \times 8 = 163616$ observations, i. e. 8 observations (periods) for each store. We created a variable which equals 1 if the store was present in the market in a particular period, and 0 if it was absent. We then replaced with 1 all the 0's that are between the other 1's in the dataset ordered by store ID and period.

Finally, we merged all of our store-period pairs obtained with the data from the guide created earlier. In this way, we got a balanced panel in a long format, which can be used to analyze the dynamics in the industry in a consistent way.

4.3 Socioeconomic characteristics of local markets

Working with socioeconomic data was fairly straightforward. The original data files were obtained separately for each indicator in *.xlsx or *.txt files. The original data was presented in cross-sectional or wide panel format. After initial processing, the data, still separately for each indicator, were saved in long panel format as *.csv files.[35] The main problem with this data is that it is presented for different time periods, which do not match each other or the other blocks of data.

[35]There is only one exception. The data on population and number of households were originally presented in a single file and saved to a single file.

4.4 Sales data

The sales data was originally provided to us in a fairly well-structured form in *.csv format. The initial processing involved simply renaming variables and categories to be more informative. However, the specifics of OFD data impose certain restrictions on the interpretation of available variables and operations with them. In this subsection, we describe our work with sales variables taking these specifics into account.

The key problem is that individual data for each store cannot be provided for commercial confidentiality concerns. For the same reason, the data received by Geointellect from the OFD were computed according to the following algorithm. First, the entire territory of St. Petersburg, as well as the Leningrad region, was divided into cells using a hexagonal grid (the radius of each cell is about 100 metres). A buffer with a radius of 1000 metres was then placed around the centroid of each cell. If the buffer received 3 or more stores with the given format and selling a particular product category, then the sales data were calculated based on information about all stores with the given format and selling a particular product category that were put in the buffer. If 0, 1 or 2 sales outlets of a given format selling a certain category of goods fell within the radius, then a zero value was taken for the corresponding cell.

Figure 3 provides a sample visualization of OFD data construction. Hexes within the boundaries of a particular postal code zone are highlighted in color. Values in cells with blue borders cannot be calculated because of commercial confidentiality concerns mentioned above. The sales variable values in the yellow cells are the equal and are calculated based on the same three stores. Postcode-level data are calculated based on the values in the yellow cells only.

In the general case, first, total number of checks, the average number of checks per store and the average check were calculated for each hex whose buffer zone

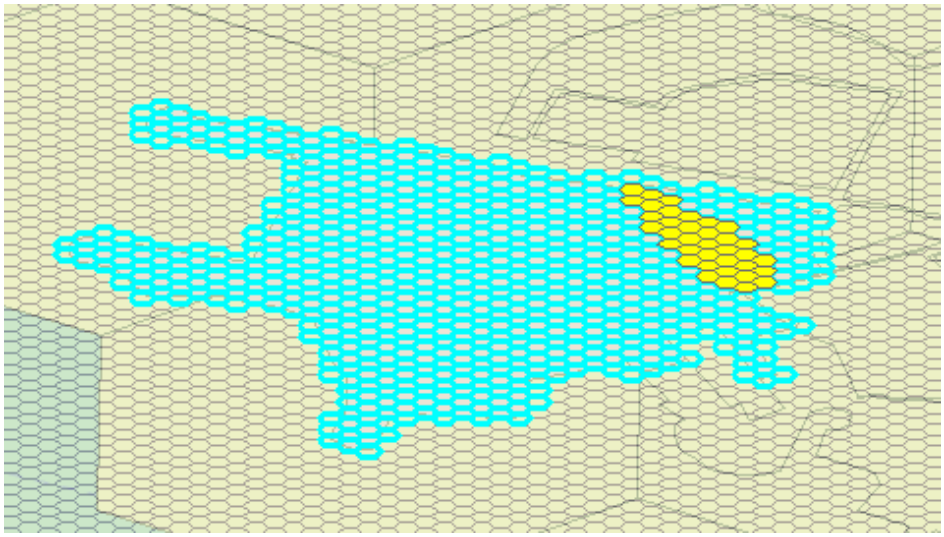

Figure 3: Sample visualization of OFD data construction

included at least three stores of a given format with data on the sales of a particular product category. Then, obtained data was aggregated at the postcode-level for cells within particular postcode zone. With this procedure it turned out that many of the cells next to each other had the same value, as they had the same stores in their buffers. Therefore, only relative indicators, namely the average number of checks per store and the average check, are suitable for analysis at postcode zone level, since the total number of checks indicator suffers from the problem of multiple counting, summing up checks from the same stores many times. Another problem is that, in fact, the data for a particular postal code can be calculated taking into account stores from neighboring zones, because the hex buffer zones can extend beyond the boundaries of the postal code zone.

Nevertheless, even taking into account all of the above features, we consider that the application of these measures is not entirely meaningless. So we use them to calculate the local market shares held by store formats in each product category. To do this, we first calculate the total revenue in the local market of each format in each product category as the product of the total number of checks and the average check. After that, we calculate the total revenue of each product

category in each local market as the sum of the revenue of the three formats. We then calculate each format's share of the total revenue for each product category. We understand that total revenue measures calculated on the basis of the total number of checks also suffer from the problem of multiple counting. However, we believe that our calculated shares are less affected by this problem, because they are relative indicators, and therefore can be used in further analysis.

We also calculate the average revenue of each store format in each product category as the product of the average number of checks and the average check, keeping in mind the limitations in the use of this indicator.

The listed problems associated with the aggregation of postcode-level data encourage us to use hex-level data directly in our analysis. However, this is associated with a heavy computational burden. Defining unit of observation as a combination of postal code, store format, and product category, we have 5697 observations, while combination of hex with store format and product category give us 1744499 observations in St. Petersburg.[36]

5 Data description

This section presents some descriptive analysis of the datasets we created.

5.1 Store location data

5.1.1 Industry-wide dynamics

Table 3 shows the dynamics of the number of grocery stores in St. Petersburg during the period under review, including the dynamics of chain stores. As one can see, the industry is quite dynamic, because the growth rate of the number of stores from period to period can be substantial, while the intervals between

[36] Both values are reported with the removal of missing values caused by commercial confidentiality concerns described above.

Table 3: Dynamics of the number of grocery stores in St. Petersburg

Period	Number of grocery stores		
	Major chains	Minor chains	Non-chain
Jan 2017	1141	1480	5626
Jun 2017	1199	1484	5995
Jan 2018	1292	1530	6141
Jul 2018	1357	1471	6186
Apr 2019	1415	1394	5816
Sen 2019	1472	1267	5607
May 2020	1300	1306	5748
Aug 2020	1329	1363	5956

periods are relatively small. An interesting observation is that the share of chain retailers, including both major and minor chains, in the total number of stores almost does not change over time and ranges from 31 to 33%. Also, one can notice that the number of major chain stores is increasing over time, although with slight fluctuations, while the share of minor chains is decreasing. These trends are quite consistent with our knowledge of the industry. For example, the decrease in the number of major chains' stores in August 2020 can be explained by the final departure of Intertorg from the market. We would also remind you that Prodovol'stvennaja birzha, which owned the Polushka chain among others, was closed in 2019. Its retail space was partially occupied by the major chains' stores.

We now turn to the analysis of the dynamics of store openings and closings presented in Table 4. The dynamics shown raises certain doubts, especially for non-chain stores, the number of entries and exits of which are subject to too many fluctuations for such a short period. These fluctuations can partially be explained by the problems described in the section 4, namely that the appearance of the store on the interactive map depends on both the representatives of the store itself and the specialists of the interactive map provider. Thus, excessive fluctuations can be

Table 4: Entries and exits of grocery stores in St. Petersburg

Period	Entries			Exits		
	Major chains	Minor chains	Non-chain	Major chains	Minor chains	Non-chain
Jun 2017	116	171	1110	58	167	741
Jan 2018	170	255	1232	77	209	1086
Jul 2018	117	232	774	52	291	729
Apr 2019	119	264	1096	61	341	1466
Sen 2019	92	115	683	35	242	892
May 2020	112	188	998	284	149	857
Aug 2020	49	98	431	20	41	223

explained, apparently, by the cycle of updating interactive maps. This seems to be the case for both chain and non-chain stores. For example, January 2018 and April 2019 have seen increases in the number of openings and closings for both chain and non-chain stores, although we are not aware of any objective economic processes that could have affected this.

At the same time, the dynamics of the entries and exits of the major retailer chains seem to be more credible compared to non-chain stores and stores of minor chains. Apparently, the managers of the large chains follow the representation of their stores on interactive maps more closely, and it is easier for the interactive map provider's staff to track the dynamics of the famous retailers. Thus, we believe that at least the dynamics of chain stores' entries and exits reflect objective economic processes.

5.1.2 Major chains' dynamics

Table 5 presents the dynamics of the number of stores for chains that were among the Top-10 grocery retailers in St. Petersburg at least in one of the periods. Note that the dynamics presented in the table adequately reflect the economic processes taking place in the periods under consideration. Recall from the section 2 that several major retailers owning, among others, such chains as Narodnaja sem'JA,

SPAR, Polushka, and Estnyj were closed around 2019. Another important observation is the decreasing rate of new store openings in 2020. This is not surprising, as many major retailers have explicitly declared that they will not open new stores in a pandemic. Recall also that Russia's largest retailer, X5 Group, had a market share over 25% in St. Petersburg in 2018, 2019 and 2020. Because of this, the Federal Antimonopoly Service banned the company from opening new stores in St. Petersburg in 2019 and 2020 in accordance with the law on trade. However, we can see from the table that the Pyaterochka and Perekrestok chains owned by X5 Group opened new stores during these periods. According to the company itself, this is not a violation of the law because, at least in 2019, these numbers include stores opened before the company even knew it had crossed the threshold.

It is also noteworthy that Table 5 does not include data on such large chains as Lenta and O'KEY. Both retailers are in the Top-5 in terms of turnover, but are not even in the Top-10 in terms of number of stores, as they traditionally rely on the hypermarket format, although this strategy has changed somewhat during the pandemic.

5.1.3 Summary of local markets' structure and dynamics

Now we turn to the analysis of the structure and dynamics of local markets, defining the local market as the postal code zone in a specific time period. In this case, we have $242 \times 8 = 1936$ observations about local markets. Figure 4 shows the distribution of local markets in terms of the number of chain and non-chain stores. Obviously, non-chain retailers dominate in terms of the number of stores, as well as have a wider range of values. That is, the structure of the local markets is quite diverse in terms of the representation of non-chain stores. At the same time there are fewer chain stores on average, and the spread in the structure of local markets is noticeably narrower in comparison with non-chain stores. This may indicate the presence of certain strategies of chain retailers in terms of their representation on

Table 5: Dynamics of Top-10 grocery retail chains by number of stores in St. Petersburg

Chain	2017		2018		2019		2020	
	Jan	Jun	Jan	Jul	Apr	Sen	May	Aug
Pyaterochka	316	342	380	399	431	464	495	490
Magnit	176	190	220	238	254	283	328	338
Dixy	287	293	295	296	293	290	294	310
VkusVill	0	0	0	2	20	51	114	112
Perekrestok	45	48	59	80	90	96	103	110
Velikolukskij mjasokombinat	138	130	128	124	127	125	115	105
Belorusskij dvorik	55	60	79	80	98	110	104	103
Semishagoff	62	65	72	81	86	88	93	96
Ermolino	10	18	46	58	86	93	93	95
Vernyj	48	47	51	60	75	77	81	88
Ankom	62	61	58	60	64	63	61	64
Narodnaja sem'JA	169	166	164	170	162	155	0	0
SPAR	53	63	82	89	98	99	0	0
Polushka	123	128	152	119	62	0	0	0
Estnyj	141	132	112	24	0	0	0	0

Table 6: Number of grocery stores in local markets by chain affiliation

Chain affiliation	count	mean	std	min	25%	50%	75%	max
Major chains	1768	5.9	3.9	0	3	6	8	27
Minor chains	1752	6.4	4.5	0	3	6	9	27
Non-chain	1936	24.3	16.7	0	13	23	33	124

the local markets. Figure 5 and Table 6 confirm that both major and minor chains have certain natural limits on their representation in local markets.

Tables 7, 8 summarize the patterns of behavior of major chain retailers on local markets in St. Petersburg. It can be seen from Table 7 that the maximum number of stores on the market does not exceed 10 for any chain. However, one can note that grocery retail chains usually have either 0 or 1 store in the local market. The exceptions are the three largest chains, for which it is quite normal to have 2 or 3 stores in the local market. At the same time, Table 8 shows that it is typical for all large chains not to change the number of stores in the local market. If a chain

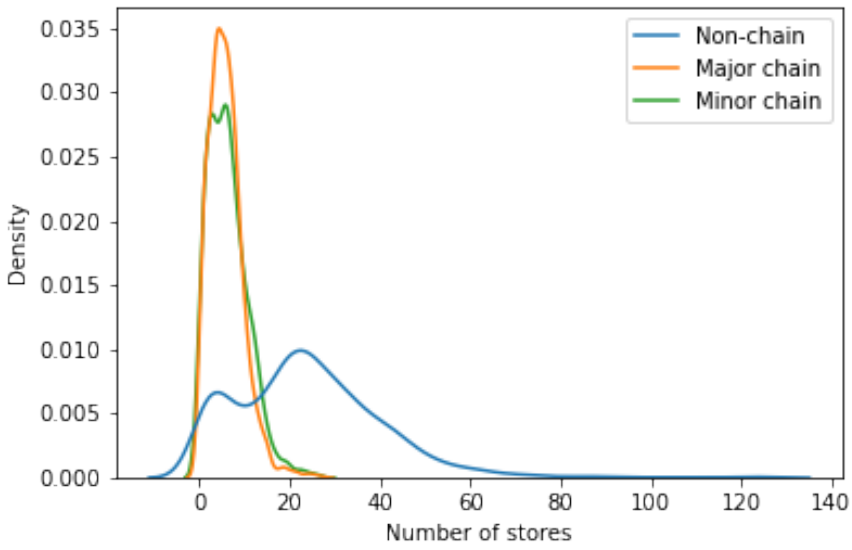

Figure 4: Local markets' size distribution in St. Petersburg

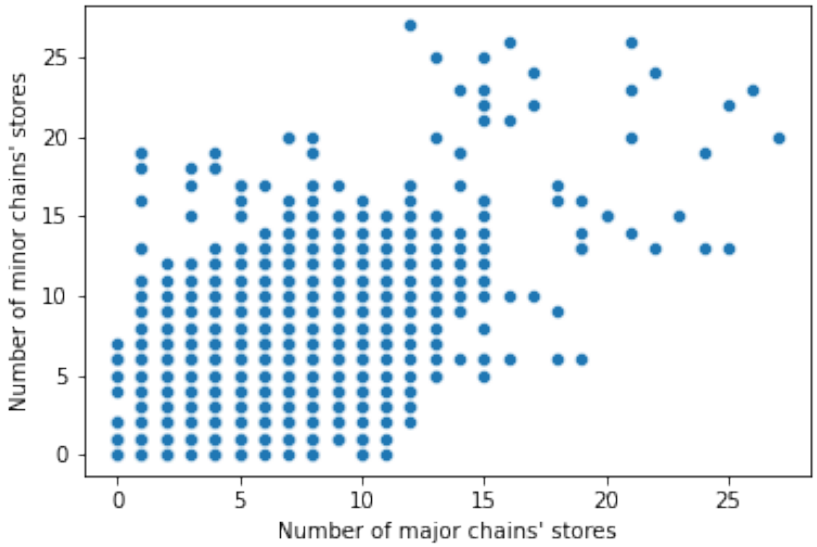

Figure 5: Representation of major and minor grocery retail chains in local markets in St. Petersburg

Table 7: The presence of major chains in local markets

Chain	Markets	Firm's local size									
		0	1	2	3	4	5	6	7	8	9
Pyaterochka	1417	519	513	390	240	139	92	20	18	3	2
Dixy	1155	781	490	361	169	81	29	12	6	7	0
Magnit	1123	813	534	369	162	35	13	6	3	1	0
Narodnaja sem'JA	634	1302	398	148	60	28	0	0	0	0	0
Perekrestok	478	1458	343	117	18	0	0	0	0	0	0
SPAR	404	1532	336	58	8	2	0	0	0	0	0
Lenta	284	1652	268	15	1	0	0	0	0	0	0
O'KEY	178	1758	173	5	0	0	0	0	0	0	0
Karusel	98	1838	98	0	0	0	0	0	0	0	0

Table 8: Dynamics of major chains in local markets

Chain	Changes	Changes in firm's local size							
		-4	-3	-2	-1	0	1	2	3
Pyaterochka	200	0	0	1	20	1130	163	15	1
Magnit	193	0	0	1	23	997	152	15	2
Narodnaja sem'JA	173	4	8	24	109	667	28	0	0
SPAR	169	1	2	13	85	475	68	0	0
Perekrestok	67	0	0	0	2	486	63	2	0
Dixy	54	0	0	1	15	1031	36	2	0
O'KEY	23	0	0	0	19	222	4	0	0
Lenta	20	0	0	0	3	288	17	0	0
Karusel	9	0	0	0	9	89	0	0	0

decides to enter or leave the market, it usually changes the number of stores by 1 and no more.

5.2 Socioeconomic characteristics of local markets

Table 9 shows the socioeconomic characteristics of the local markets, which we can potentially use in further analysis. As one can see, these data are quite heterogeneous in terms of available periods. However, we believe that with enough imagination and the right imputation techniques, these data can be adapted to

Table 9: Descriptive statistics of local markets' socioeconomic characteristics

Variable	Periods	count	mean	std	min	max
pop	2017, 2018, 2019, 2020, 2021	1200	24250.9	19635.3	0	161094
flats	2017, 2018, 2019, 2020, 2021	1200	9754.2	7530.2	0	54308
devices	Jul 2017, Oct 2017, Apr 2019, Oct 2019	988	109.8	144.1	0	1354
price_sale	Jul 2018, Oct 2019, Dec 2019, Feb 2020, Oct 2020, Dec 2020	1245	108800.3	61261.8	9826.3	406376.6
price_rental	Jul 2018, Oct 2019, Dec 2019, Feb 2020, Oct 2020, Dec 2020	1313	1040.5	503.8	167.7	4372
price_rental_res	Jun 2018, Mar 2020	262	559.4	137.4	306.7	1154.3
income_real	Jun 2018, Mar 2020	262	118197.8	29747.5	62104.3	245588.7
income_model	2017	237	100206.5	25868.3	7293.9	166975
auto	Oct 2020	235	231946.4	226083.7	35.9	1183286
stops	Jul 2018, Feb 2020	466	17.1	12.9	1	69

match each other correctly, as well as to refer to openings and closings data, and so forth.[37]

Because of this temporal inconsistency in the data, it is difficult to make straightforward comparisons of the variations in the available indicators. In addition, since the data we obtained initially have different sources of origin, we also face a varying coverage of St. Petersburg and a different number of missing values. All this is reflected in the varying number of observations for available indicators.

The good news is that we have data on the two indicators most commonly used to measure the socioeconomic characteristics of local markets, namely population

[37]For more detailed descriptive statistics, including their dynamics, refer to our previous paper (Gaivoronskaia et al., 2021).

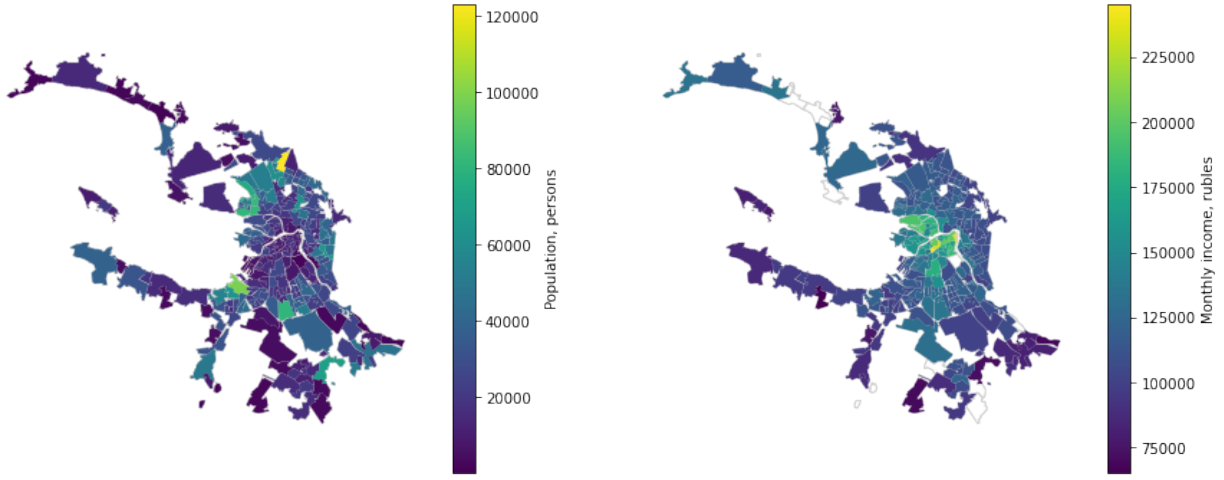

Figure 6: Population and income by postcode zones in St. Petersburg, 2018

and income. The map in Figure 6 shows the distribution of these indicators by postal code zones in St. Petersburg. The data presented are consistent with logic and common sense. The maps show that the city center is characterized by a smaller population and higher incomes, while the peripheries are more densely populated and typically described by lower incomes.[38]

Note, however, that the obvious advantage of the data presented, compared to previous studies, is a broader set of indicators beyond the standard characteristics used.

5.3 Sales data

Finally, we move on to the analysis of the sales data originally sourced from the OFD. Table 10 presents the descriptive statistics of the two main variables aggregated at the level of postal code zones for all product categories across store formats, while Figures 7, 8 present more detailed comparative characteristics across different product categories and store formats. In addition, the maps in Figures 9, 10 show the spatial distribution of the data available to us.

[38]The Table 9 also shows that local markets with zero population are present in the data. Note that there is only one such case in our data.

Recall that for reasons of commercial confidentiality, this data is available to us only in aggregate form. Roughly speaking, we need data on the sales of a particular category of products in at least three stores of a given format in a certain zone, in order to have the right to use the aggregated data without violating trade secrets (for a more detailed description, we refer everyone to section 4). Therefore, the amount of available data varies depending on the format. For example, hypermarkets are the largest format and they are relatively few in each local market, which leads to many situations with the inability to disclose data for this format. This explains the small number of observations for this format in Table 10 and Figure 7 and the wider confidence intervals in Figure 8. The same logic applies to supermarkets, which are more numerous than hypermarkets, but significantly fewer than discounters and convenience stores.

The resulting descriptive statistics are consistent with common sense and allow us to empirically compile profiles of grocery retail formats in St. Petersburg.

Once again, in terms of the number of stores, discounters and convenience stores have a huge advantage over other formats, while losing out on the size of the average check and the average number of checks per month, which is quite logical.

At the same time, hypermarkets have the highest average check values (Table 10). Since a hypermarket is a large store with a wide product range, the main model of consumer behaviour in this case seems to be a car trip to buy groceries for one or two weeks in advance. Figure 8 shows that this trend persists for almost all product categories with the exception of *cakes*, *instant food*, and *non-food*. Differences in check values between product categories also make sense. Regardless of the type of store, the highest values of the average check are shown in the *alcohol* category, and the lowest values are shown for *bakery products*. The only surprising thing is perhaps the low average check values in the *dietary and health food* category.

Table 10: Descriptive statistics for sales data by store type (Saint Petersburg, December 2019)

Store format	Variable	count	mean	std	min	max
Discounters, convenience stores	*avg_check*	245	928.9	337.0	289.0	2557.0
	avg_traffic	245	1678.7	617.4	131.3	4626.7
Hypermarkets	*avg_check*	37	1351.1	936.5	182.9	2799.6
	avg_traffic	37	70212.5	45905.0	319.5	182482.0
Supermarkets	*avg_check*	125	610.5	257.8	277.9	1551.2
	avg_traffic	125	42925.2	17640.1	10272.0	102602.0

Note. The unit of observation is the postcode zone. Thus, the number 37 in table means that there are 37 postcode zones where at least three hypermarkets are located.

The leading type of stores in terms of average number of checks is also hypermarkets with an average of 70996.1 checks per month for all product categories, compared to 42973.3 checks for supermarkets and 1679.1 checks for discounters (see Table 10). This difference between store types is also maintained by product category (see Figure 8), with the exception of *tobacco* and *cakes*. The leading category in terms of the average number of receipts is *bakery products*, while the lowest number of receipts is observed in the *cat and dog food* category, which is also quite logical.

The scatter plots in Figure 7 confirm these trends and also provide some information about clustering patterns in the performance of different store formats. The maps in Figures 9, 10 show that these trends are mostly confirmed in all local markets in St. Petersburg, with only a few exceptions.

6 Research applications

The first look at potential research opportunities for the grocery retailing industry in Russia was presented in our previous paper (Gaivoronskaia et al., 2021). Here we specify some possible directions of research, taking into account the specifics of the available data. Note that the selection of directions presented below is not complete or representative, but rather reflects the research interests of our team.

Figure 7: Avg. check value vs. avg. monthly number of checks scatter plot (by product category and store format)

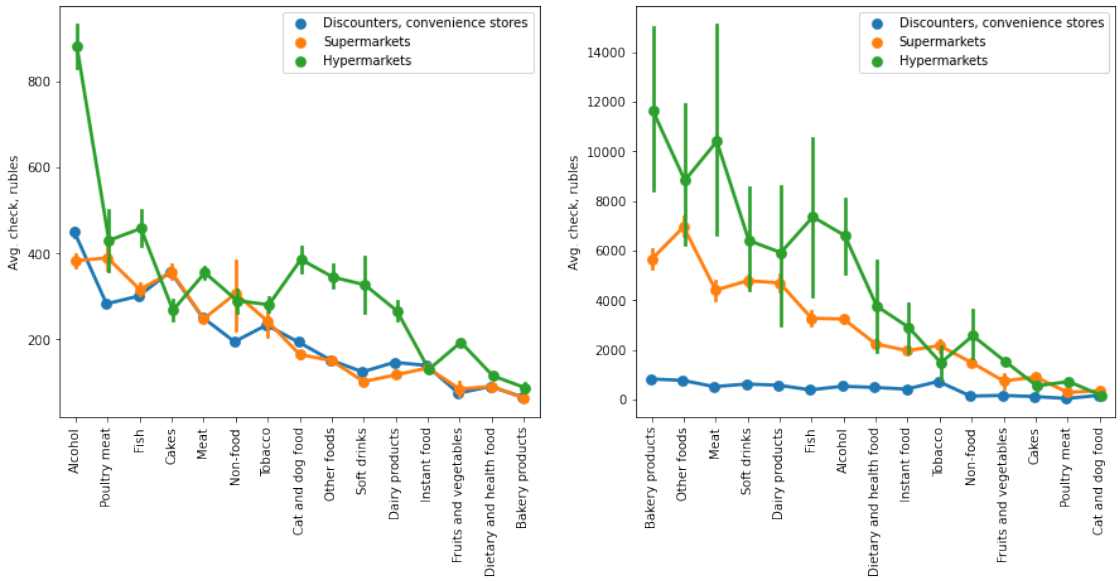

Figure 8: Avg. check value and avg. monthly number of checks (by product category and store format)

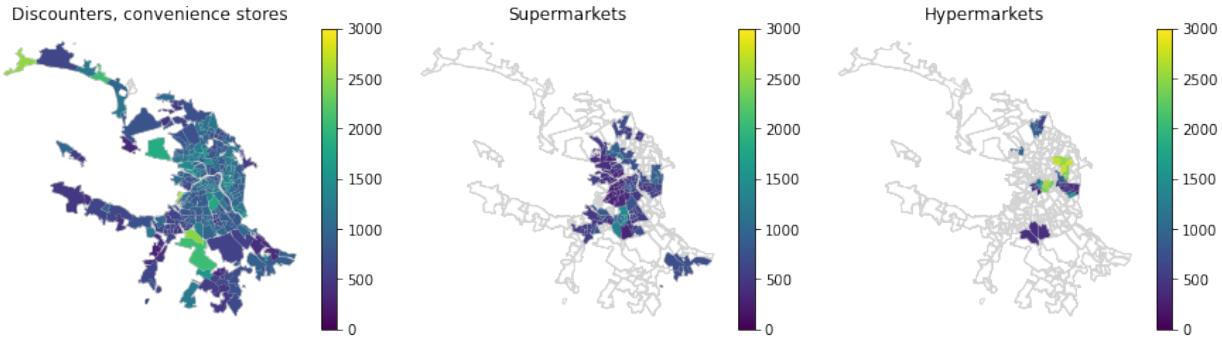

Figure 9: Avg. check value for all product categories (by postcode zone and store format)

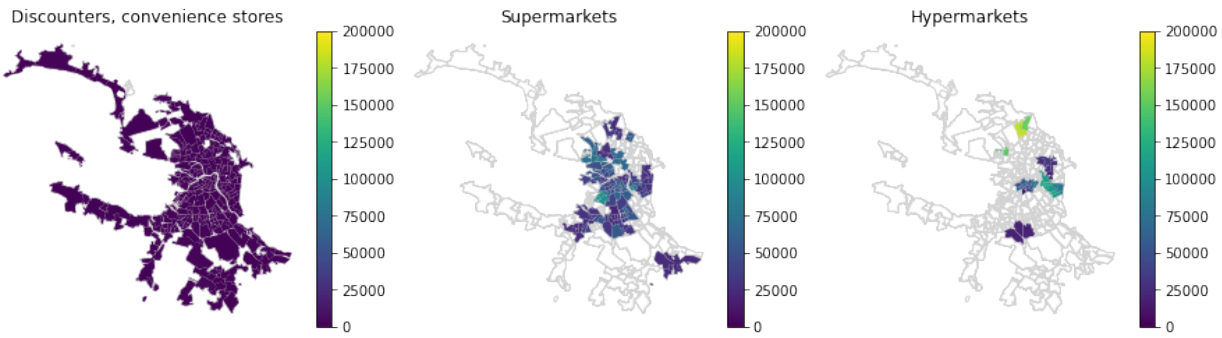

Figure 10: Avg. monthly number of checks for all product categories (by postcode zone and store format)

Thus, given previous research experience, our data can be used for research in several major areas.

As a starting point, we must recognize that there is a lack of even descriptive studies that provide a general overview of competition trends in Russian retailing. In this case, the data on St. Petersburg, although not covering the whole of Russia, is a rather interesting example due to the size of the city and the spatial differentiation of local markets there, as well as the highest level of concentration in the industry, which makes it a regular object of attention of the Federal Antimonopoly Service. Our data, which includes full statistics on market entries and exits by the major retailers in St. Petersburg, may allow us to close this gap.

Beyond that, our interests are mostly focused on building structural models of demand and competition in grocery retailing and their estimation using available data.

Starting with demand, we note that our data allow us to adapt the demand model introduced by Holmes (2011) and extended by Ellickson et al. (2020) for competition among several supermarkets using sales data. Since we are able to quantify the shares aggregated both on the level of product category and on the level of local market, our model can be definitely related to these studies. Note, however, that unlike from the latter paper, we do not have store-level sales data, which implies the need to make adjustments to the estimation procedure in our case. Still, we have available data at the level of 100-metre hexagonal cells, which can provide a sufficient level of detail.

Having estimated demand, we can use it in combination with static and dynamic models of competition between retailers. Decisions to enter or leave the market are always a tradeoff between demand and the intensity of competition. So, we can estimate the static model of spatial competition in the spirit of Seim (2006), Zhu and Singh (2009), Orhun (2013), among others. In addition, we can incorporate some dynamics in the model following Aguirregabiria and Vicentini

(2016) or Igami and Yang (2016). Using similar framework allows us to address such problems as cannibalization between stores of the same retail chain, as well as preemptive behavior of major retailers.

7 Concluding remarks

This paper presents a description of new data that provides previously unattainable opportunities for grocery retail industry research. The use of the data presented allows for the first serious empirical analysis of this market in Russia. More than that, our data is rich enough to make a significant contribution to the existing literature, regardless of the geographic context.

However, typically in economic research, we have to work with data collected from different sources, data that have been collected by other people and for other purposes. Thus, one of the goals of this paper was to provide sufficient justification for the validity of our data. We have done this by outlining all of the stages of data collection and processing in a systematic way.

Our data include three main blocks, including (i) store location data, (ii) socioeconomic characteristics of local markets, and (iii) sales data. Some dynamics are available for the first two blocks. Nevertheless, it is worth noting that there are some problems and limitations in the use of these data. Consumer and sales data are available only in an aggregated form, primarily at the postal code zone level. This is consistent with research conventions, but ideally one would like to have more disaggregated data. At the same time, store location data are available, although for several periods, but within a limited time frame, which does not allow the entire history of industry development to be monitored. These problems must be taken into account in the research.

References

AGUIRREGABIRIA, V. AND G. VICENTINI (2016): "Dynamic spatial competition between multi-store retailers," *The Journal of Industrial Economics*, 64, 710–754.

AVDASHEVA, S. AND A. SHASTITKO (2011): "Russian anti-trust policy: power of enforcement versus quality of rules," *Post-Communist Economies*, 23, 493–505.

AVDASHEVA, S. B., O. S. KHOMIK, AND M. N. KHRAMOVA (2015): "The impact of Russian retail chains business practices on suppliers performance before and after the new legislative regulation: microdata assessment," *Russian Management Journal*, 13, 3–38.

EINAV, L., E. LEIBTAG, AND A. NEVO (2010): "Recording discrepancies in Nielsen Homescan data: Are they present and do they matter?" *Quantitative Marketing and Economics*, 8, 207.

ELLICKSON, P. B., P. L. GRIECO, AND O. KHVASTUNOV (2020): "Measuring competition in spatial retail," *The RAND Journal of Economics*, 51, 189–232.

GAIVORONSKAIA, E., F. ISKHAKOV, M. KARMELIUK, S. KOKOVIN, E. OZHEGOV, A. OZHEGOVA, AND D. TERESHCHENKO (2021): "Competition among Russian grocery stores: facts and hypotheses to explore," *Available at SSRN 3957743*.

HOLMES, T. J. (2011): "The diffusion of Wal-Mart and economies of density," *Econometrica*, 79, 253–302.

IGAMI, M. AND N. YANG (2016): "Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in Canada," *Quantitative Economics*, 7, 483–521.

ORHUN, A. Y. (2013): "Spatial differentiation in the supermarket industry: The role of common information," *Quantitative Marketing and Economics*, 11, 3–37.

PENNERSTORFER, D. AND B. YONTCHEVA (2021): "Local market definition in competition analysis: An application to entry models," *Economics Letters*, 198, 109678.

RADAEV, V. (2006): "The evolution of organizational forms in Russian retailing," *Voprosy Ekonomiki*, 41–62.

——— (2018): "A rise of state activism in a competitive industry: The case of Russian retail trade law of 2009," *Communist and Post-Communist Studies*, 51, 27–37.

ROBINSON, T. (1998): "The role of retailing in a Russian consumer society," *European Business Review*, 98, 276–281.

SEIM, K. (2006): "An empirical model of firm entry with endogenous product-type choices," *The RAND Journal of Economics*, 37, 619–640.

YANG, N. (2020): "Learning in retail entry," *International Journal of Research in Marketing*, 37, 336–355.

ZHU, T. AND V. SINGH (2009): "Spatial competition with endogenous location choices: An application to discount retailing," *Quantitative Marketing and Economics*, 7, 1–35.