Pontifícia Universidade Católica
do Rio de Janeiro

Johann Marques Viana Freitas

High Frequency Online Prices

Dissertação de Mestrado

Master's dissertation presented to the Programa de Pós–graduação em Economia, do Departamento de Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia.

Advisor: Prof. Dr. Lucas Lima

Rio de Janeiro
March 2025

Johann Marques Viana Freitas

High Frequency Online Prices

Master's dissertation presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia. Approved by the Examination Committee:

Prof. Dr. Lucas Lima
Advisor
Departamento de Economia – PUC-Rio

Prof. Dr. Yvan Becard
Departamento de Economia – PUC-Rio

Prof. Dr. Marcelo Sant'Anna
FGV EPGE

Rio de Janeiro, March 21st, 2025

Johann Marques Viana Freitas

Economist, Petróleo Brasileiro S.A. (Petrobrás).
B.A., Economics, Universidade Federal Fluminense (UFF), 2022.

Bibliographic data

Freitas, Johann

High Frequency Online Prices / Johann Marques Viana Freitas; advisor: Lucas Lima. – 2025.

75 f: il. color. ; 30 cm

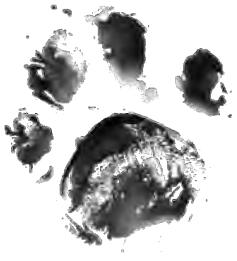
Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2025.

Inclui bibliografia

1. Economia – Teses. 2. Raspagem de Preços. 3. Bairro. 4. Aprendizado Profundo. 5. Classificação de Produtos. I. Lima, Lucas. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 004

For Pretinho (in memorian)



Acknowledgments

Johanna Maria, Isabela e Vanessa had fundamental role in this Project, by acting on data collection and product classification tasks. To your immeasurable contributions, my most sincere and profound thanks.

To all friends, loved ones, and family members, who provided me with the emotional and operational foundations to fulfill this Program.

To Professor Lucas Lima, who believed in this Project.

To the Colleagues and Managers of Petrobras who encouraged the fulfillment of this Project.

Ao Pretinho.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

Abstract

Freitas, Johann; Lima, Lucas (Advisor). **High Frequency Online Prices**. Rio de Janeiro, 2025. 75p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

This research explores the use of online prices from the iFood platform as a proxy for socioeconomic indicators at the neighborhood level and as a data source for daily frequency inflation measures. By collecting pricing data for a açaí sorbet across multiple neighborhoods in Rio de Janeiro, a single-good price index is constructed. The relationship between iFood prices and the Social Progress Index (SPI) is examined, and it is found that there is a positive correlation between the two variables. Once this correlation pattern is established, a theoretical economic framework describing the underlying relationship is proposed. Results suggest that iFood prices can provide valuable information for policy formulation and decision-making at the local level, particularly in cities where traditional indicators are not available or outdated. The research also highlights the importance of considering cultural factors and local tastes when using online prices as a proxy for socioeconomic indicators. On inflation side, the monthly measure fits well official inflation data at national level, suggesting that iFood prices are indeed meaningful for estimating inflation and daily observations seems a valid indicator for within-month inflation dynamics.

Keywords

Scraped Prices; Neighborhood Level Data; Deep Learning; Product Classification.

Resumo

Freitas, Johann; Lima, Lucas. **Preços Online em Alta Frequência**. Rio de Janeiro, 2025. 75p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Este estudo explora o uso de preços do iFood como proxy para indicadores socioeconômicos a nível de bairro e como fonte de dados para uma medida de inflação em frequência diária. Examina-se o relacionamento entre preços e o Índice de Progresso Social (SPI), e verifica-se correlação positiva entre as duas variáveis. Propõe-se um arcabouço teórico para descrever este relacionamento. Este resultado sugere que preços online são uma alternativa factível para elaboração de políticas públicas e tomada de decisão a nível local, particularmente em municípios onde indicadores tradicionais estão indisponíveis ou desatualizados. Pelo lado da inflação, verificamos que a média das estimativas diárias em dado mês acompanha a inflação mensal oficial a nível nacional, o que sugere que preços do iFood constituem meio válido para estimativas de inflação, e observações diárias são um indicador válido para a dinâmica inflacionária ao longo do mês.

Palavras-chave

Raspagem de Preços; Bairro; Aprendizado Profundo; Classificação de Produtos.

Table of contents

List of figures

List of tables

List of algorithms

List of codes

List of Abreviations

ADI – Análise Digital de Imagens

BIF – *Banded Iron Formation*

1
Introduction

The advent of Big Data has significantly impacted the field of economics, offering both opportunities and challenges. Authors (EINAV; LEVIN, 2014b; EINAV; LEVIN, 2014a) have explored how the Data Revolution can transform economic policy by enabling the capture and processing of real-time data, enhancing government operations, and improving economic policymaking efficiency. The availability of extensive micro-level administrative and private firm data has been revolutionizing economic research (ABRAHAM et al., 2022). This paper proposes employing scrapping techniques to collect pricing data from iFood, a Brazilian digital ordering and delivery platform.

Researchers have used online pricing data for constructing price indexes (see section 1.1). iFood, tough, offers geographically detailed data, which allows for analysing pricing patterns for arbitrarily small territorial units, such as neighborhoods. One may use these prices as a proxy for economic indicators at neighborhood level, say, for policy design, in the context of unavailable official data. Such proxy may also be usefull in the context of outdated official data: for instance, Census data in Brazil provides income per capita information at neighborhood level. Nevertheless, these are usually calculated on a 10 year basis. This study's first is whether iFood pricing data ability to proxy socioeconomic indicators. We provide evidence that online data is indeed useful for this purpose.

Besides geographic detail, iFood data also emerges as an input for price index. In line with the existing literature on online price indexes, iFood allows for collecting data on daily basis. However, the platform large market share ensures data representativeness of Brazilian delivery market, which previous literature on online price indexes points out as a drawback for this approach. Analysing iFood daily inflation data is preceeded by the second research

question: whether iFood prices fit official inflation data. Once this connection is stablished, we explore iFood daily inflation as a preliminary indicator for monthly inflation, intra-month inflation dynamics and also price stickiness stylized facts.

Specifically, this project evaluates two attributes of granularity provided by iFood data: spatial and temporal granularity. The spatial objective is to explore the iFood prices database and evaluate its effectiveness in capturing neighborhood-level economic indicators. Once the statistical relationship between iFood prices and these indicators is established, an econometric analysis further explores this connection. This connection allows using iFood prices as a feasible proxy for economic indicators when these are not available or outdated for a given location. This examination also determines whether up-to-date iFood data offers additional value compared to outdated official data through a vanilla application. Such data could significantly benefit metropolitan areas beyond capital cities by providing essential insights for policy formulation and decision making.

Local-level analysis should also take into account how units interact with each other. One can model these interactions into network setups. Networks have been used to estimate spillover and peer effects (BRAUN; VERDIER, 2023; AUERBACH, 2022; JOHNSSON; MOON, 2021; BRAMOULLé; DJEBBARI; FORTIN, 2020).

For temporal granularity, this paper provides daily inflation estimates and compare these against monthly benchmarks to assess intra-month inflation dynamics, particularly during a significant local event. Specifically, an online price index is computed at states' capitals. An intriguing finding is that the online price index is far more volatile than a traditional benchmark. This paper also shed light on candidate drivers of this additional volatility.

1.1 Literature Review

For nowcasting, timely data can bypass delays in macroeconomic indicator publication. New data sources, including social media, news, internet searches, and the Internet of Things (IoT), have become valuable. The granularity and high frequency of scrapped data, combined with lower collection costs and resilience against government interventions, make them particularly useful in constructing price indexes. Machine learning algorithms further enhance their utility in tasks like product classification for price index calculation (BALDACCI et al., 2016; OANCEA, 2023; MACHADO, 2023).

Among these new data sources, tech enterprises and digital platforms play a major role. The popularization of online delivery and grocery shopping platforms in Brazil, exemplified by iFood's dominance during the Covid-19 pandemic, demonstrates the potential of these platforms. iFood's extensive reach provides geographically detailed but unstructured pricing data, which can be instrumental in economic analysis and policymaking (ANGRIST; GOLDBERG; JOLLIFFE, 2021).

Digital platforms like iFood offer alternative data sources for various fields, including epidemiology and neighborhood-level socioeconomic status (SES) measurements. These data can influence policy regarding environmental hazards (XIE; HUBBARD; HIMES, 2020), crime (STACY; HO; PENDALL, 2017), and school performance (RUIZ; MCMAHON; JASON, 2018). Real-time SES measures can guide local policies more effectively than traditional surveys. For instance, data from business review platforms have been used to measure local economic activity, serving as indicators of neighborhood gentrification and housing price changes (GLAESER; KIM; LUCA, 2018).

Over past decade, online prices have been used for inflation measuring, among other Macroeconomics topics. The *Billion Prices Project* (BPP) (CAVALLO; RIGOBON, 2016) offered daily price data for goods sold by online supermarkets across multiple countries. Its applications spanned a wide range

of studies (CAVALLO; RIGOBON, 2011; CAVALLO; CAVALLO; RIGOBON, 2014; CAVALLO, 2018; CAVALLO et al., 2018).

A natural concern regarding online pricing data is how it is related with its physical counterpart. Literature evidence indicate that price levels in websites are comparable to physical stores and do represent retail prices (CAVALLO, 2017), despite heterogeneities and the fact that most of transactions still occurring offline at that time. Also, measuring inflation through prices collected from online retailers aligns with official inflation estimates in Latin American countries (CAVALLO, 2013). A discrepancy is noted in Argentina, which can be attributed to government interventions aiming to mask inflation data. Furthermore, the alignment between online prices and official inflation improves as data is sourced from supermarkets with larger market shares and cities that are more representative of the country as a whole.

These studies highlight the key advantages of using scrapped price data. Scrapping techniques provide access to large-scale real-time information, in contrast to traditional methods that often have significant time gaps between data collection and publication. This enables analysis of short-term dynamics, seasonal patterns, and other phenomena on higher frequencies. Additionally, scrapping data comes with lower costs and offers greater granularity and flexibility. Importantly, this data is independent of official statistical offices, which may be susceptible to government intervention.

However, classifying products using large-scale data poses a significant challenge. Furthermore, online data may be scattered across various sources, such as different supermarkets that contribute to the BPP database. This dispersion adds complexity to the task of adjusting routines to process and normalize the data, as well as raises concerns about representativeness, as highlighted by Cavallo (2013). Note that, for this paper case, iFood platform serves as a centralized online environment that brings together multiple vendors. This enables the use of a common scraper to gather data from various

vendors.

Also, while literature raised concerns regarding vendor's representativeness in terms of market share, iFood has a dominant position in the Brazilian delivery market, thus making the data obtained from this platform highly representative. Moreover, just like the existing literature on digital platforms and local-level economic indicators, this paper leverages flexibility in terms of spatial and product coverage. iFood allows for searching for arbitrary keywords associated with specific food or beverage items, based on given latitude and longitude coordinates.

2
Exploratory Analysis

In recent years, the use of delivery services in Brazil has grown considerably (PIGATTO et al., 2017). This expansion can be attributed not only to the increasing availability and popularity of digital platforms but also to the influence of the Covid-19 pandemic and social distancing measures. Although the adoption of delivery services was already on the rise before 2020[1], the pandemic further accelerated this trend as it prompted restaurants and consumers to embrace this mode of service (BOTELHO; CARDOSO; CANELLA, 2020).

The significant growth in this sector can largely be attributed to the widespread adoption of iFood. iFood is a Brazilian technology company that operates an online platform for food ordering and delivery services. Established in 2011, the company has received substantial investments during the latter half of the 2010s. In 2022, iFood was recognized as the most valuable startup in Brazil and the second most valuable in Latin America[2].

iFood boasts an user base of 40 million users, facilitating over 60 million orders per month. The platform operates with a network of 200,000 couriers and serves approximately 300,000 merchants. Its coverage extends to more than 1,200 municipalities across Brazil[3]. According to market reports, iFood held an 83% share of the Brazilian delivery market by 2021. This market dominance has likely been further solidified following the exit of Uber Eats, one of iFood's main competitors, from the Brazilian market in 2022[4].

The platform experienced a relevant growth during the pandemic period, reaching 39 million orders per month[5]. Estimates suggest that in 2022, iFood contributed to approximately 0.53% of Brazil's GDP (HADDAD et al., 2023)[6].

[1] See iFood (2024).
[2] See Bloomberg Línea Brasil (2022).
[3] See iFood para parceiros (2024).
[4] See Measurable AI (2021) and TecMundo (2023).
[5] More information can be found here.
[6] Accounting for both direct and indirect effects

As mentioned above, iFood offers a wealth of diverse and representative data on various goods. We employ web scraping techniques to collect iFood pricing data and calculate a price index at the neighborhood level for Rio de Janeiro. Then, we delve into exploring the relationship between this price index and the Social Progress Index (SPI)[7]. Rio de Janeiro is the second largest Brazilian city in terms of population, with more than 6 million inhabitants. The city's weight in the National Consumer Price Index is 9.43%.

The argument for using SPI is the fact that this indicator stands out as the most up-to-date and comprehensive socioeconomic measure available at neighborhood level in Rio de Janeiro municipality (PULICI et al., 2022; STERN et al., 2014; GREEN et al., 2024). For the purpose of this study, one can interpret the SPI as essentially playing the same role as income per capita, for example.

Generally, the SPI exhibits a strong correlation with income measures, such as GDP.[8]. However, it is important to note that this value may not be entirely reliable since analysis relies on data from the 2010 Census, which is considerably outdated compared to most of the components used in calculating the SPI.

Therefore, establishing a statistical relationship between a price index constructed using iFood data and the SPI would indicate that these online prices serve as a practical proxy for economic development, well-being, and potentially income indicators at a detailed level in various cities. This is particularly valuable in situations where data on indicators like the SPI or GDP is either unavailable or significantly delayed. Furthermore, these indexes possess the advantageous properties of scalability, adaptability, and real-time availability.

[7]It is calculated by averaging three subindexes: Basic Human Needs, Well-being Fundamentals, and Opportunities. More detailed information about the components and sources of these subindexes can be found on the [official website].

[8]Pulici et al. (2022) estimate a 55% Pearson coefficient.

2.1 Web scraping procedure

The platform operates in the following manner: a seller, such as a restaurant, supermarket, or pet shop, registers her business on the platform. She has the option to choose between two available plans: the Basic plan, where each store handles its own couriers to serve customers, or the Delivery plan, which allows the store to utilize iFood's network of courier partners. These partners use their own vehicles, predominantly bikes and motorcycles, to deliver orders.

Once a store is registered, nearby users will be able to discover the store and view its menu when searching for specific products or categories. Users can place an order through the app and typically make the payment within the app as well. iFood charges the merchant a commission and a payment fee based on the transaction value.

After receiving the order, the store begins to prepare it. In cases where the seller has chosen the Delivery plan, a courier located nearby the store is notified to pick up the order. The courier collects the order from the store and delivers it to the user's designated location.

The web scraping software utilized in this study emulates a search on the iFood platform. To execute the search, the software requires a query specifying the item to be searched, as well as latitude and longitude coordinates to simulate the user's location. The software collects all available items from the search response, along with their prices, and retrieves additional merchant information such as delivery fees and distances from the provided coordinates. The scraping process is executed sequentially for each centroid, and the results are combined to provide an overview of the entire city.

Figure 2.1 below illustrates an example of an iFood response as seen on the platform's user interface. The corresponding response from the perspective of the web scraper is presented in Code 1.

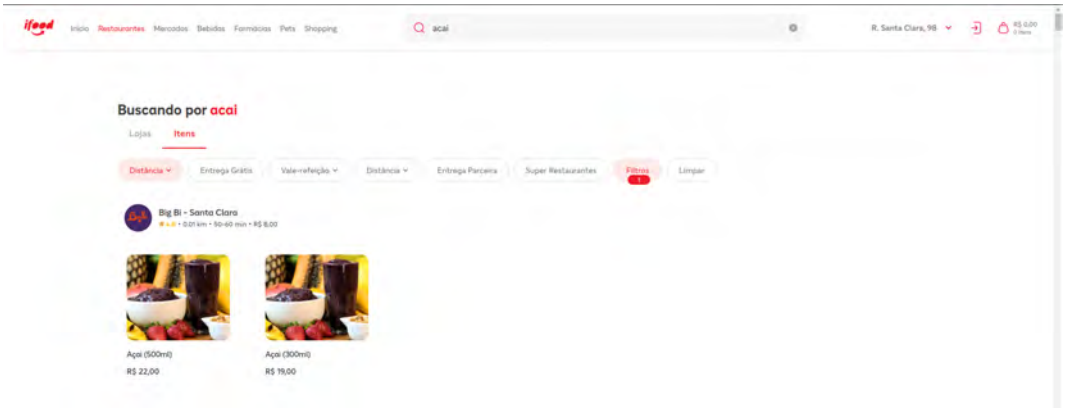


Figure 2.1: Example of iFood response - User interface


```
{'action': 'merchant?identifier=6b962527-6e4d-49d0-8f69
    -0519aec47ae5&name=Big%20Bi%20-%20Santa%20Clara&slug=
    rio-de-janeiro-rj/big-bi---santa-clara-copacabana', '
    available': True, 'currency': 'BRL', 'deliveryInfo':
    {'fee': 800, 'timeMaxMinutes': 60, 'timeMinMinutes':
    50, 'type': 'DELIVERY', 'deliveryMode': 'DEFAULT', '
    formattedDeliveryFee': '[#666666: R$ 8,00]'}, '
    distance': 0.05, 'id': '6b962527-6e4d-49d0-8f69-0519
    aec47ae5', 'isFavorite': False, 'isNew': False, '
    isSuperRestaurant': False, 'items': [{'action': '
    catalog-item?itemUuid=d0a2e64b-494e-4786-aad4-82
    c11a7121b6&merchantUuid=6b962527-6e4d-49d0-8f69-0519
    aec47ae5', 'available': False, 'currency': '', 'id':
    'd0a2e64b-494e-4786-aad4-82c11a7121b6', 'imageUrl':
    ':resolution/pratos/a2a7e38b-b644-40eb-8d8d-83257
    c7b0004/201704241700_26037439.jpg', 'name': 'Açai
    (500ml)', 'pricing': {'type': 'REGULAR', 'unitPrice':
     2200}}, {'action': 'catalog-item?itemUuid=3b5049bb-6
    ece-4123-9008-6de37b10084c&merchantUuid=6b962527-6e4d
    -49d0-8f69-0519aec47ae5', 'available': False, '
    currency': '', 'id': '3b5049bb-6ece-4123-9008-6
    de37b10084c', 'imageUrl': ':resolution/pratos/
```

```
a2a7e38b-b644-40eb-8d8d-83257c7b0004/201704241700
_26037439.jpg', 'name': 'Açai   (300ml)', 'pricing':
{'type': 'REGULAR', 'unitPrice': 1900}}], 'logoUrl':
':resolution/logosgde/logo%20big_BIGBI_CLARA.jpg', '
name': 'Big Bi - Santa Clara', 'formattedName':
'[#141414: Big Bi - Santa Clara]', 'userRating':
4.800000190734863}
```

Code 1: Ifood scrapper response

2.2 Data description

Using Rio de Janeiro's neighborhoods boundaries data (Data Rio, 2024), I computed a list of centroids that will be used to represent a user sending a query in each neighborhood.[9]

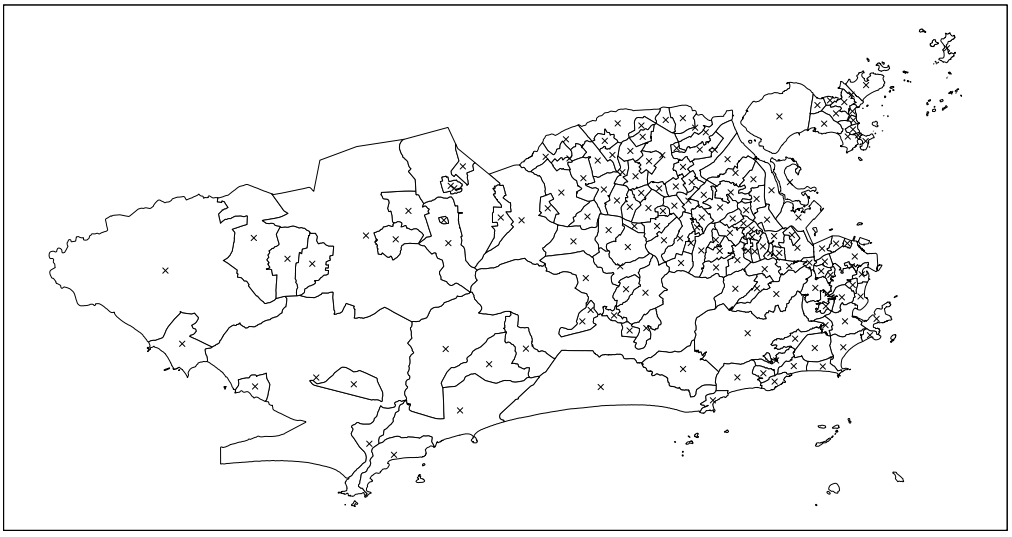


Figure 2.2: Neighborhood boundaries and centroids positioning

The chosen query was *acai*, expecting to connect with açai berry sorbet offers. Its consumption is highly connected with local culture, thus providing a robust to neighborhood-level heterogeneities, as it is reasonable to state that

[9]An important observation is that some neighborhoods encompass uninhabited small islands, which affect the centroids computation. To circumvent this issue, the centroid is obbtaineed on the larger polygon associated with each neighborhood, as seen in Figure 2.2.

social profiles may affect *how* this good is consumed, for example, its quality, but not neighborhood average demand, that is homogeneous.

Collecting the entirety of the data brings impurities, manifesting as undesired items. Examples of undesired items found during the research include açai berry shampoo, announced in iFood by supermarkets, and *combos* containing açai berry sorbet and other goods, like burgers. This issue is addressed by data cleansing. I compute a frequency table on the words found in items' labels. Then, I manually look for words that are possibly associated with items other than açai berry sorbet, and then delete entries containing this words. Additional filters are imposed to control bowl sizes, kept at 400ml.

Ceteris paribus, individuals tend to prefer merchants closer to their location. Moreover, there are high interdependencies across neighborhoods: a single merchant may be reached by users in different neighborhoods. Therefore, in order to ensure the indexes indeed reflect the prices practiced in a particular neighborhood, I normalize (Equation 2-1) the prices of each firm j (P_j) by the square root of the distance between the user's i location (neighborhood centroid) and the merchant. This normalization places more weight on merchants located closer to the users (centroids). The option for the square root was data-driven.

$$p_{ij} = \frac{P_j}{\sqrt{Distance_{ij}}}. \tag{2-1}$$

I used an intuitive aggregation for computing an index for each neighborhood. The index $\bar{p}_i^{\text{geom}}$ defined in Equation 2-2 is simply the geometric mean, which provides a natural interpretation for its logarithm - the average of $\{\log(p_{ij})\}_{j=1}^{\#K_i}$, where K_i denotes the set of firms connected to i and $\#K_i$ is its cardinal.

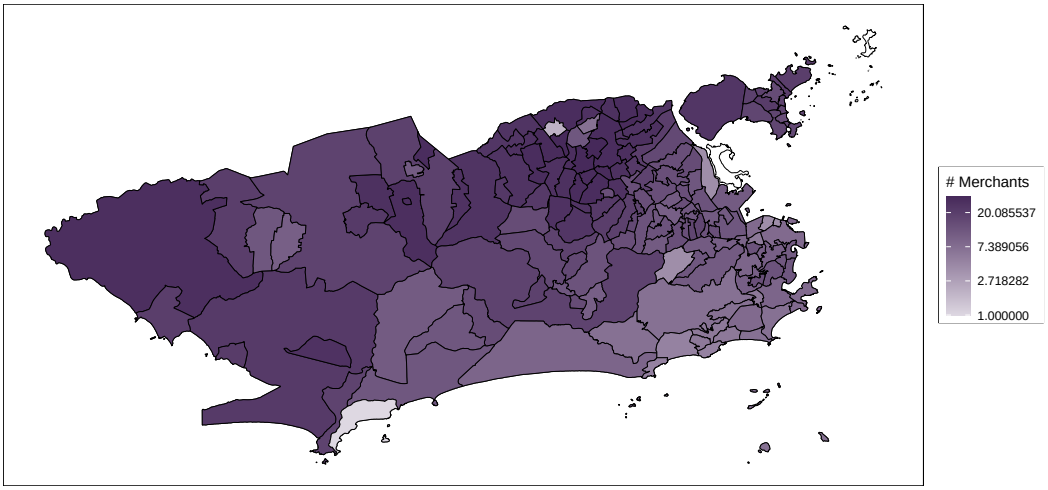


Figure 2.3: Number of different merchants connected in each neighborhood

$$\bar{p}_i^{\text{geom}} = \sqrt[\#K_i]{\prod_{j=1}^{\#K_i} \frac{P_j}{\sqrt{Distance_{ij}}}} = \sqrt[\#K_i]{\prod_{j=1}^{\#K_i} p_{ij}}. \tag{2-2}$$

I used data collected on a Friday[10]. The scraping procedure started at 3 PM. Figure 2.3 shows the number of achieved merchants in each neighborhood.

The missing neighborhoods for SPI are Paquetá Island, Lapa, Gericinó, Jabour, Vila Kennedy and Guaratiba Island, while for $\bar{p}_i^{\text{geom}}$ and $\log(\bar{p}_i^{\text{geom}})$ the missing observations are Paquetá Island and Cidade Universitária. Histograms in Figure 2.4 suggests that the number of different merchants connected in each neighborhood has a bimodal distribution.

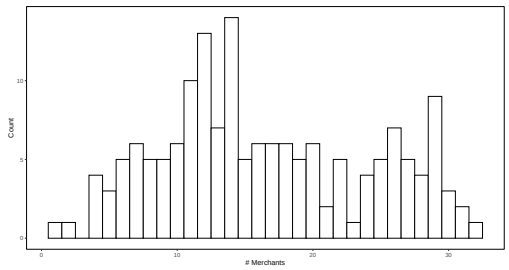
Figure 2.5 displays the spatial distribution for both SPI and $\bar{p}_i^{\text{geom}}$. It illustrates the presence of spatial autocorrelation in data. Barra da Tijuca presents the highest SPI (79.29) in the sample, while Cidade Nova has the worst index (50.43). Ipanema has the greatest $\bar{p}_i^{\text{geom}}$ (36.84), and Grumari has the lowest (4.89).

Given the spatial autocorrelation structure, it is convenient to nest the observations into larger territorial units. The Rio de Janeiro Municipality is

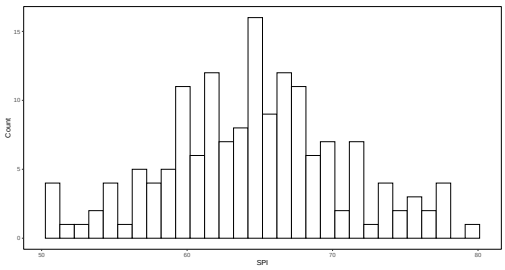
[10] March 1st, 2024.

Table 2.1: Descriptive statistics

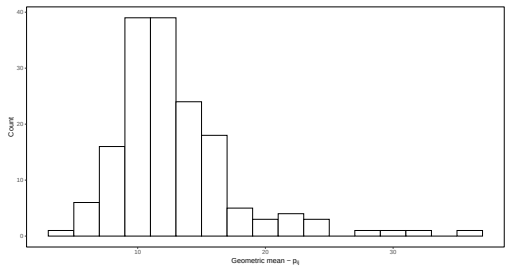
		Connected		Pric	
		Neighborhoods	Merchants	Avg.	St
Centro and Região Portuária	Neighborhood (Avg.)	-	9.33	23.18	
	Merchant (Avg.)	3.59	-	22.43	
	AP	15	39	23.44	
Zona Sul and Grande Tijuca	Neighborhood (Avg.)	-	8.56	24.21	
	Merchant (Avg.)	3.24	-	22.71	
	AP	25	66	23.78	
Zona Norte - Méier and Surroundings	Neighborhood (Avg.)	-	19.39	18.76	
	Merchant (Avg.)	7.74	-	17.16	
	AP	79	198	18.90	
Zona Oeste - Barra and Jacarepaguá	Neighborhood (Avg.)	-	13.63	21.27	
	Merchant (Avg.)	2.91	-	20.35	
	AP	19	89	21.62	
Zona Oeste - Bangu and Santa Cruz	Neighborhood (Avg.)	-	21.88	18.34	
	Merchant (Avg.)	3.02	-	17.09	
	AP	24	174	18.34	
Whole City	Neighborhood (Avg.)	-	16.48	20.24	
	Merchant (Avg.)	6.15	-	18.34	
	AP	162	434	19.81	



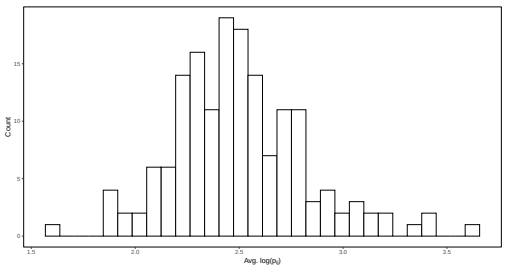
(a) Number of different merchants connected in each neighborhood



(b) SPI

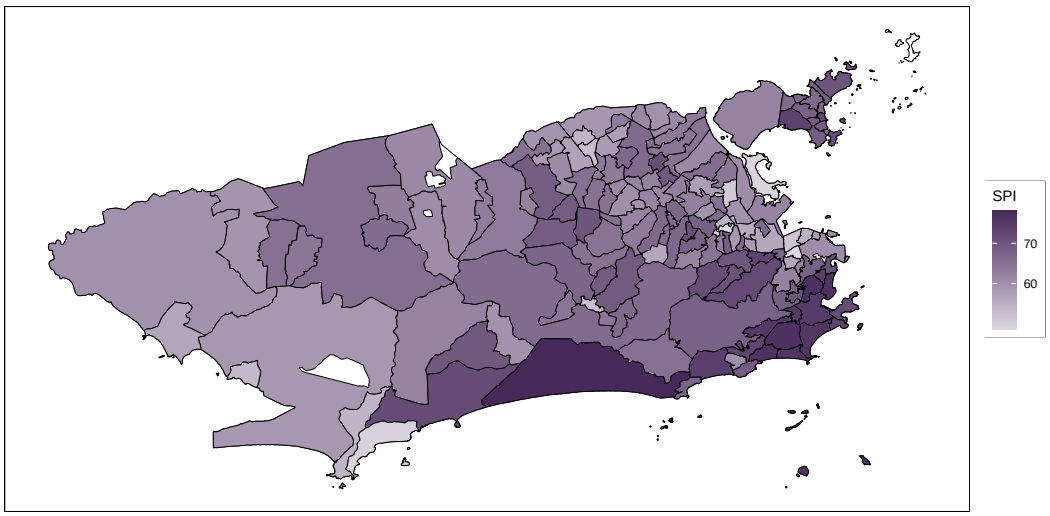


(c) $\bar{p}_i^{\text{geom}}$

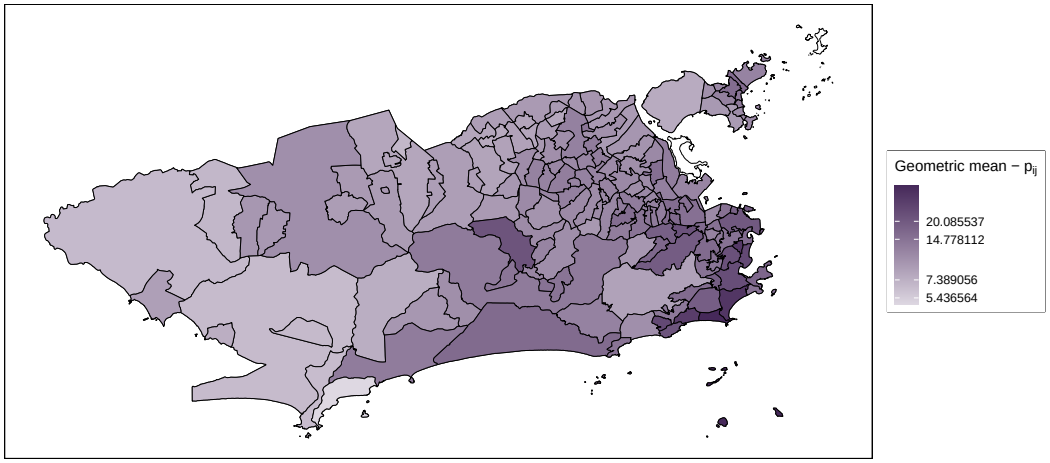


(d) $\log\left(\bar{p}_i^{\text{geom}}\right)$

Figure 2.4: SPI and $\bar{p}_i^{\text{geom}}$ distributions



(a) SPI distribution by neighborhoods

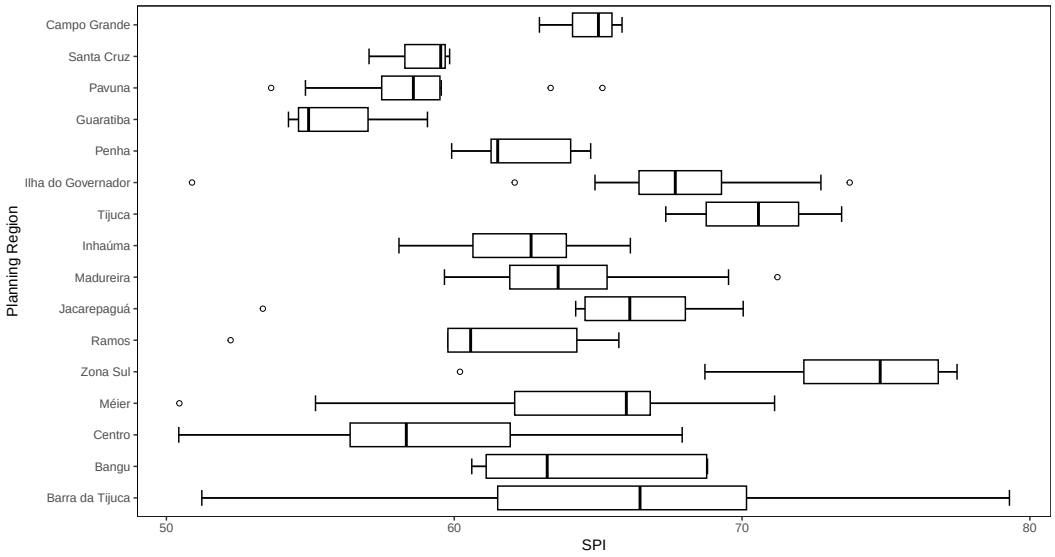


(b) $\bar{p}_i^{\text{geom}}$ distribution by neighborhoods

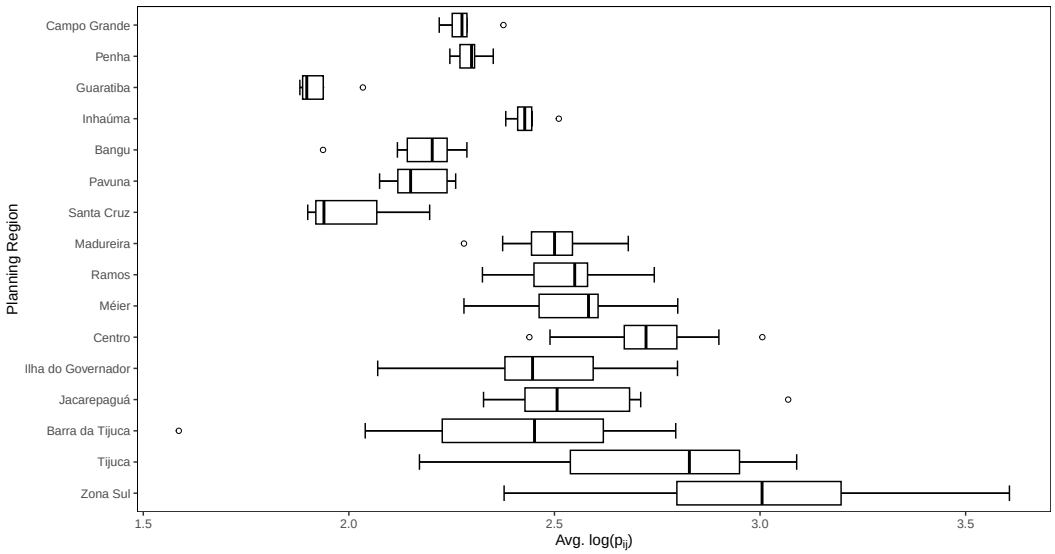
Figure 2.5: SPI and $\bar{p}_i^{\text{geom}}$ distributions

divided into Planning Areas (AP); Planning Regions (RP); Administrative Regions (RA); Neighborhoods; Watersheds and Sub-watersheds; and Air Basins and Micro-basins (Prefeitura da Cidade do Rio de Janeiro, 2024). There are 16 RPs, representing subdivisions of APs and groups of RAs, following specific homogeneity criteria. Analysing the distributions within RPs through the boxplots in Figure 2.6 unveils that four RPs - Centro, Barra da Tijuca, Méier and Bangu - displays particularly high variability for SPI. Barra da Tijuca (RP), for example, encompasses the highest SPI (Barra da Tijuca) and the

third lowest SPI (Grumari, 51.23). The highest median SPI and the highest median $\log(\bar{p}_i^{\text{geom}})$ coincides (Zona Sul), and the same happens for the lowest (Guaratiba).



(a) SPI



(b) $\bar{p}_i^{\text{geom}}$

Figure 2.6: Boxplots by planning region

These RPs likely reflects specific phenomena that the online prices tool is not able to account for. Therefore, for evaluating iFood prices ability to proxy the SPI purposes, these observations can be discarded, in a step analogue to outliers pruning.

2.3
iFood prices and SPI

As seen in Figure 2.6(a), Centro, Barra da Tijuca, Méier and Bangu have specially volatile SPI. I consider analysing the relationship between $\bar{p}_i^{\text{geom}}$ and SPI both with and without these RPs. Figure 2.7 displays the scatterplots of SPI against $\log(\bar{p}_i^{\text{geom}})$, as well as OLS results.

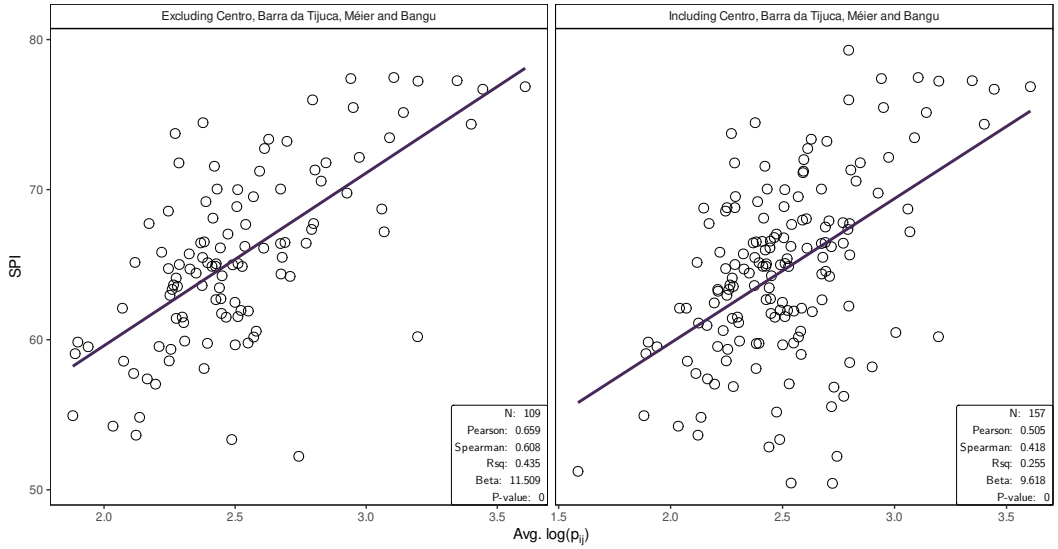


Figure 2.7: Avg. $\log(p_{ij}) \times$ SPI

There is a clear positive correlation between the SPI and $\log(\bar{p}_i^{\text{geom}})$. Most of the observations that do not follow this trend belongs to the four excessively volatile RPs. Specially, there are clusters characterized by excessively low SPI given prices. Indeed, when these observations are not excluded, correlation measures become weaker and the explanatory power decreases by about 41%.

As argued before, there are specificities that are not captured by the prices, and these specificities are reflected in high in-RP variation - recall Figure 2.6(a). Some possible explanations are the presence of lower SPI neighborhoods in each RP, or even smaller locations in some neighborhoods - like shantytowns. As prices have high spatial autocorrelation, these locations interact mutually: a shantytown may drive SPI and - hypothetically - prices disproportionately. By the other hand, a user located in a lower SPI neigh-

borhood connects to higher SPI and hypothetical higher prices close location, which in turn increases her location's average price.

2.4 Other cities

As an attempt to establish an external validation argument, I performed the same exercise, using the same good, for evaluating the correlation between the price index and a socioeconomic indicator for São Paulo. São Paulo is the largest Brazilian city in terms of population, with more than 11 million inhabitants according to 2022 Census. The city's weight in the National Consumer Index is 32.28%.The chosen indicator was the Income per capita, calculated using 2010 Census data. One can argue that using this indicator is appropriate given the fact that it is usually highly correlated with the SPI, specifically through Opportunities dimension, which is the most correlated with prices. It was calculated within districts, which is usually above neighborhoods. There are 91 subprefeituras in São Paulo's sample from 96 existing. The five missing subprefeituras are Jardim Paulista, Liberdade, Marsilac, Vila Mariana and República, for which the scrapped returned no data on açaí.

Figure 2.8 shows a positive correlation between Income and the price index. However, correlation measures are substantially weaker than in Rio de Janeiro. There are differences between experiments performed in Rio de Janeiro and in São Paulo, notably the chosen indicator and the territorial unit. Another possibility is that cultural factors may drive these differences. We test this cultural factors by scraping data for a different item in São Paulo, under the hypothesis that it is a better representative of local taste. Scraping data for pizza in São Paulo returns a clearer correlation between prices and Income (Figure 2.9).

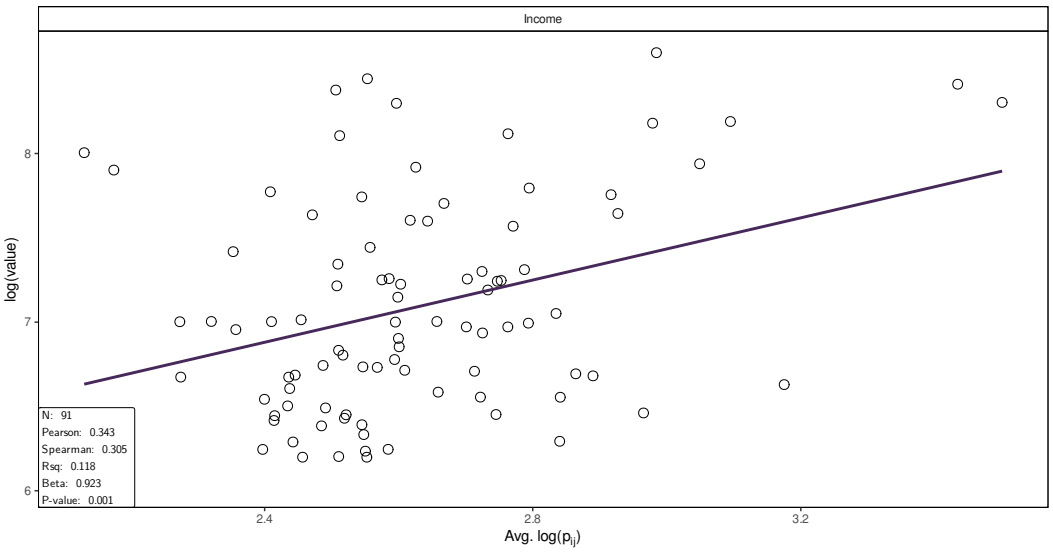


Figure 2.8: Avg. log(p_{ij}) $\times$ Income - São Paulo

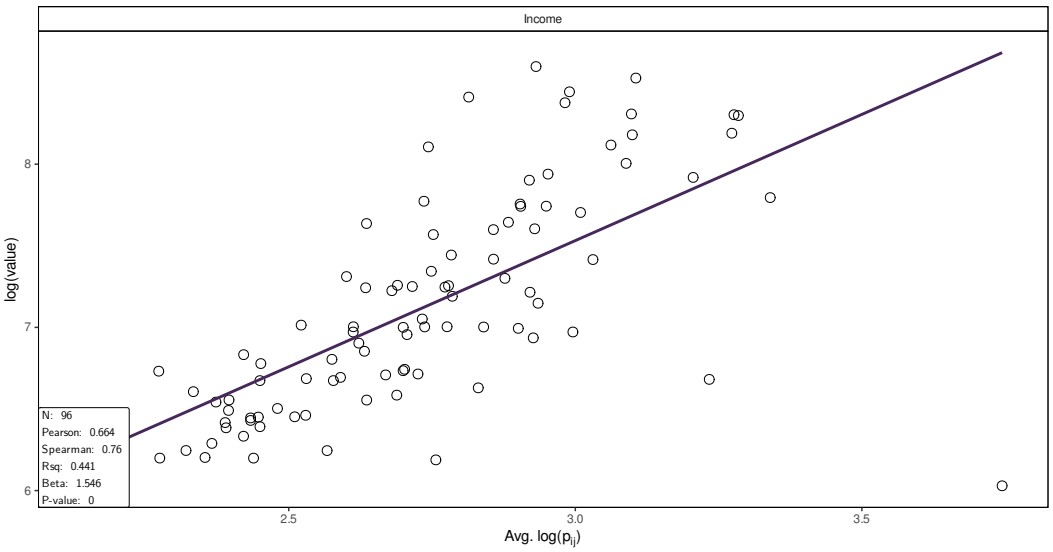


Figure 2.9: Avg. log(p_{ij}) $\times$ Income - São Paulo - Pizza

2.5
A toy application

In order to assess iFood data usefulness in providing up-to-date socioeconomic information at neighborhood level, we compute a neighborhood level[11] index based on pizza prices for 205 municipalities in Brazil[12]. The reason for

[11] Neighborhood level data is used when available. However, other reference units are used when neighborhood data is not available, such as subdistricts (e.g., for Brasília) and districts (e.g., for São Paulo).

[12] These are municipalities (1) with available shapefiles and income data at neighborhood level; (2) covered by iFood; (3) has at least 10 neighborhoods with schools (as well as academic achievement and enrolled students data available) laying inside it's respective polygons, as we trim municipalities sample so each municipality has at least 10 neighborhoods in estimation

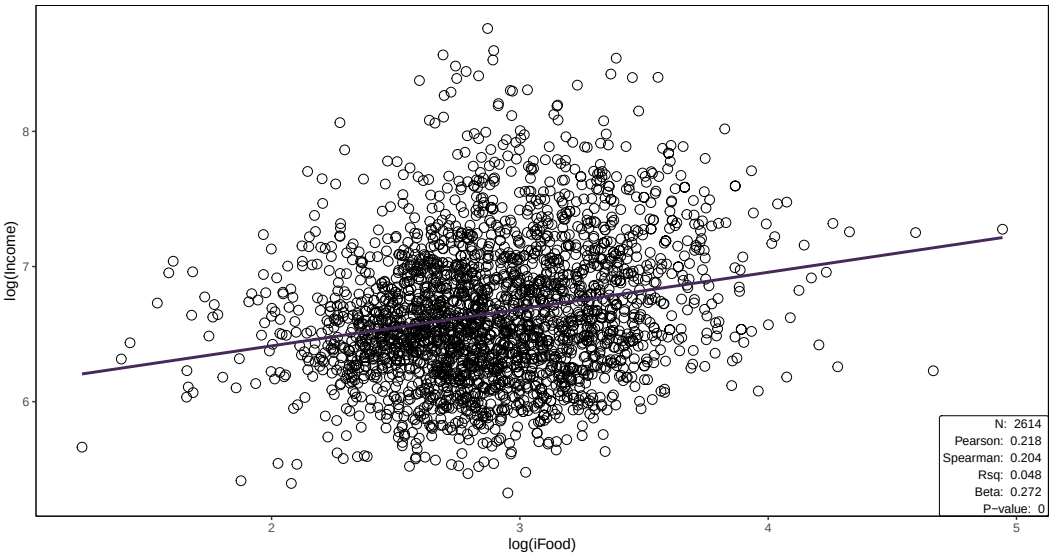


Figure 2.10: Brazilian neighborhoods - Scatterplot - iFood prices and Income per capita

choosing pizza relies on Chapter 2 results regarding the importance of choosing an adequate good for each municipality. It is assumed that pizza better approaches a general representative good than açaí does.

Once price indexes are calculated, we employ it as a control for a naïve regression specification for illustrating the relationship between school's number of enrolled students and academic achievement - potentially useful in, for example, supporting local policy-makers in deciding whether or where to open a new school. For each municipality j, We estimate Equation 2-3 and Equation 2-4 by OLS, where $Performance_{ij}$ is the average *performance index* in 2021 and MAT_{ij} is the average number of enrolled students according to 2021 Scholar Census, both at neighborhood i of municipality j. The performance index, in turn, is calculated at school level as the harmonic mean of pass rates (INEP, 2007; INEP, 2021).

$$Performance_{ij} = \alpha_{0j} + \alpha_{1j} MAT_{ij} + \alpha_{2j} \log(Income)_{ij} \tag{2-3}$$

$$Performance_{ij} = \tilde{\alpha}_{0j} + \tilde{\alpha}_{1j} MAT_{ij} + \tilde{\alpha}_{2j} \log(p_{ij}) \tag{2-4}$$

sample.

We compute the r-squareds, $R^2_{Income,j}$ and $R^2_{iFood,j}$, and then compare Income per capita and prices ability to cover variations in academic achievement. By doing such comparison, it is possible to evaluate whether using a timely available proxy is better than the outdated income measure. We also provide distinction between stages: Anos Iniciais (Elementary School), Anos Finais (Middle School/Junior High), and Ensino Médio (High School/Senior High).

Table 2.2 depicts the share of municipalities where each measure (Income or iFood's pizza prices) has a better fit. Table 2.3 presents descriptive statistics for the goodness of fit improvement driven by using iFood data, $R^2_{iFood,j}$ - $R^2_{Income,j}$. Controlling using the iFood's proxy provides a better fit for approximately half of the municipalities.

	Anos Iniciais	Anos Finais	Ensino Médio
Equal	0.00		
Income	0.48	0.48	0.54
iFood	0.51	0.52	0.46

Table 2.2: Better fits share across municipalities

Anos	N	Sd.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Anos Iniciais	201	0.11	-0.39	-0.03	0.00	0.01	0.05	0.56
Anos Finais	174	0.14	-0.58	-0.05	0.00	-0.02	0.04	0.43
Ensino Médio	114	0.10	-0.25	-0.06	-0.00	-0.00	0.05	0.36

Table 2.3: Difference between R^2_{iFood} and R^2_{Income} - Descriptive statistics

A natural following step is conditioning municipalities where iFood performs better than lagged Income. Because the median is essentially zero, and the interquartile range is reasonably symmetric with a roughly 5 percentage points distance to the median, a discontinuity at R^2_{iFood} - $R^2_{Income} = 0$ is not informative, as contains observations in which differences between iFood and Income in terms of goodness of fit are negligible. Alternatively, we look at the lower 25% quantile (25% worse iFood fits when compared to Income) against upper 25% (25% best iFood fits when compared to Income).

Table 2.4: Winning models distribution by Capital/Non-Capital

	Anos Iniciais		Anos Finais		Ensino Médio	
	Upper 25	Lower 25	Upper 25	Lower 25	Upper 25	Lower 25
Non-Capital	**0.961**	0.882	**0.955**	0.886	**0.931**	0.759
Capital	0.039	**0.118**	0.045	**0.114**	0.069	**0.241**

Table 2.5: Winning models distribution by Metropolitan/Non-Metropolitan Region

	Anos Iniciais		Anos Finais		Ensino Médio	
	Upper 25	Lower 25	Upper 25	Lower 25	Upper 25	Lower 25
Non-Metropolitan region	**0.412**	0.373	**0.455**	0.364	**0.276**	0.172
Metropolitan region	0.588	**0.627**	0.545	**0.636**	0.724	**0.828**

Table 2.4 shows that the share of non-capital municipalities increases among those in the upper 25% quantile, specially for Ensino Médio data. Similarly, the share of municipalities that do not belong to a metropolitan region also increases (Table 2.5).

We also analyse population data, income per capita, academic achievement, prices, average enrolled students and prices measured at municipality level. Population data is measured in 2022 Census, and Δ Population is the percentual change in population between 2022 and 2010 Census. Figure 2.11 shows that municipalities where iFood perfoms better are more likely to have smaller populations. These results support the argument that the iFood proxy is particularly relevant for smaller municipalities. Note, however, that income, population changes and academic achievement have heterogeneous behavior across stages specifications.

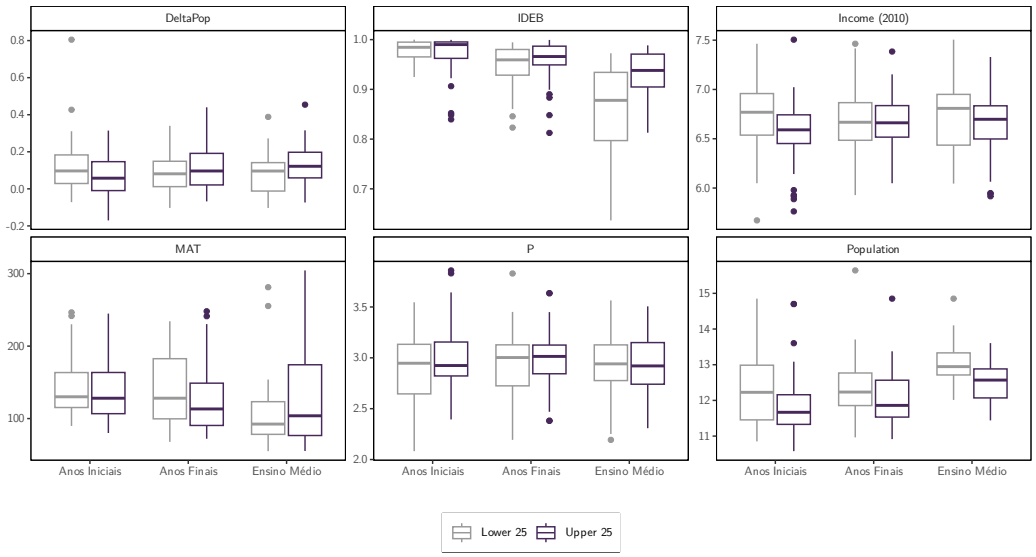


Figure 2.11: Difference between R^2_{iFood} and R^2_{Income} - Characteristics boxplots by quantiles

3
Bipartite Networks setup

Data from Rio de Janeiro suggest that iFood prices can be useful for unveiling local socioeconomic characteristics, at least for some location clusters. However, there is also a non-negligible set of observations which are not accounted by prices itself. Explaining why iFood prices do not fit some observations and clarifying which sources of variability it can not cover, as well as stating a formal economic link between these variables remains an open issue. This chapter contributes on this issue by proposing a more formal framework for better understanding how the network structure in the city interacts with prices and local characteristics.

3.1
Econometric framework

As argued in Section 2.3, iFood prices in a given location is affected by its surroundings. The goal for this section is modelling the firm-consumer relationship as a network structure, decomposing the effects of neighborhoods characteristics over price into *own-effects* - characteristics of neighborhood i over its own price - and *network-effects* - effects driven by neighborhoods $i' \neq i$ connected to i.

In the data, we observe for each centroid i the set of firms connected to i, which we denote by K_i. Now we introduce the other direction: each firm j is connected to a set C_j of centroids. This division generates a bipartite network, divided between firms and consumers (centroids) whom interact with each other. A bipartite network is a bipartite graph $\{V_1, V_2, E\}$, with V_1 and V_2 partitions of a set V of points, connected by lines E (BONHOMME, 2020).

Consider an additive model:

$$Y_{ij}^{*} = X_i'\beta + \alpha_i + \psi_j + \varepsilon_{ij}, \tag{3-1}$$

where Y_{ij}^{*} are potential outcomes that are realized if the link (i, j) exists, i.e., $D_{ij} = 1$. The observed Y_{ij} and potential outcome Y_{ij}^{*} coincide when $D_{ij} = 1$. Equation 3-1 can be consistently estimated through usual regression techniques as long as Assumption 3.1 holds.

Assumption 3.1 *Covariates* $X = \{X_{ij}\}_{i,j\in V}$*, existence of links* $D = \{D_{ij}\}_{i,j\in V}$*, and unobserved characteristics* $\alpha = \{\alpha_i\}_{i\in V_1}$ *and* $\psi = \{\psi_j\}_{j\in V_2}$ *are strictly exogenous:*

$$\mathbb{E}\left[\varepsilon_{ij}|D, X, \alpha, \psi\right] = 0. \tag{3-2}$$

For our particular problem, assume that firms set prices P_j based on the characteristics of connected centroids, C_j. Note that our previous approach in Section 2 predicts socioeconomic status - characteristics - given prices. Previously, our central issue implicitly writes these characteristics on the left-hand side. Now, Economic reasoning requires switching sides to clarify this relationship.

Let prices have a separable functional form,

$$P_j = \sum_{i\in C_j} w_{ji}\left(X_i'\beta + \xi_i\right) + z_j'\delta + \epsilon_j. \tag{3-3}$$

where X_i are observable characteristics and ξ_i is an unobserved one-dimensional characteristic of centroid i. Buyer's (centroid) i weight in firm's j price (KRAMARZ; MARTIN; MEJEAN, 2020) is denoted by w_{ji}, z_j are observable characteristics of firm j and ϵ_j accounts for its respective unobserved characteristics.

The observed characteristics should include (potentially previous) Census data information, or any other data that is available for a large sample of cities.

In this setting, we can understand our current exercise as constructing for each centroid i the normalized prices p_{ij},

$$p_{ij} = \tilde{w}_{ij} P_j, \tag{3-4}$$

if $j \in C_i$, and where $\tilde{w}_{ij} = \frac{N_{ij}}{\sqrt{Distance_{ij}}}$, as a particular case, weighting by the number of items offered by j to i, N_{ij}.

The average price observed at centroid i is then

$$\begin{aligned}\tilde{P}_i &= \frac{1}{\#K_i} \sum_{j \in K_i} p_{ij} \\ &= \frac{1}{\#K_i} \sum_{j \in K_i} \tilde{w}_{ij} \left(\sum_{i' \in C_j} w_{ji'} \left(x'_{i'} \beta + \xi_{i'} \right) \right) + \underbrace{\frac{1}{\#K_i} \sum_{j \in K_i} \tilde{w}_{ij} z'_j}_{\tilde{z}_i} \delta + \underbrace{\frac{1}{\#K_i} \sum_{j \in K_i} \tilde{w}_{ij} \epsilon_j}_{\tilde{\epsilon}_i}\end{aligned} \tag{3-5}$$

One can express this formulation in matrix notation. Let

$$\widetilde{\mathbf{W}} = \begin{bmatrix} \frac{1}{\#K_1} \sum_{j \in K_1} \tilde{w}_{1j} w_{j1} & \frac{1}{\#K_1} \sum_{j \in K_1 \cap K_2} \tilde{w}_{1j} w_{j2} & \cdots & \frac{1}{\#K_1} \sum_{j \in K_1 \cap K_I} \tilde{w}_{1j} w_{jI} \\ \frac{1}{\#K_2} \sum_{j \in K_2 \cap K_1} \tilde{w}_{2j} w_{j1} & \frac{1}{\#K_2} \sum_{j \in K_2} \tilde{w}_{2j} w_{j2} & \cdots & \frac{1}{\#K_2} \sum_{j \in K_2 \cap K_I} \tilde{w}_{2j} w_{jI} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{1}{\#K_I} \sum_{j \in K_I \cap K_1} \tilde{w}_{Ij} w_{j1} & \frac{1}{\#K_I} \sum_{j \in K_2} \tilde{w}_{Ij} w_{j2} & \cdots & \frac{1}{\#K_I} \sum_{j \in K_I} \tilde{w}_{Ij} w_{jI} \end{bmatrix}, \tag{3-6}$$

then

$$\widetilde{\mathbf{P}} = \widetilde{\mathbf{W}} \left(\mathbf{X} \beta + \xi \right) + \widetilde{\mathbf{Z}} \delta + \widetilde{\epsilon}. \tag{3-7}$$

This provides a decomposition between how much of $\widetilde{\mathbf{P}}$ is explained by centroid characteristics, $\widetilde{\mathbf{W}} \left(\mathbf{X} \beta + \xi \right)$, and how much is explained by firm characteristics, $\widetilde{\mathbf{Z}} \delta + \widetilde{\epsilon}$.

Although the previous finding suggests that $\widetilde{\mathbf{P}}$ is correlated with SPI_i, this framework allows for investigating at least two additional sources of variation in $\widetilde{\mathbf{P}}$: first is explained by $\widetilde{\mathbf{Z}}$, that is, the characteristics of firms connected to those centroids, for example, ratings or whether a firm has joined

in iFood recently[1]. The second is the network effect, i.e. centroids $i' \in C_j$ other than i connected to $j \in K_i$. The latter effect addresses a possible explanation for the SPI and $\tilde{P}$ mismatch in some locations.

3.2 Empirical exercise

Turning back on data, the bipartite networks setup relevance can be visualized through a networks diagram (Figure 3.1).

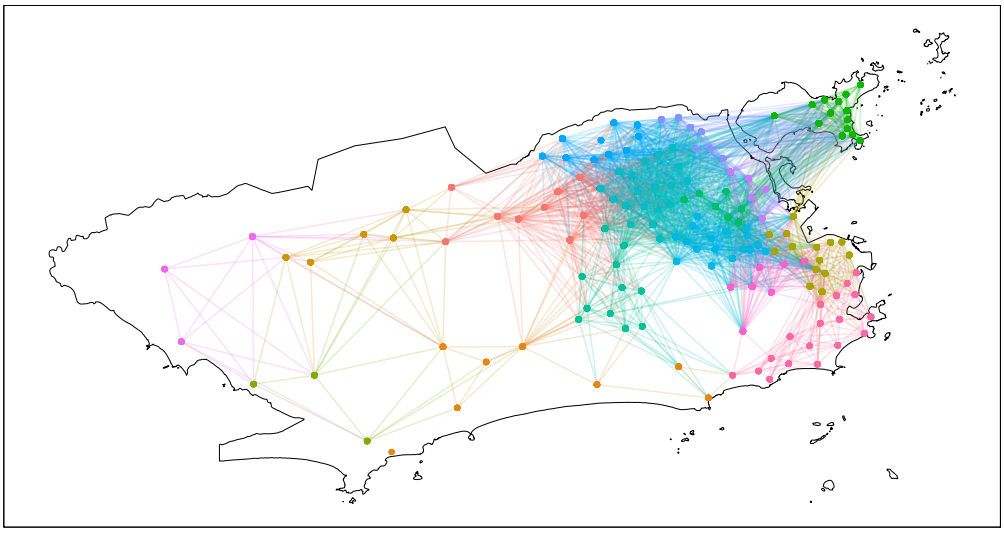


Figure 3.1: Neighborhoods interconnection, colored by RP

Because Equation 3-7 is additive, we can decompose $\widetilde{\mathbf{W}}\mathbf{X}$ into $\widetilde{\mathbf{W}}\mathbf{X}_i = \mathbf{Idiag}(\widetilde{\mathbf{W}})\mathbf{X}$, a diagonal matrix which accounts for the effects of the characteristics of neighborhood i over i's own price, and analogously $\widetilde{\mathbf{W}}\mathbf{X}_{-i} = \left[\widetilde{\mathbf{W}} - \mathbf{Idiag}(\widetilde{\mathbf{W}})\right]\mathbf{X}$, which diagonal entries are all zero, and accounts for the characteristics of all neighborhoods $\{i' \neq i | \exists j : i', i \in C_j\}$. I set $w_{ij} = 1$, assuming that firms weights buyers uniformly. Table 3.1 shows regression results. SAR specifications were estimated by maximum likelihood.

Columns (5) in Table 3.1 depicts Rio de Janeiro estimates for Equation 3-7. The networks setup extended with firm characteristics outperforms plain OLS setups, with and without firm characteristics (Columns 5 and 2

[1] A similar in spirit consideration appears in (BILAL, 2023), which uncovers local determinants of unemployment.

respectively). One can also note that coefficients weight are symmetric for the neighborhood itself ($\widetilde{\mathbf{W}}\text{SPI}_i$) and its surroundings ($\widetilde{\mathbf{W}}\text{SPI}_{-i}$).

Table 3.1: OLS, Networks, and Spatial Autoregression Results

	Dependent variable:						
	$\widetilde{\mathbf{P}}$ (in logs)						
	OLS	OLS	SAR	OLS	OLS	SAR	SAR
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SPI	0.026***		0.012***	0.007**		0.003*	
	(0.005)		(0.003)	(0.003)		(0.002)	
$\widetilde{\mathbf{W}}\text{SPI}_i$		0.010***			0.006***		0.004***
		(0.0004)			(0.001)		(0.001)
$\widetilde{\mathbf{W}}\text{SPI}_{-i}$		0.007***			0.006***		0.003***
		(0.001)			(0.002)		(0.001)
isNew				−0.313	−0.490***	0.068	−0.152
				(0.284)	(0.158)	(0.143)	(0.127)
userRating				0.265***	0.121***	0.198***	0.114***
				(0.020)	(0.031)	(0.012)	(0.015)
Constant	0.798**	1.395***	−0.150	1.251***	1.397***	0.505***	0.764***
	(0.313)	(0.079)	(0.169)	(0.174)	(0.084)	(0.117)	(0.101)
Observations	157	157	157	157	157	157	157
Adjusted R^2	0.250	0.823		0.794	0.871		
Akaike Inf. Crit.	43.708	−181.629	−61.329	−157.490	−229.997	−239.208	−286.474

Note: *p<0.1; **p<0.05; ***p<0.01

HC1 errors for OLS estimates

3.2.1
Networks and conventional spatial methods

One can interpret Equation 3-3 as a specific method to address spatial dependence. By leveraging spatial dependence structure, it is convenient to compare Equation 3-7 against a plain spatial autoregressive (SAR) model:

$$\widetilde{\mathbf{P}} = \rho \mathbf{W}^{\text{S}} \widetilde{\mathbf{P}} + \widetilde{\mathbf{X}}\beta + \widetilde{\mathbf{Z}}\delta + \varepsilon, \tag{3-8}$$

where $\mathbf{W}^{\text{S}}$ is a spatial weights matrix, computed by assigning 1 to entries i, j if i and j are neighbors - i.e., share an edge or vertex - and normalizing rows to sum 1:

$$w_{ij}^{\mathrm{S}} = \begin{cases} \dfrac{\mathbb{1}\left\{i \text{ and } j \text{ are contiguous}\right\}}{\sum_{j=1}^{N} \mathbb{1}\left\{i \text{ and } j \text{ are contiguous}\right\}} & \text{, if } j \neq i \\ 0 & \text{, otherwise.} \end{cases} \tag{3-9}$$

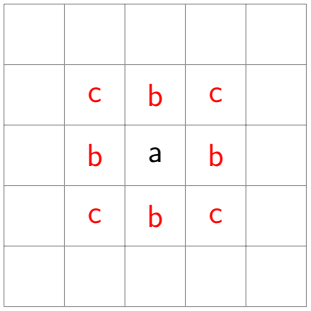


Figure 3.2: Radius (queen) contiguity, as in Anselin (1988)

Columns (3), (6) and (7) in Table 3.1 depicts results for SAR specifications. Results suggest that the networks setup with firm characteristics drives comparable results in terms of AIC comparable to the SAR specification with firm characteristics. Moreover, a SAR specification combined with a network structure (Column 7) significantly improves the fitting. A possible explanation for this latter result is that the SAR side captures characteristics other than the SPI.

3.2.2 Alternative scenarios

The analysis of data from São Paulo reveals a different interaction between territorial units, which diminishes the networks' role in determining prices. This stands in contrast to the Rio de Janeiro data, where a high level of interconnectivity and significant network effects are observed. An initial hypothesis is that the lower granularity of the São Paulo data, characterized by larger territorial units, contributes to a reduced interconnectivity. To test this hypothesis, an approach could involve aggregating the Rio de Janeiro data into larger territorial units to examine interconnectivity and the potential effects of networks.

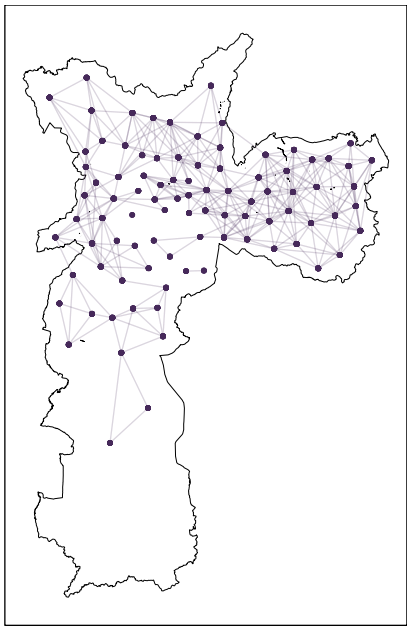


Figure 3.3: Neighborhoods interconnection - São Paulo

We propose aggregating Rio de Janeiro's neighborhoods into Administrative Regions (RA). We compute RA level SPI by averaging neighborhoods SPI. Prices, on the other hand, are collected by scrapping prices for centroids calculated at RA level. Figure 3.4 shows that this approach heterogeneously enlarges territorial units: notably, RAs are larger on the west side of the city, while units remain relatively small and highly interconnected on the east side. Overall, network still relevant source of variation in prices, as seen in Table 3.2, and this aggregation substantially diminishes spatial lags relevance: the networks specification estimated with OLS outperforms the networks setup with spatial lags.

In principle, one could suggest aggregating neighborhoods following a mathematical criterion - e.g., such that units have uniform area - instead of the existing territorial division, as an attempt to mitigate the granularity on the east side of the city. Completely decoupling from existing territorial setup, however, brings interpretability issues. Hence, we consider handling this mismatch as city-level particular characteristics. From a materialistic point of view infrastructure or geographic restrictions may vary across the city. However, our understanding is that testing this hypothesis exceeds the scope

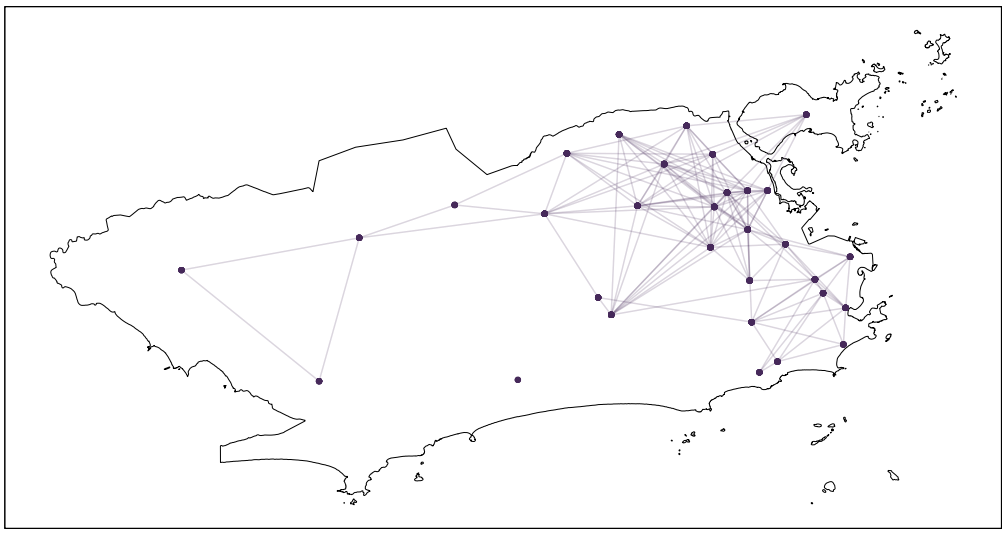


Figure 3.4: Neighborhoods interconnection - RA level

Table 3.2: OLS, Networks, and Spatial Autoregression Results - RA level

	Dependent variable:						
	$\widetilde{\mathbf{P}}$ (in logs)						
	OLS	OLS	SAR	OLS	OLS	SAR	SAR
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SPI	0.036**		0.035***	0.011		0.010	
	(0.016)		(0.011)	(0.007)		(0.007)	
$\widetilde{\mathbf{W}}$SPI$_i$		0.007***			0.009***		0.009***
		(0.001)			(0.002)		(0.002)
$\widetilde{\mathbf{W}}$SPI$_{-i}$		0.004**			0.004**		0.004
		(0.002)			(0.002)		(0.003)
isN$\tilde{}$ew				0.289	−0.560**	0.225	−0.591**
				(0.402)	(0.282)	(0.313)	(0.290)
userR$\tilde{}$ating				0.176***	−0.019	0.180***	−0.011
				(0.029)	(0.046)	(0.029)	(0.052)
Constant	0.157	1.795***	0.054	1.153***	1.805***	1.076**	1.710***
	(0.955)	(0.089)	(0.753)	(0.420)	(0.069)	(0.429)	(0.165)
Observations	31	31	31	31	31	31	31
Adjusted R^2	0.231	0.826		0.755	0.833		
Akaike Inf. Crit.	27.215	−17.897	28.904	−6.496	−17.501	−5.388	−16.401
Note:							*p<0.1; **p<0.05; ***p<0.01
	HC1 errors for OLS estimates						

of this work.

As in Chapter 2, we look into pizza data for São Paulo. It returns a slightly more interconnected map (Figure 3.5).

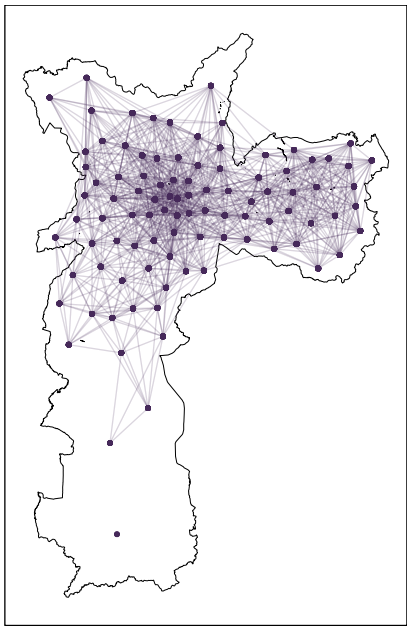


Figure 3.5: Neighborhoods interconnection - Pizza in São Paulo

Table 3.3: OLS, Networks, and Spatial Autoregression Results - São Paulo - Pizza

	Dependent variable:						
	$\tilde{\mathbf{P}}$ (in logs)						
	OLS	OLS	SAR	OLS	OLS	SAR	SAR
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Income	0.286***		0.158***	0.196***		0.054*	
	(0.044)		(0.036)	(0.047)		(0.031)	
$\widetilde{\mathbf{W}}$Income$_i$		0.041***			0.027***		0.015***
		(0.006)			(0.009)		(0.004)
$\widetilde{\mathbf{W}}$Income$_{-i}$		0.142***			0.077		−0.032
		(0.054)			(0.055)		(0.021)
isN$\tilde{\text{e}}$w				−0.063	0.710	−0.464	−0.309
				(1.593)	(1.474)	(0.832)	(0.654)
userR$\tilde{\text{a}}$ting				0.254***	0.203**	0.256***	0.189***
				(0.052)	(0.081)	(0.040)	(0.041)
Constant	0.699**	1.051***	0.152	0.634**	1.157***	0.030	0.050
	(0.324)	(0.334)	(0.233)	(0.297)	(0.338)	(0.202)	(0.181)
Observations	96	96	96	96	96	96	96
Adjusted R^2	0.436	0.444		0.551	0.482		
Akaike Inf. Crit.	−18.938	−19.484	−37.833	−38.907	−24.350	−68.793	−103.862

Note: *p<0.1; **p<0.05; ***p<0.01

HC1 errors for OLS estimates

4
High frequency data and Price Indexes

Previous chapter leveraged spatial granularity provided by iFood prices. As discussed in Chapter 1, constructing and assessing properties of price indexes played a major role in online prices research agenda. This chapter explores iFood data ability in providing a price index on daily basis.

First, data collection and product classification procedures are introduced. Then, iFood inflation estimates are compared against official numbers. This monthly basis comparison is useful for validating what iFood prices actually measure. On the top of positioning iFood as a feasible source of real time, cheaper inflation data, this validation step permits analysing daily dynamics with certain reliability.

It is verified that iFood inflation estimates are notably more volatile than the official one. Two hypotheses for this additional volatility are considered: one is that changes in the vendors set drives additional volatily; other uses price stickiness statistics to evaluate whether prices indeed vary more. Finally, we use daily estimates for evaluating how inflation evolves within a month, specially in the context of Climate change, using the Porto Alegre historical flood case.

4.1
Data and Methods

This section goal is to describe the data collection procedure and the product classification step. The scrapping software was executed in daily frequency sequentially for each subitem α belonging to the Extended National Consumer Price Index (IPCA) basket for city l.

For computational simplicity, analyses in this chapter renounces degrees of freedom in spatial dimension and treats cities as the relevant spatial unit. Thus, the scrapping software is executed for each subitem in a representative

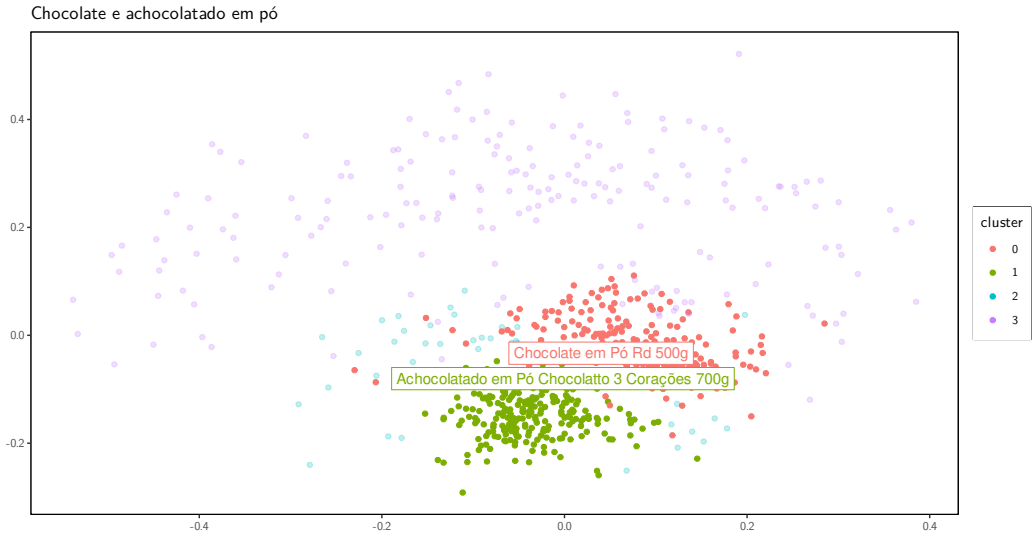


Figure 4.1: Product clusters - Chocolate e achocolatado em pó

neighborhood for each location[1]. The full list of representative neighborhoods and IPCA weights for each location is available in Table B.1. Data collection started on March 24$^{\text{th}}$, 2024.

Once data was gathered, a inspection shows that many observations are not related to the subitem of interest. A semi-supervised learning approach is employed for ruling-out undesired entries[2].

Figure 4.1 provides clusters and Principal Coordinate Analysis for a illustrative subitem, *Chocolate e achocolatado em pó* (Powdered chocolate and Powdered chocolate drink). Opaque clusters are those kept in sample after product classification procedure, and text boxes depicts the label associated with clusterwise centroids. Although procedure works reasonably well for this particular subitem, it is necessary to note that performance vary substantially throughout different subitems.

Finally, indexes are calculated following the methodology proposed in IBGE (2020)[3]. Though the index is calculated on daily basis, there are different

[1]The software was set to start execution at 10 AM.

[2]See Appendix - Product Classification

[3]The mean prices for each product, in each location, at each period are smoothed through Kalman filtering, and the relative for product j belonging to subitem α, in location l at day t of month m, is calculated as follows: $R_{j,\alpha}^{l,m,t} = \bar{p}_{j,\alpha}^{l,m,t} / \bar{p}_{j,\alpha}^{l,m,1}$. Monthly estimates for month m are simply the daily index at last day of month m.

approaches for defining the relevant step for analysing price variation: a 1-day approach, comparing day t prices against $t-1$ (1-day step), or comparing t prices against its counterpart on the previous month - $t-30$, or $t-28$ (four weeks). This specific exercise uses the former approach.

Location	Retailers	Observations (M)	Products (K)	Days	Initial Date	Final Date
AJU	1113	2.81	14.76	214	2024-03-24	2024-11-01
BEL	1655	3.48	18.85	214	2024-03-24	2024-11-01
BH	2100	2.65	18.81	214	2024-03-24	2024-11-01
CG	1560	6.82	20.99	214	2024-03-24	2024-11-01
CUR	2357	6.36	27.85	214	2024-03-24	2024-11-01
DF	1583	3.02	16.68	214	2024-03-24	2024-11-01
FOR	1580	2.93	14.72	214	2024-03-24	2024-11-01
GOI	2050	3.84	20.20	214	2024-03-24	2024-11-01
POA	1965	3.10	22.58	214	2024-03-24	2024-11-01
RB	777	0.81	8.62	214	2024-03-24	2024-11-01
REC	1615	6.41	18.72	214	2024-03-24	2024-11-01
RJ	1926	4.44	23.06	214	2024-03-24	2024-11-01
SAL	1443	2.56	15.43	214	2024-03-24	2024-11-01
SL	1442	1.60	12.02	214	2024-03-24	2024-11-01
SP	4544	3.15	30.92	214	2024-03-24	2024-11-01
VIT	1740	4.32	19.24	214	2024-03-24	2024-11-01

Table 4.1: Database Description

4.2 Comparison against official inflation measures

Once indexes are calculated, iFood estimates are compared against official inflation. Beacause official inflation is released on monthly frequency, the sample is too short for any inference through usual statistical methods at national level. This issue is bypassed by comparing indexes at city level through a cross-city panel.

Figure 4.2 shows daily inflation estimates for each location. Figure 4.3 shows a comparison between IPCA for the Food and Beverage group and iFood price index at local level, and a confusion matrix of the signs is depicted in Table 4.3. Table 4.2 indicates that iFood index is far more volatile than IPCA, by spanning a wider range than IPCA does, and is, on average, lower than IPCA.

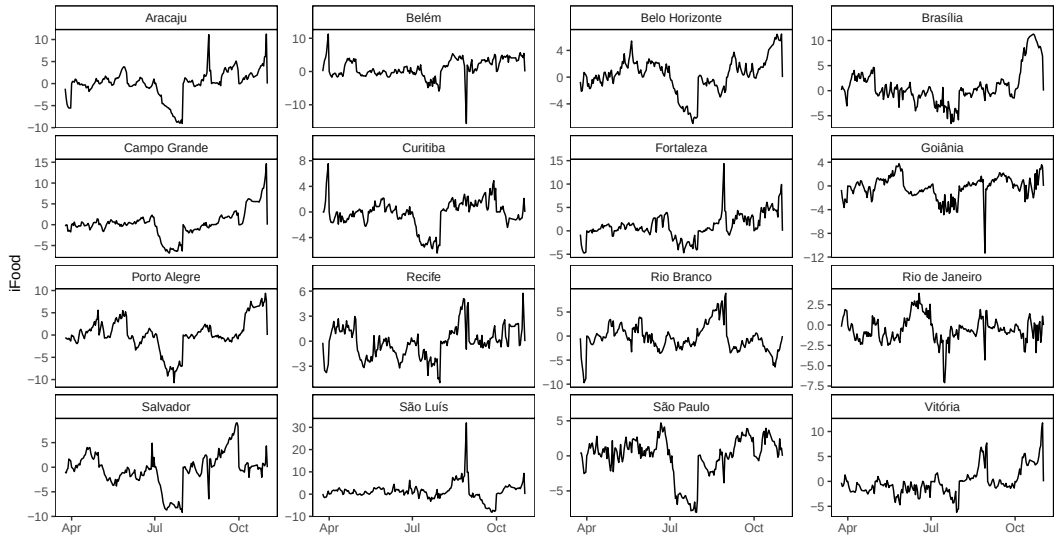


Figure 4.2: iFood price index, Daily variation (%)

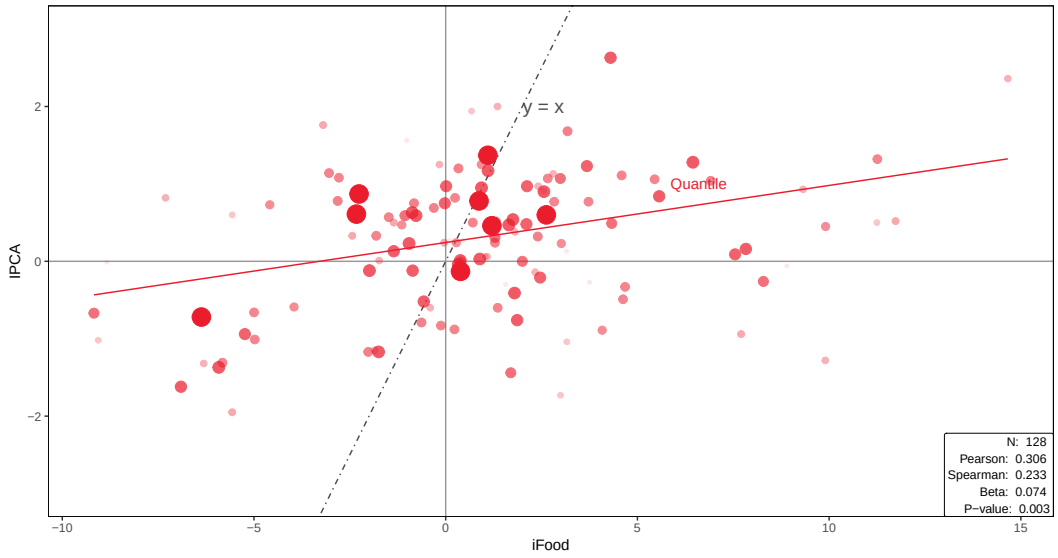


Figure 4.3: iFood price index and IPCA Food and Beverage comparison, Monthly variation (%). Bubble sizes represent location's weight on national index

	N	T	N × T	Std. Dev.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
IPCA	16	8	128	0.88	-1.95	-0.28	0.33	0.25	0.84	2.63
iFood	16	8	128	4.31	-9.17	-1.35	0.88	0.91	2.81	14.66

Table 4.2: Descriptive statistics - iFood price index and IPCA

	Negative	Positive
Positive	0.21	0.45
Negative	0.17	0.16

Table 4.3: iFood price index (columns) and IPCA Food and Beverage (rows) confusion matrix, Monthly variation (%)

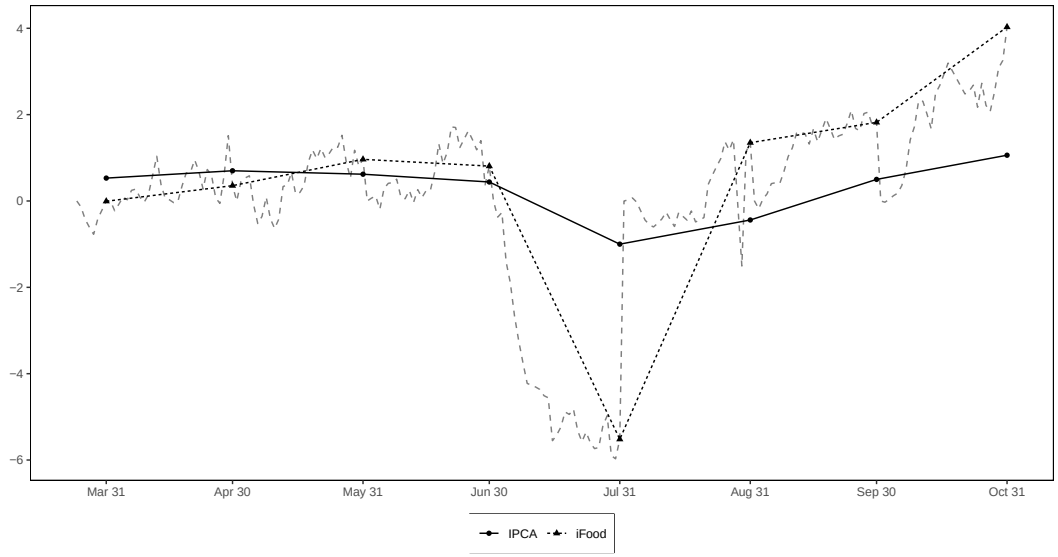


Figure 4.4: iFood price index and IPCA Food and Beverage comparison, national index, Monthly variation (%), and iFood price index, national index, Daily variation (%)

Looking at national level, Figure 4.4 shows that differences between two indexes appear to be smoothed. iFood exacerbated deflation in July. Table 4.4 and Table 4.5 show subitems with lowest and largest Mean Absolute Deviation (MAD) between iFood and IPCA.

4.3 Exploring volatility sources

As discussed above, iFood index is remarkably more volatile than official inflation data. One may argue that additional variability in iFood index could be due to vendors participation, that is, changes in data composition in terms of suppliers participation throughout the sample. Another possibility is that

	Subitem	MAD
1	Sardinha em conserva	0.81
2	Milho-verde em conserva	0.87
3	Leite em pó	1.25
4	Pepino em conserva	1.35
5	Massa semipreparada	1.35
6	Abacaxi	1.90
7	Sorvete	1.98
8	Pão francês	1.99
9	Queijo	2.13
10	Iogurte e bebidas lácteas	2.15

Table 4.4: iFood price index and IPCA Food and Beverage, 10 lowest MAD, national index

	Subitem	MAD
1	Salsicha em conserva	65.01
2	Suco em pó	58.15
3	Colorau	49.17
4	Tangerina	31.28
5	Peito	22.82
6	Maracujá	22.37
7	Abobrinha	19.72
8	Manga	19.59
9	Inhame	19.52
10	Sopa desidratada	18.72

Table 4.5: iFood price index and IPCA Food and Beverage, 10 largest MAD, national index

this volatility is that adjustments are more frequent or have larger magnitude. This section explores these two possibilities.

4.3.1
Decomposing into changes in vendors set

This first hypothesis for additional volatility is evaluated by calculating iFood index on a restricted sample, imposing that observations for vendor j at day t are kept if and only if j is observed on $t-1$.

Figure 4.5 compares National iFood index in daily frequency, with and without this restriction. Table 4.6 shows regression results of iFood index on restricted one (i$\tilde{\text{Fo}}$od), including dummies for April and July periods. It suggests that indexes are reasonably identical, except for April and July.

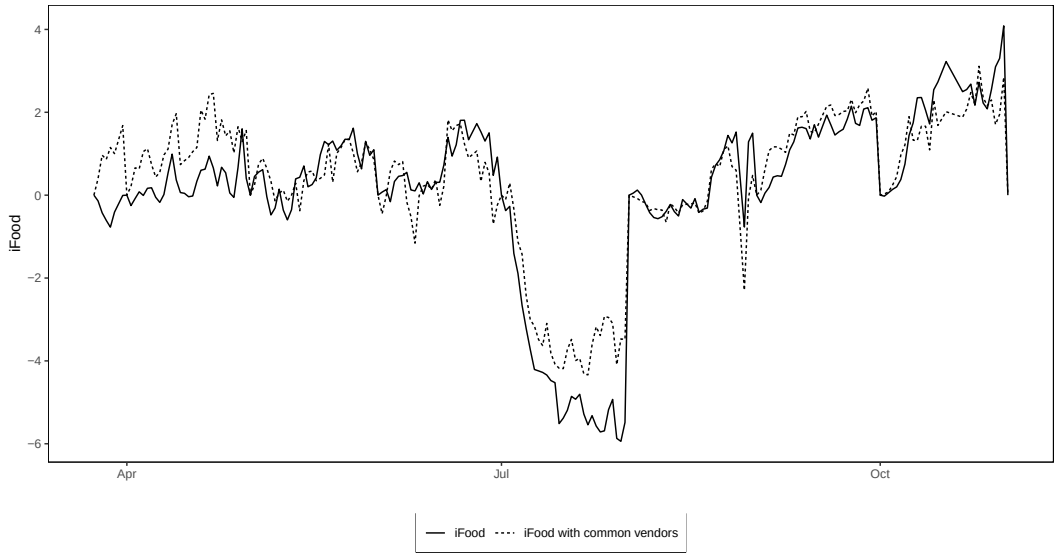


Figure 4.5: Inflation Decomposition - Changes in vendors set

Table 4.6: Inflation Decomposition

	Dependent variable:			
	iFood			
	Panel Data	Panel Data	National Index	National Index
D(April or July)		-0.803^{***}		-1.080^{***}
		(0.039)		(0.103)
$\widetilde{\text{iFood}}$	0.958^{***}	0.924^{***}	1.158^{***}	1.021^{***}
	(0.006)	(0.006)	(0.030)	(0.028)
Constant	-0.227^{***}	0.014	-0.339^{***}	0.013
	(0.018)	(0.021)	(0.051)	(0.053)
Observations	3,424	3,424	214	214
R^2	0.867	0.882	0.873	0.917
Note:				*p<0.1; **p<0.05; ***p<0.01

National iFood index is 93% correlated with its restricted counterpart. Moreover, assuming that regressions residuals approximate the inflation due to changes in sample composition in terms of merchants, this effect drives only 8% to 14% of changes in inflation index, and hence can't explain iFood's excess of volatility.

4.3.2 Price Stickiness statistics

Once changes in available vendors does not accounts for the bulk of variation in iFood compared to IPCA, the latter can be likely explained by differences in price changes iteself. Estimates for price spells frequency and implicit duration expressed in months are shown in Table 4.7[4].

Macro Item	Frequency	Duration
Median	0.05	0.65
Alimentação fora do domicílio	0.03	1.24
Açúcares e derivados	0.03	0.94
Enlatados e conservas	0.04	0.89
Sal e condimentos	0.04	0.87
Panificados	0.04	0.84
Pescados	0.04	0.82
Farinhas, féculas e massas	0.04	0.79
Bebidas e infusões	0.05	0.69
Óleos e gorduras	0.05	0.59
Aves e ovos	0.06	0.55
Carnes e peixes industrializados	0.06	0.55
Carnes	0.06	0.51
Leites e derivados	0.07	0.48
Cereais, leguminosas e oleaginosas	0.07	0.48
Hortaliças e verduras	0.07	0.43
Tubérculos, raízes e legumes	0.08	0.40
Frutas	0.08	0.39

Table 4.7: Frequency and Monthly Durations

Unfortunately, calculating respective statistics for IPCA (or for physical stores) counterpart exceeds the scope of this work. Literature (BARROS,

[4]Frequencies are calculated first on product name (label in platform) and merchant level as the number of price changes over the numbers of days in which such pair is observed. Then frequencies are progressively averaged up to name, cluster and item level, and, finally, median frequency across items and macro item average frequencies are reported. Implicity durations are calculated as $-1/\log(1 - frequency)$ and converted to month units.

2009) estimate for monthly duration using daily online data for Brazil is 1.8 months. Meanwhile, other estimate using daily online data (CAVALLO, 2018) for Brazil is 1.27 months, which suggests a roughly 50% drop in duration by using iFood data, though, comparing these two statistics is not straightforward due to differences in time periods (2007-2010). Also, later author's estimates for Brazil are based on a single retailer, what, besides motivating robustness concerns, directly decreases price change likelihood, so does frequency, and increases duration. By all means, such result is consistent with an hypothesis that digital platforms increase price changes frequency, which may reflect, for instance, large-scale algorithmic adjustments which were not as popular by the 2000's.

In terms of change size, estimates on Table 4.8 are similar to those for Brazilian CPI data (13% mean, 8.3% median) (BARROS, 2009) and online data (11.52 mean) (CAVALLO, 2018). This finding provides additional argument for price changes frequency and changes size driving additional volatility in iFood index.

An additional insight from Table 4.8 relies on the fact that goods characterized by less value-aggregation steps, such as Fruit and Vegetables, Meats and Fishes, has larger and more frequent price changes. On the other hand, Food Outside Home, including final meals and fast-food, which naturally encompasses more intermediary steps, has both smaller (median) size and frequency of price changes.

4.4 Climate change and daily inflation

After comparing iFood inflation against its official counterpart and exploring candidate drivers of its addional volatility, this section uses iFood daily data in the context of Porto Alegre 2024 flood to evaluate how climate change and major events associated with this scenario may affect daily inflation.

Rio Grande do Sul state was hitten by severe floods in May/24, which

Table 4.8: Change Size and Absolute Change Size (%)

	Price Change		Absolute Price Change	
Macro Item	Median	Mean	Median	Mean
All	1.56	5.78	9.10	17.20
Hortaliças e verduras	0.00	5.00	17.19	25.26
Tubérculos, raízes e legumes	0.00	3.03	13.58	20.26
Frutas	2.00	4.79	12.69	19.81
Carnes	2.71	8.20	11.27	21.86
Pescados	2.12	3.80	10.72	16.31
Panificados	2.27	3.94	10.22	16.58
Farinhas, féculas e massas	1.92	3.35	9.60	15.08
Açúcares e derivados	1.99	3.40	9.52	14.79
Carnes e peixes industrializados	1.40	3.48	9.23	16.21
Enlatados e conservas	2.00	2.79	9.10	13.81
Óleos e gorduras	1.65	1.99	8.07	12.27
Aves e ovos	0.00	2.46	8.01	14.13
Alimentação fora do domicílio	4.18	141.30	8.00	148.33
Leites e derivados	1.77	3.30	7.92	12.74
Cereais, leguminosas e oleaginosas	1.43	2.55	7.50	12.31
Sal e condimentos	1.11	3.41	7.37	13.51
Bebidas e infusões	0.00	3.24	7.29	12.05

Table 4.9: Change Size and Absolute Change Size (%) and Durations (months)

	Duration	Price Change		Abs. Price Change	
Macro Item	Median	Median	Mean	Median	Mean
iFood	0.65	1.56	5.78	9.10	17.20
Cavallo (2018)	1.27				11.52
Barros (2009)	1.8			8.3	13

resulted in more than 170 deaths and 420 thousand people homeless, besides impacts over power and clean water supply, communication services, and infrastructure, such as roads and bridges. In Porto Alegre, state's capital, Guaíba river rose up to 5.31 meters, beating previous the previous record (4.76 meters) achieved in 1941 historical flood. Several locations became inaccessible due to the flood. The International Airport was also hitten by the flood, being closed for indeterminate period[5].

Both IPCA (2.63%) and iFood (4.31%) predict a high inflation for Porto

[5]See Povo (2024b), Defesa Civil do Rio Grande do Sul (2024), Zanatta, Rigue e Lauxen (2024) and Povo (2024a)

Figure 4.6: iFood price index for Porto Alegre, 28 days basis monthly variation (%), and Guaíba river level, in meters

Alegre in May. Using daily inflation estimates, it is possible to asses how inflation evolved along the month. For each day t in May 24, a daily estimate for the monthly inflation is calculated as the 28-days basis inflation.

National Water and Sanitation Agency (ANA) data provides 15-minutes frequency telemetry measurements for Guaíba level[6]. In this exercise, the daily maximum level is employed as a daily measure for Guaíba water level. Figure 4.6 shows inflation path for Porto Alegre in May 24, as well as Guaíba level in daily frequency. For reference, two relevant points are also highlighted: the Flood reference Level and the Peak Level observed in 1941 flood, the worst in history until 2024. Inflation rise as Guaíba level does in early May, and variations keep in line until half of June. Table 4.10 depicts regression results for iFood price index on Guaíba level for the sample between April 2024 and June 2024, and for the sample restricted to May 2024. It shows that during flood month variations in Guaíba level substantially explained variations in inflation.

[6]See HIDRO - Dados Hidrometeorologicos por estação.

Table 4.10: Guaíba level and daily inflation

	Dependent variable:	
	Δ log(iFood)	
	Apr/24 - June/24	May/24
Δ Level	0.149	0.244***
	(0.207)	(0.042)
Observations	68	30
Adjusted R^2	−0.007	0.526
F Statistic	0.515 (df = 1; 66)	33.223*** (df = 1; 28)
Note:		*p<0.1; **p<0.05; ***p<0.01

5
Concluding remarks

This study analyzed price data scraped from iFood, finding that in Rio de Janeiro, açaí prices effectively approximate the Social Progress Index (SPI) at the neighborhood level. Since SPI correlates with other economic indicators like income, this suggests that online prices from local delivery services can serve as a low-cost, feasible proxy for neighborhood-level socioeconomic measures. This approach is especially useful in Brazilian cities where timely and reliable socioeconomic statistics are lacking. A noteworthy drawback, though, is that this study overlooks a structural pricing theory.

On the time granularity side, high-frequency data from iFood was employed to create an online Consumer Price Index (CPI) and daily inflation estimates using machine learning to address product classification challenges. While, in our window, iFood data is positively correlated with official inflation bechmarks on a monthly basis at local level, the national-level online indicator aligns with official numbers, albeit over a limited evaluation window.

These findings highlight the potential of iFood data for inflation measurement at significantly lower costs than traditional methods. Monthly estimates are available immediately at the end of the reference month, often preceding official releases. Moreover, daily data can construct preliminary monthly inflation estimates and offer insights into intra-month dynamics.

The index calculated using online prices is remarkably more volatile than official inflation estimates. This study provide evidence that this additional volatility is not driven by changes in the suppliers set. Price stickiness statistics indicate that price changes in our dataset are more frequent than in other data sources. Also, the size of price changes is larger in iFood data. This findings provide evidence on the nature of additional volatility. Further research may provide more insight into this question by understanting *why* iFood prices are

less sticky.

Daily inflation estimates are crucial for understanding responses to atypical events. This paper illustrated this with examples from May 2024, showing how daily inflation in Porto Alegre (following a natural disaster) revealed dynamics not captured by monthly observations. As climate change increases the frequency of such events, accurate measurement and understanding of their economic impacts become increasingly important.

Finally, these encouraging results encompasses some methodological choices which can be reviewed. It's not clear how these results are responsive to changes in filtering procedures, or the time steps when calculating the index, for instance. Enriching results through a theoretical framework, further exploring alternative approaches, as well as the questions that arose from the stylized facts provided by this study appear to offer a promising direction for future literature.

6
Bibliography

ABRAHAM, K. G. et al. Introduction: Big data for twenty-first-century economic statistics: The future is now. In: ABRAHAM, K. G. et al. (Ed.). **Big Data for Twenty-First-Century Economic Statistics**. [S.l.]: University of Chicago Press, 2022. p. 1–22. ISBN 978-0-226-80125-4.

ANGRIST, N.; GOLDBERG, P. K.; JOLLIFFE, D. Why is growth in developing countries so hard to measure? **Journal of Economic Perspectives**, v. 35, n. 3, p. 215–42, August 2021. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/jep.35.3.215>.

ANSELIN, L. **Spatial Econometrics: Methods and Models**. [S.l.]: Springer Science & Business Media, 1988. v. 4.

AUERBACH, E. Identification and estimation of a partially linear regression model using network data. **Econometrica**, v. 90, n. 1, p. 347–365, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19794>.

BALDACCI, E. et al. Big data and macroeconomic nowcasting: From data access to modelling. **Luxembourg: Eurostat. Doi: http://dx. doi. org/10.2785/360587**, v. 90, 2016.

BARROS, R. W. d. S. **Price setting and macroeconomic variables: Evidence from Brazilian CPI**. Tese (Doutorado) — Fundação Getúlio Vargas, 2009.

BILAL, A. The Geography of Unemployment*. **The Quarterly Journal of Economics**, v. 138, n. 3, p. 1507–1576, 03 2023. ISSN 0033-5533. Disponível em: <https://doi.org/10.1093/qje/qjad010>.

Bloomberg Línea Brasil. **iFood é avaliado em US$ 5,4 bilhões e se torna a startup mais valiosa do Brasil**. 2022. <https://www.bloomberglinea.com.br/2022/08/19/ifood-e-avaliado-em-us-54-bilhoes-e-se-torna-a-startup-mais-valiosa-do-brasil/>. Accessed: January 18, 2024.

BONHOMME, S. Chapter 5 - Econometric analysis of bipartite networks. In: GRAHAM, B.; de Paula Áureo (Ed.). **The Econometric Analysis of Network Data**. Academic Press, 2020. p. 83–121. ISBN 978-0-12-811771-2. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128117712000110>.

BOTELHO, L. V.; CARDOSO, L. d. O.; CANELLA, D. S. Covid-19 e ambiente alimentar digital no brasil: reflexões sobre a influência da pandemia no uso de aplicativos de delivery de comida. **Cadernos de Saúde Pública**, SciELO Brasil, v. 36, p. e00148020, 2020.

BRAMOULLé, Y.; DJEBBARI, H.; FORTIN, B. Peer effects in networks: A survey. **Annual Review of Economics**, Annual Reviews, v. 12, n. Volume 12, 2020, p. 603–629, 2020. ISSN 1941-1391. Disponível em: <https://www.annualreviews.org/content/journals/10.1146/annurev-economics-020320-033926>.

BRAUN, M.; VERDIER, V. Estimation of spillover effects with matched data or longitudinal network data. **Journal of Econometrics**, v. 233, n. 2, p. 689–714, 2023. ISSN 0304-4076. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304407621002827>.

CAVALLO, A. Online and official price indexes: Measuring Argentina's inflation. **Journal of Monetary Economics**, v. 60, n. 2, p. 152–165, 2013. ISSN 0304-3932. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304393212000967>.

CAVALLO, A. Are online and offline prices similar? evidence from large multi-channel retailers. **American Economic Review**, v. 107, n. 1, p. 283–303, January 2017. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/aer.20160542>.

CAVALLO, A. Scraped Data and Sticky Prices. **The Review of Economics and Statistics**, v. 100, n. 1, p. 105–119, 03 2018. ISSN 0034-6535. Disponível em: <https://doi.org/10.1162/REST_a_00652>.

CAVALLO, A.; CAVALLO, E.; RIGOBON, R. Prices and supply disruptions during natural disasters. **Review of Income and Wealth**, v. 60, n. S2, p. S449–S471, 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12141>.

CAVALLO, A. et al. Using online prices for measuring real consumption across countries. **AEA Papers and Proceedings**, v. 108, p. 483–87, May 2018. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/pandp.20181037>.

CAVALLO, A.; RIGOBON, R. **The Distribution of the Size of Price Changes**. [S.l.], 2011. (Working Paper Series, 16760). Disponível em: <http://www.nber.org/papers/w16760>.

CAVALLO, A.; RIGOBON, R. The billion prices project: Using online prices for measurement and research. **Journal of Economic Perspectives**, v. 30, n. 2, p. 151–78, May 2016. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/jep.30.2.151>.

Data Rio. **PCRJ::limite-de-bairros**. 2024. <https://www.data.rio/datasets/PCRJ::limite-de-bairros/about>.

Defesa Civil do Rio Grande do Sul. Defesa civil atualiza balanço das enchentes no rs – 10/6, 9h. 2024. Consulted on 10 June 2024. Disponível em: <https://www.estado.rs.gov.br/defesa-civil-atualiza-balanco-das-enchentes-no-rs-9-6-9h>.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: <https://arxiv.org/abs/1810.04805>.

EINAV, L.; LEVIN, J. The data revolution and economic analysis. **Innovation Policy and the Economy**, University of Chicago Press Chicago, IL, v. 14, n. 1, p. 1–24, 2014.

EINAV, L.; LEVIN, J. Economics in the age of big data. **Science**, v. 346, n. 6210, p. 1243089, 2014. Disponível em: <https://www.science.org/doi/abs/10.1126/science.1243089>.

GLAESER, E. L.; KIM, H.; LUCA, M. Nowcasting gentrification: Using yelp data to quantify neighborhood change. **AEA Papers and Proceedings**, v. 108, p. 77–82, May 2018. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/pandp.20181034>.

GREEN, M. et al. **Social Progress Index 2024**. Social Progress Imperative, 2024. Disponível em: <https://www.socialprogress.org/2024-social-progress-index/>.

HADDAD, E. et al. Impacto socioeconômico do ifood. 2023.

IBGE, C. d. I. d. P. **Sistema nacional de índices de preços ao consumidor: estruturas de ponderação a partir da pesquisa de orçamentos familiares: 2017-2018**. Rio de Janeiro: IBGE, Coordenação de Índices de Preços, 2020. v. 46. (Coleção Ibgeana; Relatórios metodológicos (IBGE), v. 46). ISSN 0101-2843. ISBN 9786587201023.

iFood. **Brasileiros Pedem Mais Delivery**. 2024. <https://institucional.ifood.com.br/noticias/brasileiros-pedem-mais-delivery/>. Accessed: February 28, 2024.

iFood para parceiros. ifood para parceiros. February 2024. Disponível em: <https://blog-parceiros.ifood.com.br/quem-somos/>.

INEP. **Nota técnica: Índice de Desenvolvimento da Educação Básica – Ideb**. [S.l.], 2007. Disponível em: <http://download.inep.gov.br/educacao_basica/portal_ideb/o_que_e_o_ideb/Nota_Tecnica_n1_concepcaoIDEB.pdf>.

INEP. **Nota informativa do Ideb 2021**. [S.l.], 2021. Disponível em: <https://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2021/nota_informativa_ideb_2021.pdf>.

JOHNSSON, I.; MOON, H. R. Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach. **The Review of Economics and Statistics**, v. 103, n. 2, p. 328–345, 05 2021. ISSN 0034-6535. Disponível em: <https://doi.org/10.1162/rest_a_00870>.

KOENKER, R.; MACHADO, J. A. Goodness of fit and related inference processes for quantile regression. **Journal of the american statistical association**, Taylor & Francis, v. 94, n. 448, p. 1296–1310, 1999.

KRAMARZ, F.; MARTIN, J.; MEJEAN, I. Volatility in the small and in the large: The lack of diversification in international trade. **Journal of International Economics**, v. 122, p. 103276, 2020. ISSN 0022-1996. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022199618301296>.

MACHADO, J. d. F. U. **New Challenges in Official Statistics: Big Data Analytics and Multi-level Product Classification of Web Scraped Data**. Tese (Dissertação) — Faculdade de Economia, Universidade do Porto, Porto, Portugal, 07 2023. Disponível em: <https://hdl.handle.net/10216/151290>.

Measurable AI. 2021 brazil food delivery: ifood continues to lead with 80September 7 2021. Disponível em: <https://blog.measurable.ai/2021/09/07/2021-brazil-food-delivery-ifood-continues-to-lead-with-80-market-share-rappi-ubereats/>.

OANCEA, B. Automatic product classification using supervised machine learning algorithms in price statistics. **Mathematics**, v. 11, n. 7, 2023. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/11/7/1588>.

PIGATTO, G. et al. Have you chosen your request? analysis of online food delivery companies in brazil. **British Food Journal**, Emerald Publishing Limited, v. 119, n. 3, p. 639–657, 2017.

POVO, C. do. Guaíba passa dos 5 metros e porto alegre amanhece com mais regiões atingidas pela enchente. 2024. Archived from the original on 7 May 2024. Retrieved 6 May 2024. Disponível em: <https://www.correiodopovo.com.br/not%C3%ADcias/cidades/gua%C3%ADba-passa-dos-5-metros-e-porto-alegre-amanhece-com-mais-regi%C3%B5es-atingidas-pela-enchente-1.1490965>.

POVO, C. do. Rs tem mais de 170 mortos pelas enchentes. 2024. Retrieved 1 June 2024. Disponível em: <https://www.correiodopovo.com.br/not%C3%ADcias/cidades/rs-tem-mais-de-170-mortos-pelas-enchentes-1.1499711>.

Prefeitura da Cidade do Rio de Janeiro. Lei complementar nº 270, de 16 de janeiro de 2024. 2024. Disponível em: <https://aplicnt.camara.rj.gov.br/APL/Legislativos/contlei.nsf/a99e317a9cfec383032568620071f5d2/0274835ddbc09b5303258aa700487674?OpenDocument>.

PULICI, A. et al. Índice de Progresso Social da Cidade do Rio de Janeiro. 2022.

RUIZ, L. D.; MCMAHON, S. D.; JASON, L. A. The role of neighborhood context and school climate in school-level academic achievement. **American Journal of Community Psychology**, v. 61, n. 3-4, p. 296–309, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ajcp.12234>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [S.l.: s.n.], 2020.

STACY, C. P.; HO, H.; PENDALL, R. Neighborhood-level economic activity and crime. **Journal of Urban Affairs**, n/a, n. n/a, 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/juaf.12314>.

STERN, S. et al. **Social Progress Index 2014**. [S.l.]: Social Progress Imperative, 2014.

TecMundo. Fim do delivery do Uber Eats: iFood tornou-se monopolista no Brasil. 2023. Disponível em: <https://www.tecmundo.com.br/mercado/231813-fim-delivery-do-uber-eats-ifood-tornou-monopolista-brasil.htm>.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

XIE, S.; HUBBARD, R. A.; HIMES, B. E. Neighborhood-level measures of socioeconomic status are more correlated with individual-level measures in urban areas compared with less urban areas. **Annals of Epidemiology**, v. 43, p. 37–43.e4, 2020. ISSN 1047-2797. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1047279719306088>.

ZANATTA, P.; RIGUE, A.; LAUXEN, N. Cidades do rio grande do sul continuam isoladas pelo 5º dia após chuvas devastadoras. **CNN Brasil**, 2024. Archived from the original on 6 May 2024. Retrieved 6 May 2024. Disponível em: <https://www.cnnbrasil.com.br/nacional/cidades-do-rio-grande-do-sul-continuam-isoladas-pelo-5o-dia-apos-chuvas-devastadoras/>.

ZHAO, Q.; HAUTAMAKI, V.; FRÄNTI, P. Knee point detection in bic for detecting the number of clusters. In: BLANC-TALON, J. et al. (Ed.). **Advanced Concepts for Intelligent Vision Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 664–673. ISBN 978-3-540-88458-3.

A.1
Appendix - Exploratory Analysis

The analysis of Rio de Janeiro data proposes discarding problematic RPs in order to avoid distortions. I also consider an additional pruning layer: dropping the outliers for each RP, i.e., the dots in Figure 2.6(a). As seen in Figure A.1, these outliers are reasonably influential. Dropping them also strengthens the relationship between the SPI and $\log\left(\bar{p}_i^{\text{geom}}\right)$, by raising R^2 by 34% in the absence of problematic RPs, and by 28% when all RPs are considered. The Pearson correlation coefficient ranges from 51% (no prune) to 76% (Excluding Centro, Barra da Tijuca, Méier and Bangu, and dropping outliers), a 51% variation.

As an alternative approach for handling these outliers, without explicitly discarding them, I consider analysing this correlation through the lens of Quantile Regression. Figure A.2 and Figure A.3 show that using all RPs to fit higher conditional quantiles partially recovers slope and measure of goodness of fit found when estimating a conditional median when the problematic RPs are discarded. Moreover, it suggests a higher correlation with prices for higher SPI levels. Note

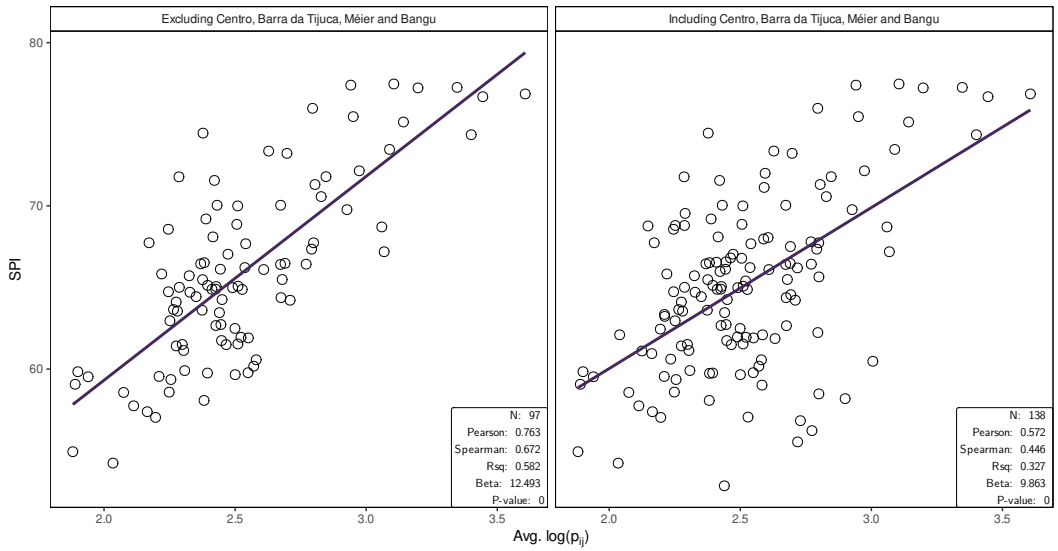


Figure A.1: Avg. $\log(p_{ij})$ $\times$ SPI - Excluding outliers

that R1 stands for Koenker e Machado (1999)'s local goodness of fit measure at the specified quantile.

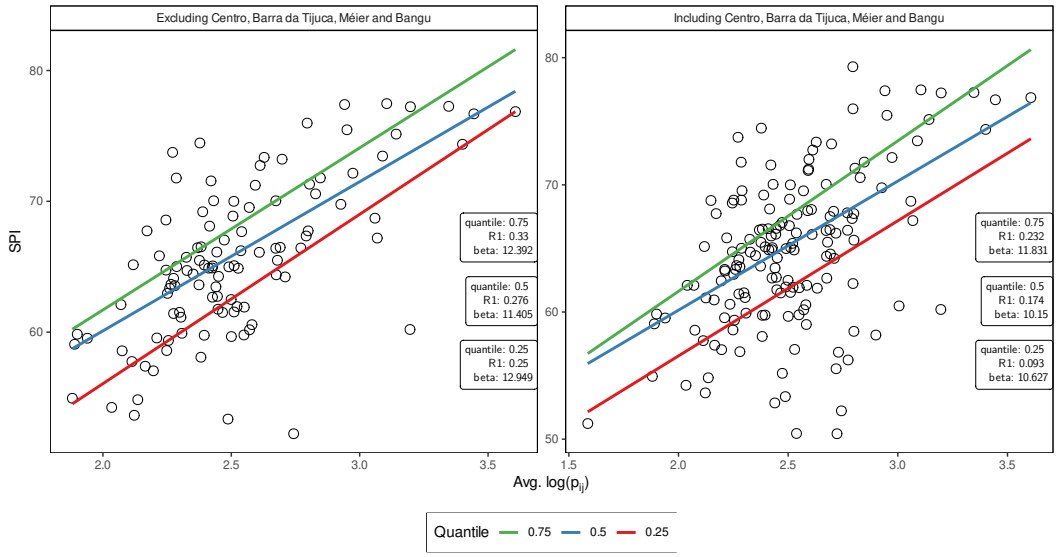


Figure A.2: Avg. $\log(p_{ij})$ $\times$ SPI - Quantile regression

At this point, the analysis was restricted to the SPI. However, data allows for analysing the subindexes that comprise the SPI. By checking Figure A.4, is verified that most of the explanatory power of iFood prices over the SPI is due to its connection with the Opportunities subindex. There is a non-negligible correlation for Opportunies even when problematic RPs are kept in sample. By the other hand, iFood prices do not capture variation in Basic Human Needs. This is an important

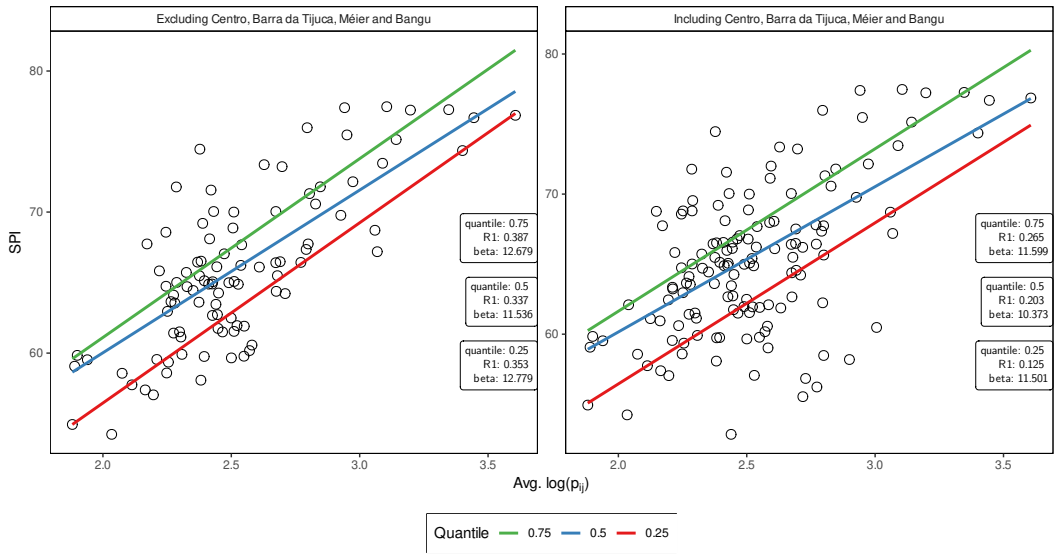


Figure A.3: Avg. log(p_{ij}) $\times$ SPI - Quantile regression - Excluding outliers

result from the policy formulation perspective, suggesting that these prices are likely inadequate for Epidemiology purposes, but might be useful in capturing economic activity, for example.

A.1.1 Seasonal variation

As mentioned earlier, results are based on data collected on a specific day. However, scraped data are subject to short run dynamics. By one hand, it allows for investigating these short run dynamics, what was unfeasible using traditional data collection methods. By the other hand, one could argue that the analysis performed along this section relies on data collected on a specific day, ignoring weekly seasonality, or even opening space for deliberately choosing a day to analyse seeking for convenient results.

In order to address this issue, and to obtain a picture of prices - and their relation to SPI - seasonal pattern along a week, I ran the scraper on daily frequency during a week, at the same hour. Results are shown below: first, I count the number of distinct sellers at each weekday, seeking to describe the operating days profile. Then, I investigate prices dynamics globally by computing the city's average log(p_{ij}). Finally, I compare the Pearson and the Spearman correlation coefficient, the R^2, slope coefficient and respective p-value across the week. Results

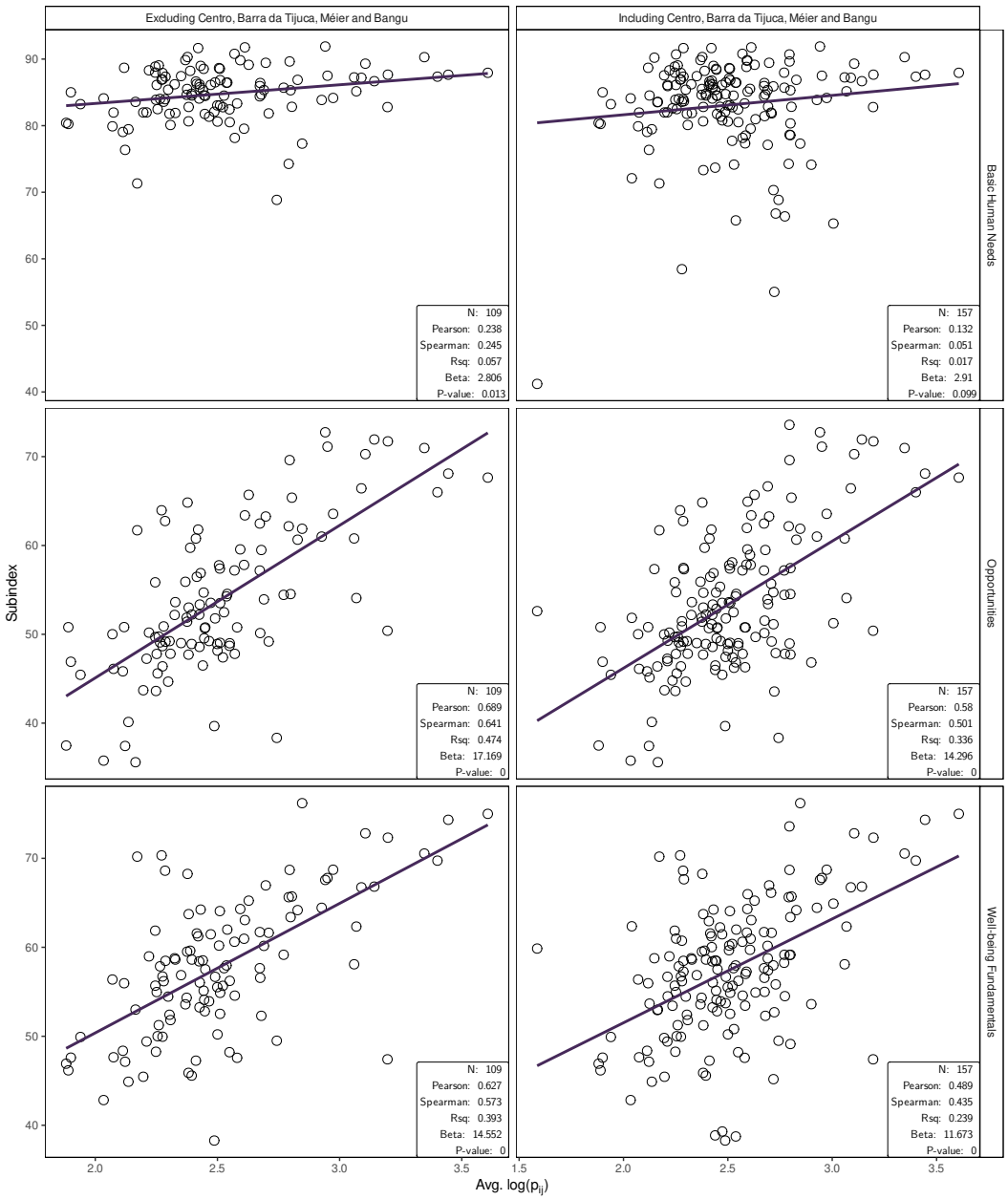


Figure A.4: Avg. $\log(p_{ij}) \times$ SPI - Subindex decomposition

are depicted on the following tables, with a vertical dashed line in the Friday, when the data afore analysed was collected.

Figure A.5 indicates that the day with most business out of operation is Sunday. The number of operating business is non-decreasing on the number of neighborhoods in sample, so the count for the full dataset is greater than or equal to the count for the pruned dataset in each weekday. Figure A.6 shows that the average price has a peak on the Thursday. Despite the lower prices on Sunday, most of variations have negligible scale.

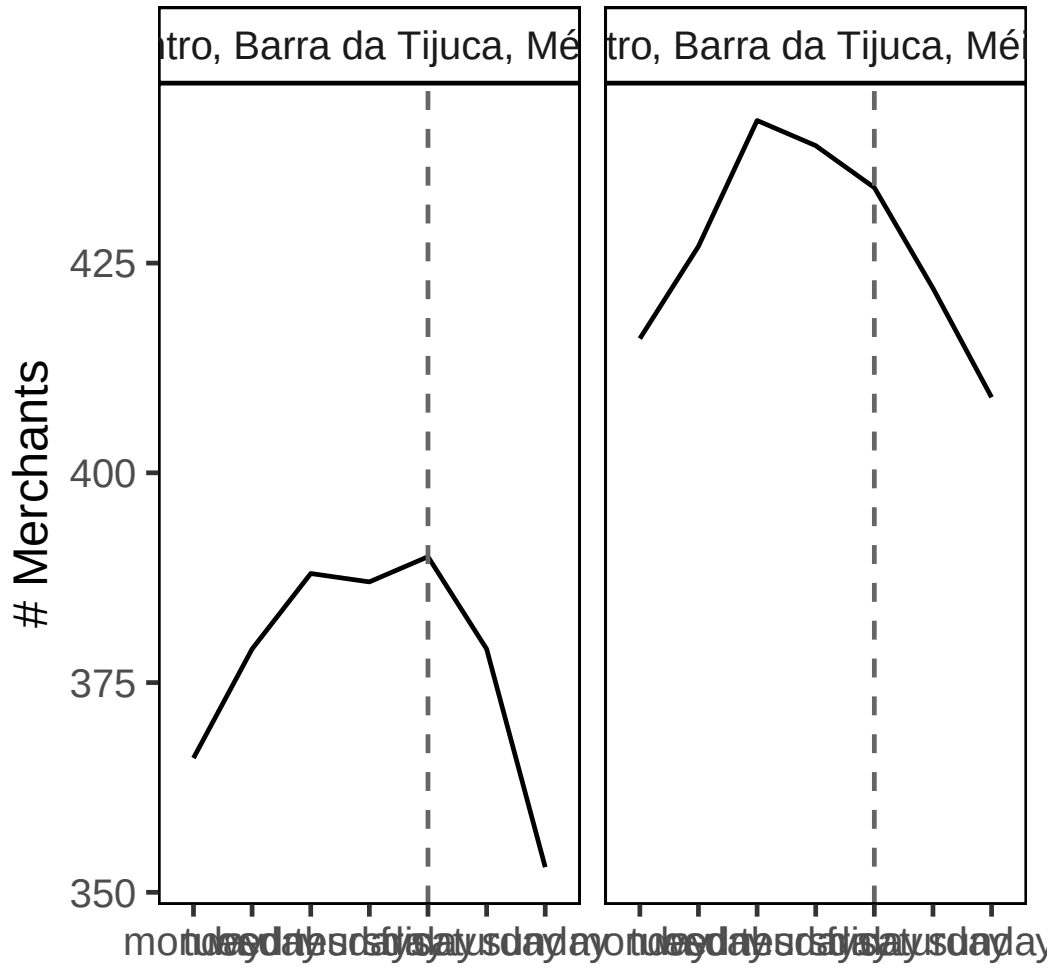


Figure A.5: Number of different merchants in the city by weekday

Like on Friday, the correlation measures are weaker when no pruning procedure is performed. Figure A.7, Figure A.8, Figure A.9 and Figure A.10 show that price connection with SPI is clearer on Thursday, and more latent on the Sunday. Overall, considering the pruned dataset, Friday seems to offer an intermediate view, and does not create any significant distortion on results. Figure A.11 shows no relevant variation when the plot scale is adjusted to cover the 0% to 1% (horizontal dashed line) range.

B.2
Appendix - High frequency data and Price Indexes

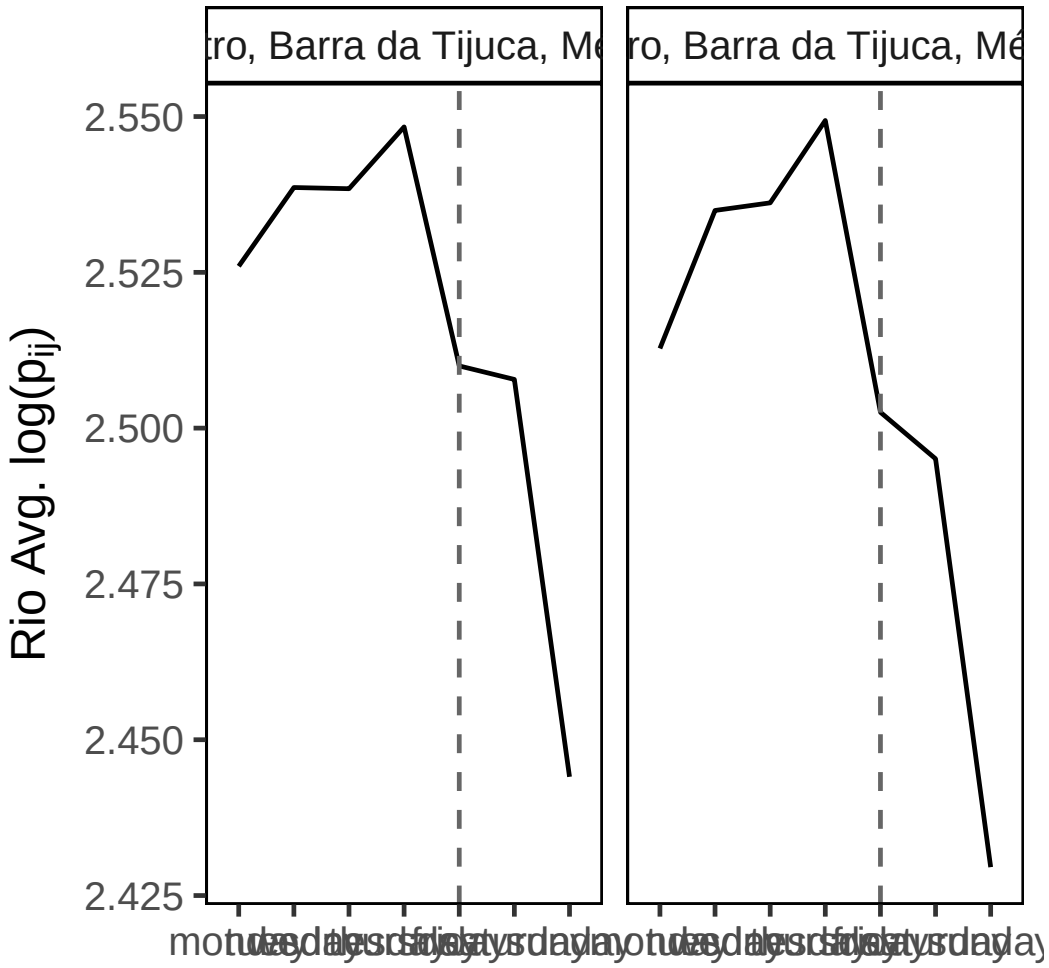


Figure A.6: Avg. log(p_{ij}) in the city by weekday

B.3
Appendix - Product Classification

The semi-supervised learning approach employed for ruling-out undesired entries is described by the following steps for each subitem α:

1. Labels were arbitrarily labelled, assigning a value 1 to labels which are indeed describing a product associated with this subitem;

2. Embedding vectors for a subsample $\{1, ..., T^*\}$ are clustered through Hierarchical Clustering;

3. Optimal number of clusters is chosen using BIC criterion following Zhao, Hautamaki e Fränti (2008) knee-point detection procedure;

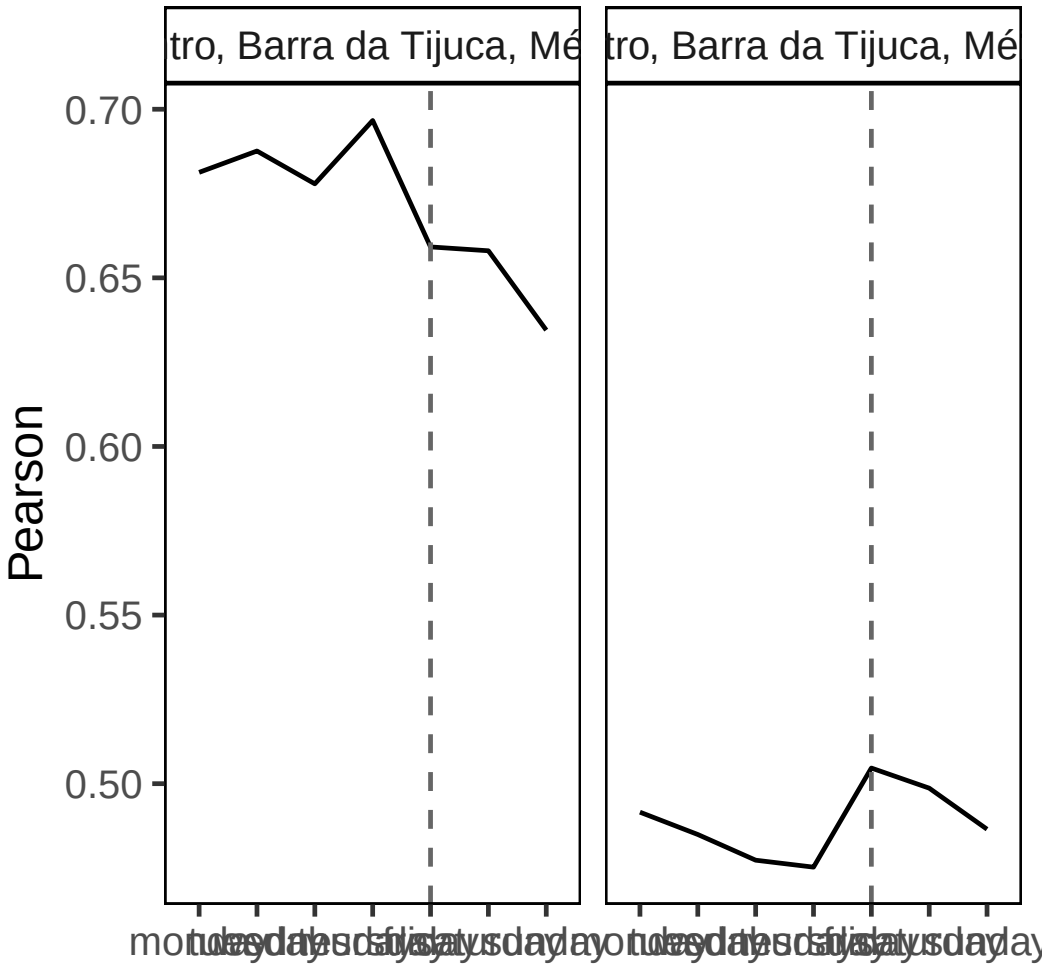


Figure A.7: Avg. $\log(p_{ij}) \times$ SPI - Pearson coefficient by weekday

4. A classification model selected by cross-validation[1] is trained for $\{1, ..., T^*\}$ and fitted for $\{T^* + 1, ..., T\}$;

5. Clusters containing labels with representative tags are kept and used for price index calculation, and these clusters are treated as products for price index calculation purposes.

Table B.2 shows the choice frequency for each classifier.

Embeddings are obtained through fine-tuning BERTimbau, a pre-trained BERT for Brazilian Portuguese (SOUZA; NOGUEIRA; LOTUFO, 2020). BERT

[1]Candidate models are Logistic Regression (LogisticRegression), Support Vector Classifier (SVC), Linear Support Vector Classifier (LinearSVC), K-nearest neighbor (KNeighbors), Decision Tree (DecisionTree), Random Forest (RandomForest) and Multilayer perceptron (MLPClassifier).

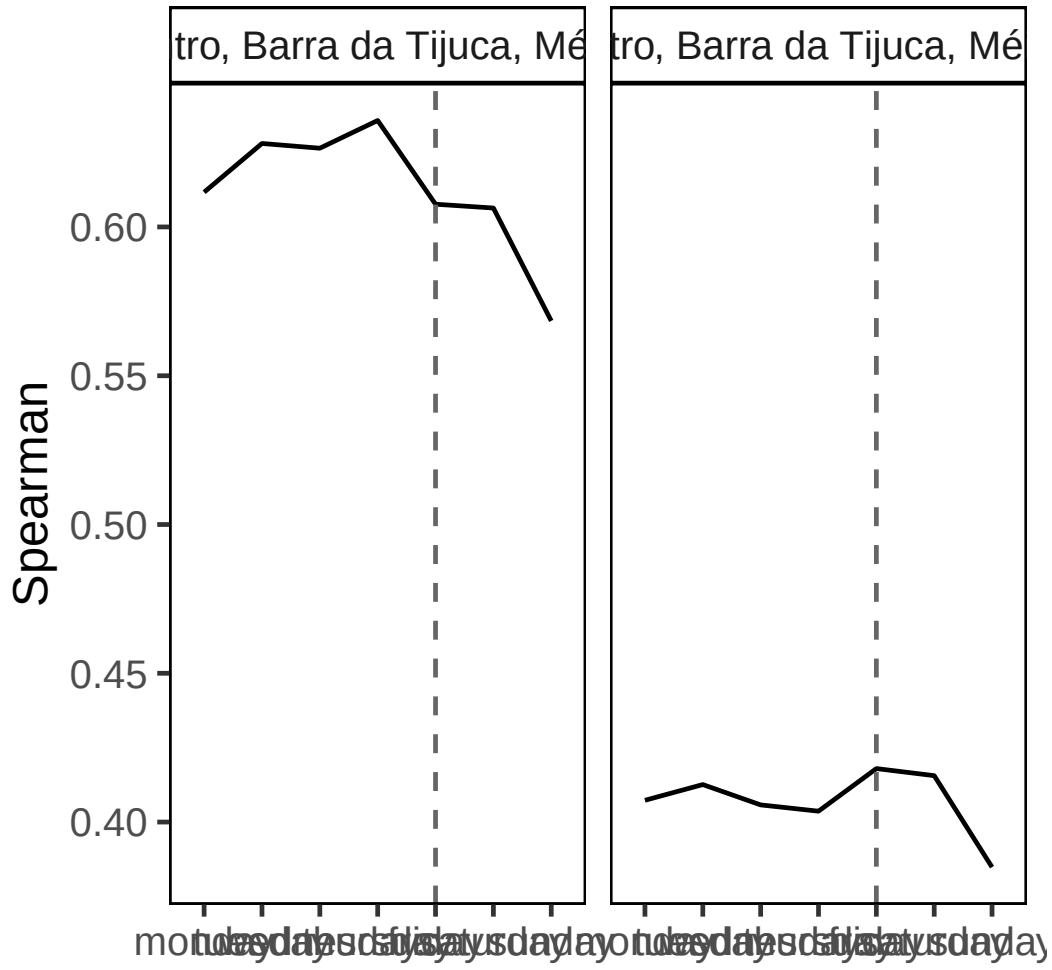


Figure A.8: Avg. $\log(p_{ij}) \times$ SPI - Spearman coefficient by weekday

(DEVLIN et al., 2019) stands for Bidirectional Encoder Representations from Transformers, which is a multi-layer bidirectional Transformer (VASWANI et al., 2017) encoder.

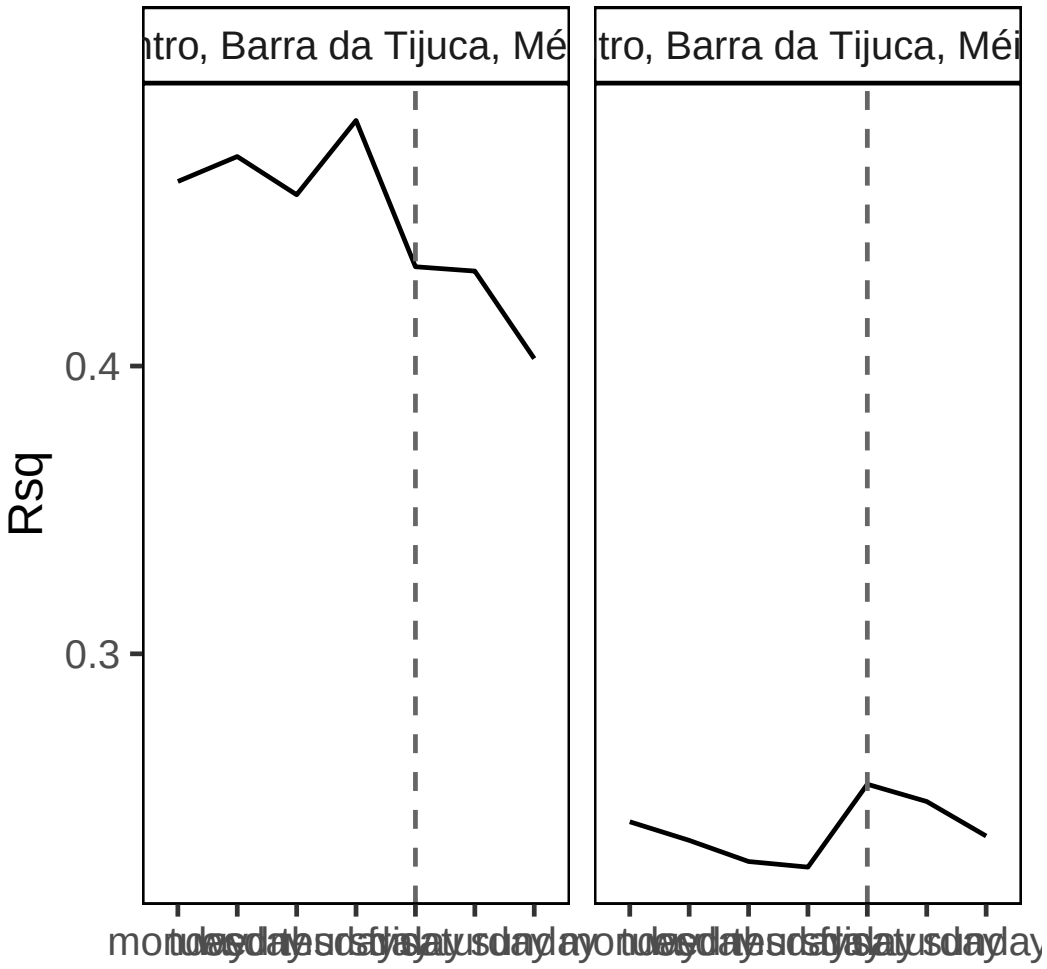


Figure A.9: Avg. log(p_{ij}) $\times$ SPI - R^2 by weekday

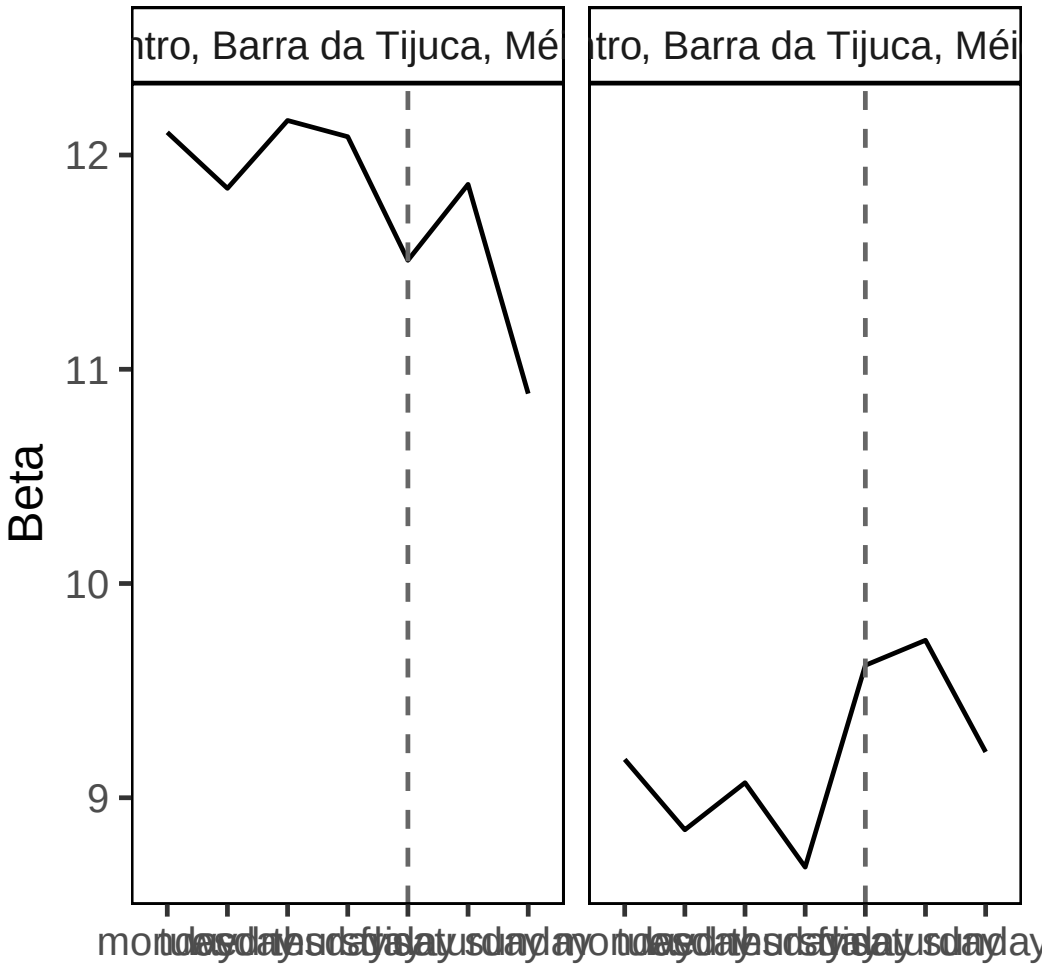


Figure A.10: Avg. $\log(p_{ij})$ $\times$ SPI - slope by weekday

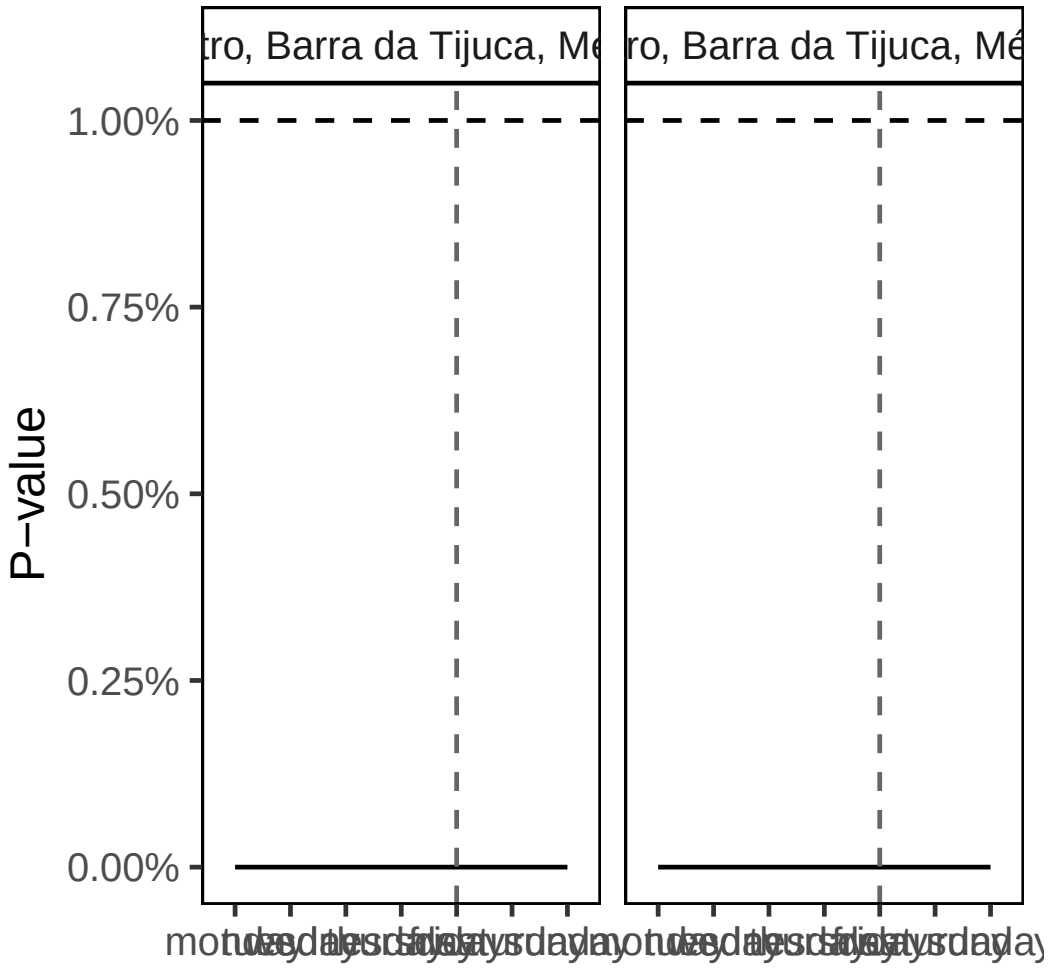


Figure A.11: Avg. $\log(p_{ij}) \times$ SPI - slope's p-value by weekday

Location	Code	Neighborhood	IPCA weight
Aracaju, SE	AJU	Atalaia	1.03
Belém, PA	BEL	Batista Campos	3.94
Belo Horizonte, MG	BH	Savassi	9.69
Campo Grande, MS	CG	Jardim dos Estados	1.57
Curitiba, PR	CUR	Batel	8.09
Brasília, DF	DF	Asa Sul	4.06
Fortaleza, CE	FOR	Aldeota	3.23
Goiânia, GO	GOI	Setor Bueno	4.17
Porto Alegre, RS	POA	Moinhos de Vento	8.61
Rio Branco, AC	RB	Bosque	0.51
Recife, PE	REC	Boa Viagem	3.92
Rio de Janeiro, RJ	RJ	Copacabana	9.43
Salvador, BA	SAL	Barra	5.99
São Luís, MA	SL	Renascença	1.62
São Paulo, SP	SP	Moema	32.28
Vitória, ES	VIT	Praia do Canto	1.86

Table B.1: Representative neighborhoods for CPI scrapping procedure and IPCA weights (%)

Model	Times chosen
LogisticRegression	40.76
MLPClassifier	24.84
SVC	19.75
LinearSVC	7.01
KNeighbors	5.73
DecisionTree	1.91

Table B.2: Chosen models by frequency, (%)