Carlos Alexandre Piccioni

Three Essays on Economics in Big Data Scenarios

Brasília

2024

Carlos Alexandre Piccioni

Three Essays on Economics in Big Data Scenarios

Doctoral Thesis presented to the Postgraduate Program in Economics of the Department of Economics at the University of Brasília, as a partial requirement for obtaining the degree of Doctor in Economics.

Universidade de Brasília - UnB

Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas - FACE

Departamento de Economia - ECO

Programa de Pós-Graduação

Supervisor Dr. Daniel Oliveira Cajueiro

Brasília

2024

Acknowledgements

I thank my parents, José Carlos and Shirley, my sisters Silmara and Elaine, my niece Patrícia and nephew Lucas, and Jandira, who was like a mother to me, for all the support and affection during this journey.

I thank my academic advisor, Daniel, for his unwavering patience, insightful guidance, constant availability, and the invaluable knowledge he imparted throughout this journey. Being his pupil has been a profound honor.

I thank Saulo for his invaluable support, guidance, and assistance in acquiring critical data essential for the execution of this work.

I thank José Luiz and Marina for all the support and motivation that were fundamental to the accomplishment of this work.

I thank Marcelo for all the help with obtaining financial data.

I thank my professors for imparting their extensive knowledge to me.

I thank my friends and colleagues who shared their time and attention with me during this journey.

I thank Lorena for her endless encouragement, support, and companionship, without which this work would not have been possible.

Abstract

This work comprises three studies on economics in big data contexts. The first analyzes the impact of ESG (Environmental, Social, and Governance) news on the stock returns of leading Brazilian companies, using an unprecedented Dictionary of ESG Terms specifically developed for this study to select and classify news according to the standards of the Sustainability Accounting Standards Board (SASB). The research indicates that only news with content that is financially material to investors influences stock returns. In other words, investors do not react for reputational or non-pecuniary reasons. The second study explores the high-frequency predictability of the Brazilian exchange rate (at the 1, 5, and 15-minute frequencies), employing both machine learning techniques and traditional linear regression for forecasting. Two types of exercises are conducted: one with contemporary predictors and another using out-of-sample data. We show that it is possible to beat the benchmark, the Random Walk, over a horizon of up to four minutes at a frequency of 1 minute. We also show that the most important predictors are those that carry local information, as well as the exchange rates of the BRICS or countries with economies similar to Brazil's. When the rates from B3's foreign exchange futures contracts are considered as predictors, we can beat the Random Walk over a horizon of up to 6 minutes. The third study measures consumption inequality at the municipal level using data from electronic payment methods, specifically data from credit card and Pix payments. Furthermore, as an application, we examine the relationship between inequality and economic complexity. We demonstrate that greater economic complexity is associated with lower consumption inequality, marking the first assessment of this kind for Brazilian municipalities.

Keywords: Big Data. Companies' Returns. Dictionary of ESG Terms. ESG News. Textual Analysis. Exchange Rates Forecasting. Machine Learning. Consumption Inequality. Electronic Payment Methods. Economic Complexity Index (ECI).

Resumo

Este trabalho compreende três estudos sobre economia em contextos de big data. O primeiro analisa o impacto das notícias de ESG (Environmental, Social, and Governance - Ambiental, Social e Governança) nos retornos das ações das principais empresas brasileiras, utilizando um Dicionário inédito de Termos ESG especificamente desenvolvido para este estudo para selecionar e classificar notícias de acordo com os padrões do Sustainability Accounting Standards Board (SASB). A pesquisa indica que apenas notícias com conteúdo financeiramente relevante para os investidores influenciam os retornos das ações. Em outras palavras, os investidores não reagem por motivos de reputação ou não pecuniários. O segundo estudo explora a previsibilidade em alta frequência da taxa de câmbio brasileira (nas frequências de 1, 5 e 15 minutos), empregando tanto técnicas de machine learning quanto regressão linear tradicional para previsão. São realizados dois tipos de exercícios: um com preditores contemporâneos e outro com dados fora da amostra. Mostramos que é possível superar o benchmark, o Random Walk, em um horizonte de até quatro minutos na frequência de 1 minuto. Também mostramos que os preditores mais importantes são aqueles que carregam informações locais, bem como as taxas de câmbio dos BRICS ou países com economias semelhantes à do Brasil. Quando as taxas dos contratos futuros do câmbio brasileiro da B3 são consideradas como preditores, conseguimos superar o Random Walk em um horizonte de até 6 minutos. O terceiro estudo mede a desigualdade de consumo no nível municipal usando dados de métodos de pagamento eletrônicos, especificamente dados de pagamentos com cartão de crédito e Pix. Além disso, como aplicação, examinamos a relação entre desigualdade e complexidade econômica. Demonstramos que uma maior complexidade econômica está associada a uma menor desigualdade de consumo, sendo esta a primeira avaliação deste tipo para municípios brasileiros.

Palavras-chaves: Big Data. Retorno das Companhias. Dicionário de Termos ESG. Notícias ESG. Análise Textual. Previsão de Taxas de Câmbio. Aprendizado de Máquina. Desigualdade de Consumo. Meios de Pagamentos Eletrônicos. Índice de Complexidade Econômica.

List of Figures

List of Tables

List of abbreviations and acronyms

AR	Abnormal Return
AuM	Assets under Management
B3	Brasil, Bolsa, Balcão (Brazilian Stock Exchange)
BRICS	Brazil, Russia, India, China, and South Africa
CBOE	Chicago Board Options Exchange
CAR	Cumulative Abnormal Returns
CE	Consumer Expenditure Survey
CNAE	Classificação Nacional das Atividades Econômicas (National Classification of Economic Activities)
CRECS	Chinese Residential Energy Consumption Survey
DI	Depósito Interfinanceiro (Interbank Deposit)
ECI	Economic Complexity Index
ESG	Environmental, Social, and Governance
FX	Foreign Exchange
GDP	Gross Domestic Product
IBGE	Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics)
IPEA	Instituto de Pesquisa Econômica Aplicada (Institute of Applied Economic Research)
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Machine Learning
MCSs	Model Confidence Sets
MDA	Mean Decrease in Accuracy
MIT	Massachusetts Institute of Technology

MSCI	Morgan Stanley Capital International
MSE	Mean Squared Error
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PWC	PricewaterhouseCoopers
PSID	Panel Study of Income Dynamics
RAIS	Relação Anual de Informações Sociais (Annual Social Information List)
RCA	Revealed Comparative Advantage
RECS	Residential Energy Consumption Survey
RF	Random Forest
RW	Random Walk
SASB	Sustainability Accounting Standards Board’s
SQL	Structured Query Language
SVR	Support Vector Regression
SVM	Support Vector Machines
USDAUD	Exchange rate of the US dollar to the Australian dollar
USDBRL	Exchange rate of the US dollar to the Brazilian real
USDCAD	Exchange rate of the US dollar to the Canadian dollar
USDCHF	Exchange rate of the US dollar to the Swiss franc
USDCZK	Exchange rate of the US dollar to the Czech koruna
USDEUR	Exchange rate of the US dollar to the euro
USDGBP	Exchange rate of the US dollar to the British pound sterling
USDILS	Exchange rate of the US dollar to the Israeli shekel
USDJPY	Exchange rate of the US dollar to the Japanese yen
USDMXN	Exchange rate of the US dollar to the Mexican peso
USDNOK	Exchange rate of the US dollar to the Norwegian krone

USDNZD	Exchange rate of the US dollar to the New Zealand dollar
USDPLN	Exchange rate of the US dollar to the Polish zloty
USDRUB	Exchange rate of the US dollar to the Russian ruble
USDSEK	Exchange rate of the US dollar to the Swedish krona
USDTHB	Exchange rate of the US dollar to the Thai baht
USDTRY	Exchange rate of the US dollar to the Turkish lira
USDZAR	Exchange rate of the US dollar to the South African rand
VIX	CBOE Volatility Index
WTI	West Texas Intermediate
XGB	Extreme Gradient Boosting

Contents

1 Introduction

This work consists of three papers on economics in big data scenarios. In the first paper, we assess the influence of ESG (Environmental, Social, and Governance) news on the stock returns of the premier Brazilian companies. Unlike earlier research that utilized data from commercial and opaque big data solutions, we developed a novel Dictionary of ESG Terms. This allows for the nuanced selection and categorization of ESG news based on the five dimensions and 26 categories delineated by the Sustainability Accounting Standards Board (SASB). In combination with a Sentiment Dictionary also created for this work, we filtered and classified ESG news from the Valor Econômico portal for the set of companies considered in this study. Using a firm-day panel, we study how positive and negative ESG news influences the stock returns of these firms. Additionally, we examine the role of financial materiality, as defined by SASB for each industry group, in influencing investor reactions. We investigate whether the dimension of the ESG topic of the news also matters. For example, do investors react the same way to news in the Environmental dimension as they do to news in the Leadership and Governance dimension?

In the second paper, we investigate the high-frequency predictability of the Brazilian exchange rate at 1, 5, and 15-minute intervals, using a mix of local and global variables as well as exchange rates from 17 other countries as predictors. For forecasting, we apply Machine Learning techniques such as Ridge, Lasso, Elastic Net, Random Forest, and Gradient Boosting, along with traditional linear regression. Our analysis includes two main exercises. The first, which we call the out-of-sample fit, employs contemporary data as predictors. In the second exercise, the out-of-sample forecasting, we attempt to predict the exchange rate at $t + h$ using information available up to time t. In both exercises, we estimate the model parameters in a rolling window for linear regression and Machine Learning methods, and perform hyperparameter tuning through cross-validation in time series with a grid search. We assess the models by comparing their Mean Squared Error (MSE) with the MSE of the benchmark, the Random Walk without drift, and use the Diebold and Mariano (1995) test to determine if the model forecasts and the Random Walk forecasts are statistically different. We also use the Model Confidence Set (MCS) procedure to determine the superior model set in each exercise, frequency, and considered horizon.

In the third paper, we measure consumption inequality at the municipal level through electronic payment methods, specifically credit cards and Pix. Traditionally, inequality assessments rely on sample surveys or census data — approaches that are resource-intensive, error-prone, and often underestimate inequality. Additionally, they are generally conducted with low frequency. For instance, the Brazilian Demographic Census, the sole source capable of providing the granular data needed for municipal-level inequality calculations, is conducted approximately once a decade. This low frequency can limit the effectiveness of proposing and evaluating public policies aimed at promoting equity. By leveraging individual payment data as a proxy for consumption, we intend to offer a new way to determine an inequality index at any given time. As a way of validating our inequality index, we compare it with the income inequality Gini index based on the IBGE Census. Going a step further, this paper proposes to examine the relationship between inequality and economic complexity. We investigate whether a higher level of economic complexity correlates with increased or decreased inequality. Existing literature primarily focuses on the relationship between inequality and economic complexity at the country level, with few studies evaluating subnational entities, and the results are still mixed. Thus, we believe we contribute to the debate by being the first study to carry out this evaluation at the municipal level in Brazil, using an unprecedented database to determine inequality.

2 The sustainability news that affects companies' returns

2.1 Introduction

Corporations and investors are increasingly considering Environmental, Social and Governance issues (ESG) in their business models (Gillan; Koch; Starks, 2021; Amel-Zadeh; Serafeim, 2018). A report by PwC[1] predicts that managers worldwide will increase their ESG-related assets under management (AuM) from US$18.4 trillion in 2021 to US$33.9 trillion by 2026. This amount is projected to represent 21.5% of total AuM[2]. Thus, companies are investing increasing amounts of resources to improve their performance on ESG issues, while regulators seek to comprehend how ESG information flows to the market and how capital market participants respond to such information (Serafeim; Yoon, 2022). In light of the potential impact of ESG factors on corporate market value, a relevant question arises: "To what extent do ESG-related news[3] items influence the market value of corporations, considering that market participants may use the stock market as a mechanism to incentivize or sanction companies based on their ESG performance?"

In this paper, we evaluate the impact of ESG news on the stock prices of the major Brazilian companies using our novel Dictionary of ESG Terms. First, we create a Dictionary of ESG Terms to enable the selection and classification of news across the five dimensions and 26 categories as defined by the Sustainability Accounting Standards Board (SASB), in addition to the three classic ESG dimensions. Second, we create a Sentiment Dictionary by assigning polarity to the terms in our Dictionary of ESG Terms, in addition

[1] PricewaterhouseCoopers is a global network of firms delivering assurance, tax, and consulting services. It is one of the largest professional services networks in the world, often referred to as one of the "Big Four" accounting firms, alongside Deloitte, EY, and KPMG. It provides insights and predictions on trends and investments in the ESG space.

[2] *"ESG-focused institutional investment seen soaring 84% to US$33.9 trillion in 2026, making up 21.5% of assets under management: PwC report"*. <https://www.pwc.com/gx/en/news-room/press-releases/2022/awm-revolution-2022-report.html>. Updated on 2022-10-10. Accessed on 2023-12-31.

[3] Extreme ESG events, like toxic substance leaks, racism scandals, and accounting fraud, have a negative effect on the market value of a company. This is well known in the academic literature (Capelle-Blancard; Desroziers; Scholtens, 2021). However, such studies have limitations: they focus on a few events and do not consider the ordinary daily operations of companies that may employ ESG practices, thereby ignoring investor reactions to such practices (Capelle-Blancard; Petit, 2019). Therefore, regular news can be a way to find out about companies' ESG practices and their consequences, not just about extreme events. Also, unlike other ESG information sources, such as company reports, certifications, or ESG ratings from specialized agencies, news allows for high-frequency analysis with low time delay. News is also less susceptible to "greenwashing" than company reports, which can be manipulated to please investors (Serafeim; Yoon, 2022). Furthermore, ESG ratings from different agencies lack standardization and are generally updated infrequently, resulting in discrepancies between agencies (Capelle-Blancard; Desroziers; Scholtens, 2021; Berg; Koelbel; Rigobon, 2022; Chatterji et al., 2016).

to those in the Harvard-IV Sentiment Dictionary[4]. Third, applying our Dictionary of ESG Terms, we select and classify ESG news articles from the Valor Econômico newspaper for the period 2014 to 2022. Fourth, using our Sentiment Dictionary, we identify sets of positive and negative news. Finally, we use a firm-day panel to verify the daily impact of news on the market value of companies, where the dependent variable is the companies' stock returns and the main independent variables are the positive and negative ESG news.

Studies providing a useful list of ESG-related words are nearly non-existent (Baier; Berninger; Kiesel, 2020), with those available focusing exclusively on the three classic ESG dimensions. There are no dictionaries providing more comprehensive and detailed classifications, such as the dimensions and categories defined by SASB. Therefore, our first contribution is proposing a dictionary that enables news selection and classification according to SASB's standards.

By using the SASB standard, we can separate news that is considered financially material from that which is not considered financially material. According to the SASB definition, "*information is financially material if omitting, misstating, or obscuring it could reasonably be expected to influence investment or lending decisions that users make on the basis of their assessments of short-, medium-, and long-term financial performance and enterprise value*". In other words, financial materiality involves identifying the ESG issues that are most relevant to the economic performance of a company and its investors. SASB has developed specific standards for different industry sectors to help companies determine which topics are materially relevant to them. Thus, unlike other studies that filter news only in the three classic ESG dimensions (Laplante; Lanoie, 1994; Klassen; McLaughlin, 1996; Dasgupta; Laplante; Mamingi, 2001; Flammer, 2013; Krüger, 2015; Capelle-Blancard; Petit, 2019), we can evaluate if investors react to all ESG news, even those considered financially immaterial, for reputational or non-pecuniary reasons[5]. This approach allows us to investigate whether the issue of financial materiality matters, potentially explaining the conflicting results in the literature (while some studies find a significant stock price response to ESG news in general, others find no such significance). Could the conflicting results in the literature be due to noise caused by news without material issues?

We demonstrate, for the Brazilian case, that stock prices react to ESG news. Unlike Krüger (2015) and Capelle-Blancard and Petit (2019), we found positive price reaction to positive news, and unlike Serafeim and Yoon (2022), we found negative price reaction

[4] The psychological Harvard-IV Dictionary, developed by Harvard University (Stone, 2002), has been prominently featured in various studies assessing sentiments in news articles (Kearney; Liu, 2014). Esteemed for its comprehensive vocabulary and precise categorization of words into emotional connotations—both positive and negative—it has become an indispensable tool for sentiment analysis. Its widespread adoption in both academic and market research underscores its effectiveness in extracting and quantifying sentiment, thereby providing valuable insights into the impact of news sentiment on investor behavior and market trends.

[5] Baker et al. (2018) shows that green bonds are issued at a premium compared to similar bonds, indicating that investors have a preference for non-pecuniary attributes.

to negative news using only one news source. We also find that the SASB concept of materiality is important: investors react only to financially material news, as classified by SASB. This means that they do not react merely to reputational or non-pecuniary issues. In addition, there are distinct reactions to the news based on the SASB dimension: for news in the Environment dimension, the reaction occurs for both positive and negative news, whereas for the Leadership & Governance dimension, a significant price reaction only occurs for negative news.

Recent studies investigating the relationship between ESG news and companies' market value on a daily basis use Big Data solutions for filtering, processing, and sentiment attribution to ESG news (Capelle-Blancard; Petit, 2019; Serafeim; Yoon, 2022). However, such solutions are proprietary and not publicly accessible, and therefore non-transparent. In contrast, our solution is open and straightforward, and can be used, supplemented, and modified by any researcher who wishes to select ESG news in any news database.

Existing research either focuses on other economies or only analyzes the largest global companies. To our knowledge, this is the first study conducted on the Brazilian economy, which is a crucial player in the global environmental landscape. Furthermore, Brazil's largest companies have been at the center of corruption scandals over the last decade, drawing attention to the importance of the SASB Leadership & Governance dimension.

The paper is structured as follows. In Section 2.2, we provide a brief introduction to the literature. In Section 2.3 we describe the database. In Section 2.4, we outline the creation process for the Dictionary of ESG Terms. In Section 2.5 we present the methodology used to verify the reaction of stock prices to ESG news. Sections 2.6 and 2.7 present and discuss our main findings. Finally, Section 2.8 concludes.

2.2 Literature review

The term ESG[6] was first used officially in 2004, in the UN Global Compact Initiative's "Who Cares Wins" report. This report was the result of a joint initiative of 20 financial institutions in response to a request by Kofi Annan, Secretary-General of the United Nations, to develop guidelines and recommendations on how to better integrate environmental, social, and governance issues into companies' business models (Gillan;

[6] The Environmental (E) dimension quantifies the effect of a company's actions on natural ecosystems. It includes everything from greenhouse gas emissions to making products that are better for the environment. The Social (S) dimension addresses the company's relationship with its workforce, consumers, and society as a whole, including everything from talent retention initiatives to the company's collaboration for the development of the communities in which it operates. Governance (G) refers to the mechanisms that ensure a company's management acts in its shareholders' best interests, such as its codes of conduct and business principles, as well as the implementation of policies to prevent illegal practices such as fraud and bribery (Liang; Renneboog, 2020).

Koch; Starks, 2021). Even before the term ESG was coined, studies were already seeking to establish a relationship between environmental news and the market value of companies, such as in Laplante and Lanoie (1994) and Klassen and McLaughlin (1996), using the Event Study methodology[7,8]. Subsequently, studies began to consider more than one ESG dimension, as demonstrated in Krüger (2015). The selection of ESG news moved from manual filtering in print newspapers, as in Laplante and Lanoie (1994), to searches in news databases based on keywords related to the ESG context, as in Klassen and McLaughlin (1996), Flammer (2013), Krüger (2015), Schmidt (2019), Ender and Brinckmann (2019). The determination of news sentiment was initially done manually, until the news databases themselves began to provide sentiment classification, as is the case with the KLD database (now part of MSCI) in Krüger (2015). In general, studies are conducted on hundreds of news items each, and Krüger (2015) is one of the first studies to break the thousand news items evaluated. Recent studies, such as Capelle-Blancard and Petit (2019) and Serafeim and Yoon (2022), use proprietary Big Data solutions for ESG news classification and sentiment determination. These solutions allow for the evaluation of more than 10,000 to 100,000 news items per study, using natural language processing and machine learning techniques to determine news sentiment from news sources. The Covalence EthicalQuote and TrueValue Labs solutions, used in Capelle-Blancard and Petit (2019) and Serafeim and Yoon (2022) respectively, are examples of such solutions. However, these solutions are closed and proprietary, making them non-transparent.

The initial conclusions are that positive and negative ESG news generate positive and negative impacts on companies' stock returns, respectively (Klassen; McLaughlin, 1996; Flammer, 2013; Dasgupta; Laplante; Mamingi, 2001). However, Krüger (2015) introduces a divergent view regarding positive news: it has a negative effect on companies' market value, providing evidence for an agency problem in relation to ESG practices[9]. On the other hand, Capelle-Blancard and Petit (2019) reach a divergent conclusion regarding positive news: it does not have an impact on companies' market value, while negative news has a negative impact. Ender and Brinckmann (2019), in turn, does not find a negative price reaction to negative ESG events. Vincentiis (2023) finds divergent results depending on the region studied (United States, Europe, and Asia-Pacific), suggesting that ESG news is interpreted differently in different geographical areas.

[7] While our study uses news as a source of information on companies' ESG practices, there is a range of studies that use company reports (Kaspereit; Lopatta, 2016; Mervelskemper; Streit, 2017) and ESG ratings from specialized agencies (Miralles-Quirós; Miralles-Quirós; Gonçalves, 2018; Yoon; Lee; Byun, 2018; Glück; Hübel; Scholz, 2021) to analyze the impact of ESG practices on firms' market value.

[8] Although the Event Study methodology is predominant in this topic, some studies use other techniques, such as constructing "good" and "bad" ESG portfolios using ESG controversy measures and comparing the performance of both portfolios to investigate the impact of ESG performance on financial performance of companies, as in studies like Franco (2020), Dorfleitner, Kreuzer and Sparrer (2020).

[9] A situation where company managers adopt ESG practices to gain a good reputation with some of the key stakeholders, such as politicians, labor unions, media, at the expense of shareholders. Thus, shareholders would receive positive news about ESG practices negatively.

Serafeim and Yoon (2022) innovates by using news classification according to the SASB standard[10], provided by TrueValue Labs. The Sustainability Accounting Standards Board is a non-profit organization that establishes specific standards for ESG reporting by companies, which identifies twenty-six categories of ESG issues organized along five major dimensions: Environment, Social Capital, Human Capital, Business Model & Innovation, and Leadership & Governance (Busco et al., 2020)[11]. The SASB standard specifies, for each of 77 industrial sectors, which of its ESG categories are financially material, i.e. must be relevant to investors' decision-making[12,13] (Jebe, 2019). Serafeim and Yoon (2022) shows that there is no significant relationship between ESG news and stock returns when only considering news classified as not financially material: such a relationship would only exist for news classified as material by SASB.

Our work is most similar to that of Serafeim and Yoon (2022), as we aim to classify news according to SASB standards and investigate whether financial materiality matters. However, we distinguish ourselves by providing a simpler and more open solution for news classification and sentiment determination. Additionally, we analyze the theme for the Brazilian scenario. We also show that with our solution, significant results can be found using only one news source, unlike Serafeim and Yoon (2022).

To the best of our knowledge, our dictionary is the first to classify news along the SASB dimensions and categories, besides the three classic ESG categories. The only works that provide dictionaries for filtering ESG topics are those prepared by Mýšková and Hájek (2018) and Baier, Berninger and Kiesel (2020). We differ from these works by allowing classification according to the SASB dimensions and categories. Moreover, we focus on ESG news, whereas these dictionaries focus on annual reports from companies[14]. We hope that our ESG Dictionary can be utilized and enhanced in research that requires selecting and classifying ESG news, whether according to the SASB standard or not.

[10] The concept of materiality according to SASB standards is also explored in other studies on other sources of information, such as company reports in Grewal, Hauptmann and Serafeim (2021).

[11] The appendix A presents the 26 categories organized in these 5 dimensions. See <https://www.sasb.org/standards/process/> for more information on the process of developing SASB standards and identifying which topics are relevant to each industry.

[12] Greenhouse Gas Emissions, for instance, is a SASB category considered financially material for the Oil and Gas Exploration industry, whereas Customer Welfare category is not. This category is financially material to the Processed Foods industry, whereas the Greenhouse Gas Emissions category is not.

[13] Khan, Serafeim and Yoon (2016) is probably the first work that seeks to distinguish material issues from immaterial issues in ESG, in this case, in relation to investments. The authors demonstrate that investments in material sustainability issues can enhance shareholder value, while investments in immaterial sustainability issues have little to no positive or negative value implications.

[14] Guo et al. (2020) uses ESG vocabulary to filter ESG news to predict stock volatility. However, the author does not disclose the specific ESG Dictionary used and does not utilize SASB classification.

2.3 Data

We used the news from the Valor Econômico portal from 2014 to 2022, specifically the Companies section, as our news source. We applied our ESG Dictionary to the news (we demonstrate the process of creating the dictionary in section 2.4.1), extracting ESG-related news and classifying it according to the SASB categories. Then, we extracted from these the news pertaining to the target companies[15]: those comprising the Ibovespa[16] index of B3 and B3's two most prominent ESG indices, the Corporate Sustainability Index - ISE[17] and the Carbon Efficient Index - ICO2[18].

We exclude all news that mentions the stock exchange so that we do not have news that may report what happened to a particular stock on a given day or in prior days[19]. We also eliminate possibly duplicate news[20]. We only retain news where companies are mentioned in the first two paragraphs, similar to Tetlock, Saar-Tsechansky and Macskassy (2008). Thus, we find at least one piece of news in the entire sample for 95 of the 109 companies considered.

Figure 1 presents the percentage of news from each SASB dimension. The categories of Leadership & Governance and Environment accounted for 57% and 21%, respectively, of the ESG news, and the remaining dimensions each accounted for up to 8%. Figure 2 presents the year-on-year percentage evolution of each dimension, and explains the high number of news items classified in Leadership & Governance: we can observe the peaks of news in this dimension between 2014 and 2018. These were the years when the country closely followed the developments of the Lava-Jato operation, which primarily investigated corruption cases involving Brazil's largest company, Petrobras.

[15] To filter news by company name, we compile a list of terms to identify them in the news. We face the third problem described in Section 2.4.1: many company names are also present in proper names. A classic example is the company "Vale", one of the largest in the country. Without proper treatment, news containing the term "Vale do Silício" (Silicon Valley) is classified as referring to the company "Vale". Thus, we expand the list of proper names that we exclude from the classification process by including proper names that contain company names.

[16] The most important index on the B3 stock exchange in Brazil. The stocks comprising the index account for approximately 80% of B3's financial volume. For more information, see <https://www.b3.com.br/en_us/market-data-and-indices/indices/broad-indices/ibovespa.htm>

[17] Corporate Sustainability Index. Its methodology aims to include businesses that are committed to corporate sustainability. More details are available at <http://iseb3.com.br>.

[18] Bovespa Carbon Efficient Index. Its methodology aims to include businesses that demonstrate a commitment to carbon emissions transparency and the transition to a low-carbon economy. More information at <https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-de-sustentabilidade/indice-carbono-efficient-ico2-b3.htm>

[19] To explain stock returns, we consider news as independent variables in the regression detailed in Section 2.5. However, some news may be dependent on returns, for example, if they describe what happened to a company's stocks on a given day due to an ESG event. To avoid news that is dependent on stock returns, we exclude news with mentions of the stock market.

[20] News referencing the same company and category, with the same number of mentions to the company and category, and the same number of terms, number of negative terms, and number of positive terms in paragraphs containing mentions of the ESG topic or the company.

Figure 1 – ESG news items by SASB dimension.

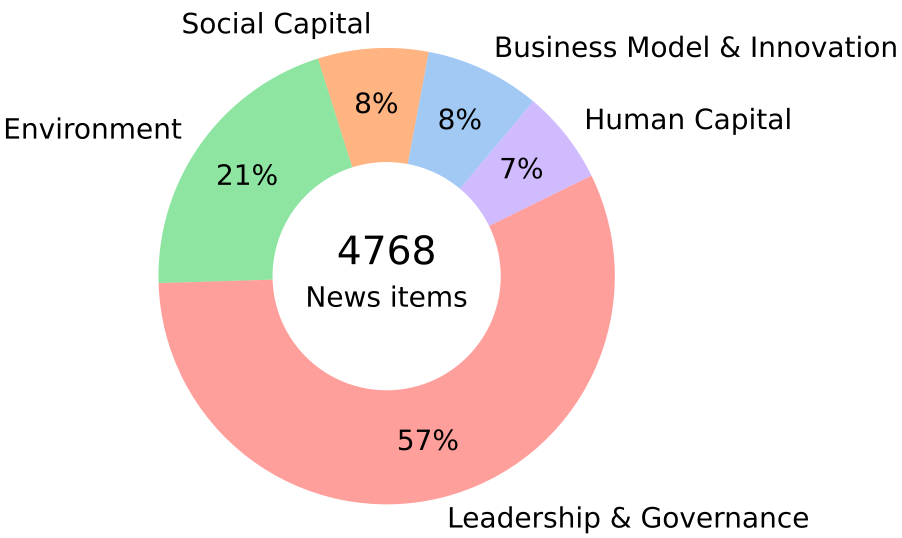


Figure 2 – ESG news items by SASB dimension / year.

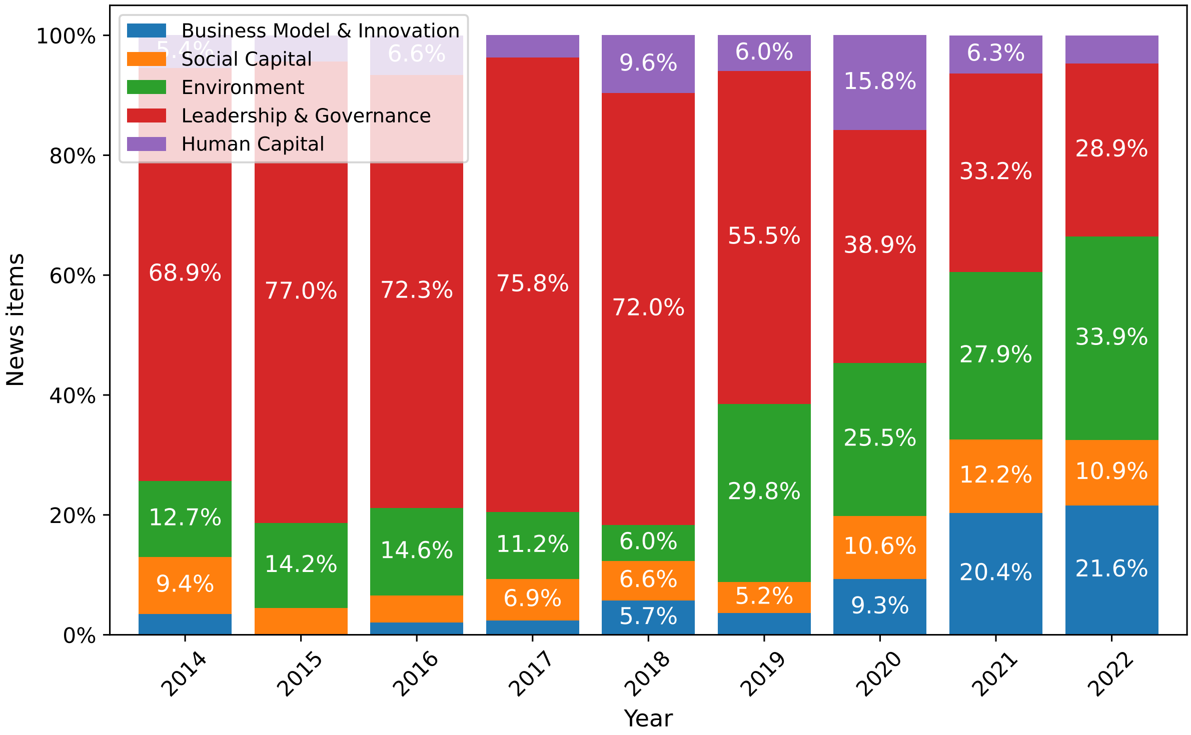


Figure 3 shows the number of news articles for each of the 26 SASB categories. This figure shows a high volume of news articles related to corruption in the Business Ethics category. Furthermore, Figure 4 presents terms related to the ESG categories (word frequency), where we see the predominance of corruption-related terms such as corruption itself, bribe and fraud.

Figure 3 – ESG news items by SASB category.

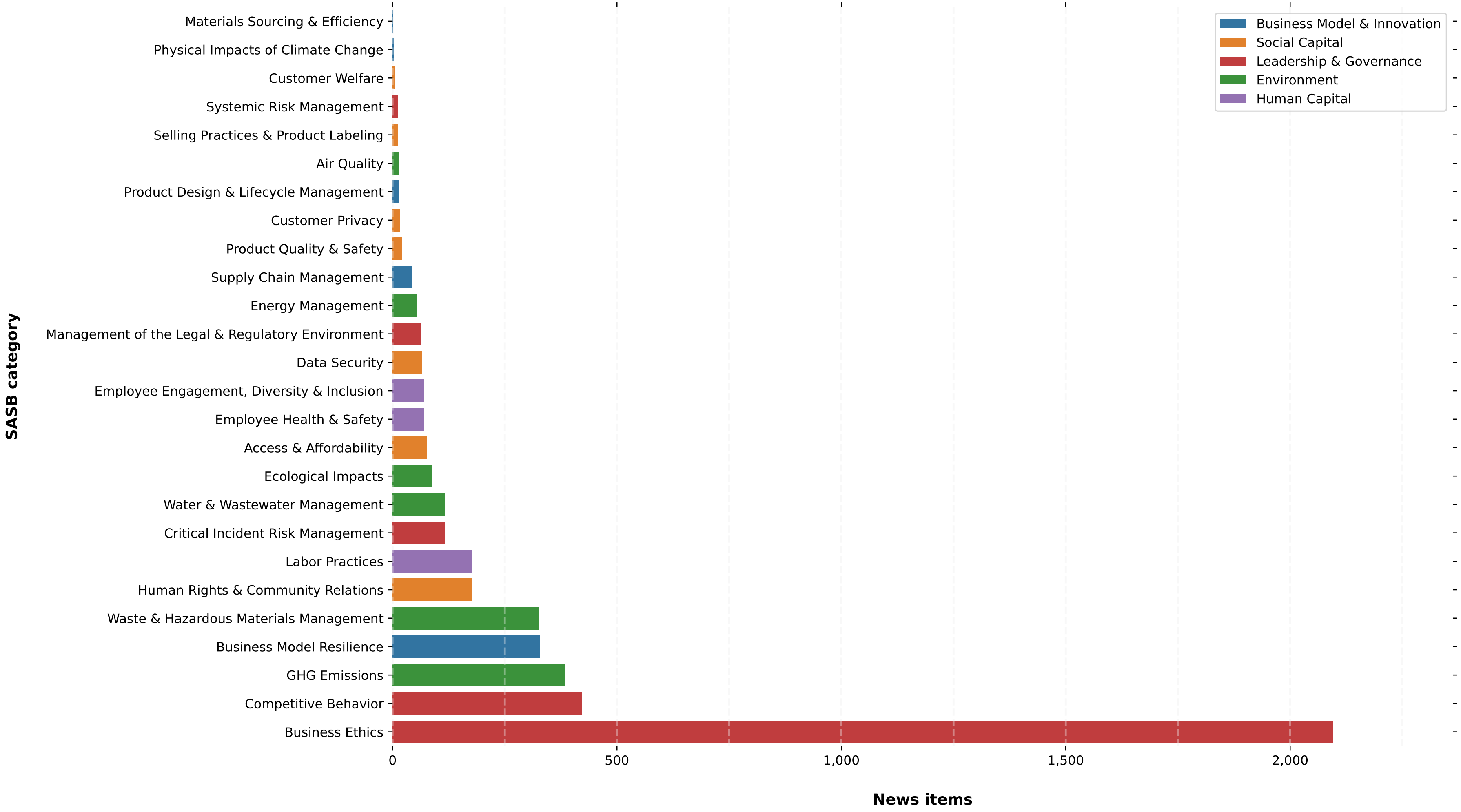
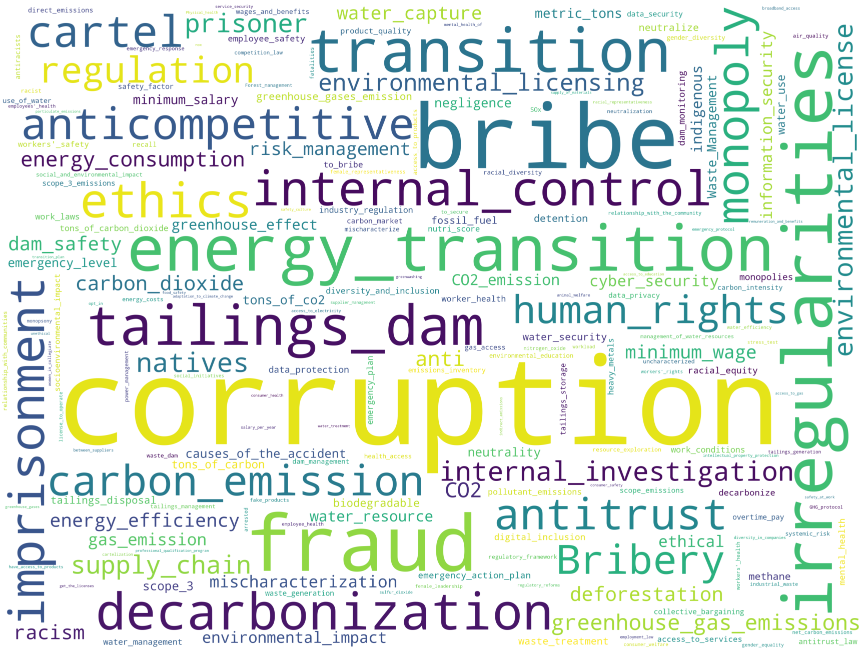


Figure 4 – ESG news word cloud (translated from the original Portuguese).



Regarding the financial data of the companies, we obtained the closing prices of the stocks from Bloomberg.

2.4 Dictionaries Construction

This section covers the development of the Dictionary of ESG Terms and the Sentiment Dictionary.

2.4.1 Dictionary of ESG Terms

We create the Dictionary of ESG Terms through an iterative process. First, we generate a list of terms for each SASB category based on our interpretation of the generic categories description provided by SASB[21] (each term also receives a classification in the classic ESG three dimensions). For example, the term 'Carbon Emissions' is associated with the 'GHG Emissions' category of SASB (and also with 'Environment' in the classic classification of the three ESG dimensions, Environment, Social, and Governance).

Second, based on the evaluation of a sample of ESG news articles from Valor Econômico, we augment our dictionary with a list of generic ESG terms that do not have a specific classification in SASB categories. We use a generic category for these terms. They are important for selecting ESG news items that may refer to SASB categories for which we do not yet have specific terms in our dictionary. By evaluating the news items after their selection, it is possible to extract the specific terms from these categories. For

[21] Available at <https://www.sasb.org/standards/materiality-finder/>

example, the term ‘ESG’ itself is a generic term that will appear in news items about ESG, but this term alone is not associated with a specific SASB category. However, this term is useful for filtering ESG news items, and from the individual analysis of each filtered news item, we can identify which SASB category it pertains to. After identifying the news items category, we check which specific terms should be included in our dictionary so that, in the filtering process, we can associate the news items with the SASB category to which they belong.

Third, we filter approximately 500 news items that contain at least one term from our initial dictionary and evaluate the performance of the classification of the news items into SASB categories, according to our judgment of the SASB classification, along with the three traditional ESG dimensions. We evaluate news by news, which terms we may add to the dictionary for each category/dimension, which ones we should modify, and which we should omit. Then, we update the dictionary and repeat the procedure twice, each time evaluating approximately 150 news items. Thus, in total, we evaluate a sample of around 800 news items to compose the dictionary[22].

Creating and designing a dictionary that covers ESG terms is a challenging endeavor. One problem is dealing with terms that can be related to ESG, but are generic enough to not be related to ESG in certain news articles. For example, the term ‘neutrality’ can refer to neutrality in terms of greenhouse gas emissions, but is not specific enough to only appear in ESG-related news (such as news about fiscal neutrality in taxation). Another example is the term ‘salary’, which can appear in news articles unrelated to ESG (such as the effect of high inflation on salaries) or refer to the implementation of fair wage policies, falling under the Labor Practices SASB category.

The second problem is dealing with terms that are specific to the ESG context but, without treatment, can be classified under more than one SASB category. For instance, the term ‘pollute’ can refer to both air and water pollution, indicating that it can fall under at least two distinct SASB categories.

To address these problems, we have two possible solutions that can be used individually or in combination. The first solution uses what we call conditional terms: terms that must appear in the same paragraph as the term being evaluated so that we classify the news into a specific category. For example, in order to classify a piece of news with the term ‘neutrality’ as news about Greenhouse Gas Emissions, we require that the same paragraph contain terms such as carbon, CO_2 and carbon dioxide.

The second solution is to use compound terms instead of single terms. For instance, we use the term ‘fair wage’ and some of its variations instead of the single term ‘salary’. Or, to identify news items that deal with the mental health of employees, we can use

[22] The proposed final ESG Dictionary is available at <https://doi.org/10.5281/zenodo.7789973>.

the compound term 'mental health' together with the conditional terms 'employees' and 'workers' to classify the news item in the SASB Employee Health & Safety category.

A third problem is terms that belong to the ESG context but are also used in proper names. For instance, the term "CH_4" (methane) falls under the SASB Greenhouse Gas Emissions category. However, there is a Brazilian company called CH_4 Energia, and without proper treatment, we misclassify news about this company as belonging to a SASB category. As a solution, we create a list of proper names that we ignore in the process of classifying news into ESG/SASB categories.

2.4.2 Sentiment Dictionary

After defining the Dictionary of ESG Terms, we assign a positive, negative or neutral polarity to each term based on our experience evaluating ESG news. Then, we attach Portuguese translations of positive and negative terms from the Harvard IV-4 dictionary. Due to potential nuances lost in translation, we manually evaluate each term and suggest changes to conform to the ESG context when we deem it necessary. In our dictionary, 2,876 terms are categorized as positive and 2,856 as negative.

2.5 Methods

In this section, we present the methods used to calculate abnormal returns of company stocks, how we define the set of positive and negative news, and how we assemble the data panel that will be used to determine if ESG news affects the market value of companies.

2.5.1 Calculation of abnormal returns

We estimate the abnormal return AR_t^i (or market-adjusted return) on day t for each company i as:

$$AR_t^i \equiv R_t^i - R_t^m,$$

where R_t^i is the log return on the stock company i in t, that is, $R_t^i = \log P_t^i - \log P_{t-1}^i$, with P_t^i being the company i stock closing price in t, and R_t^m the log return of the Ibovespa index.

Consider that a news item is published at time t. The impact of an ESG news on stock returns may not occur only at t. If the market has access to the information before the news is published, the market's response may occur before t. Alternatively, even if the news is timely, the market may respond only in the days following the news

Table 1 – Summary statistics of the number of positive and negative terms per news item (all observations).

	News	Min	q_1	Median	Mean	q_3	Max	St. Dev
$Positive\ Terms_j$	4768	0	5	10	12.95	17	106	10.92
$Negative\ Terms_j$	4768	0	3	7	9.22	13	86	8.39
$Total\ Terms_j$	4768	9	77	122	150.90	199	1015	105.36
S_j	4768	-0.286	-0.022	0.021	0.021	0.065	0.364	0.068

$Positive\ Terms$ ($Negative\ Terms$) is the number of positive (negative) terms per news item, in paragraphs that mention the company or ESG terms. $Total\ Terms$ is the number of terms in paragraphs that mention the company or ESG terms. S is the news sentiment index, calculated according to equation (2.1).

publication, meaning that the stock response may occur after t. Thus, for news at time t, we also evaluate market-adjusted returns at $t-1$ and $t+1$, and the cumulative abnormal returns from $t-5$ to $t-2$ and from $t+2$ to $t+5$ ($CAR^i_{t-5:t-2}$ and $CAR^i_{t+2:t+5}$, where $CAR^i_{t+a:t+b} = \sum_{p=t+a}^{t+b} AR^i_p$, $b > a$). Therefore, we can evaluate the returns in these windows around the release date t of the news.

Similar to Krüger (2015), we exclude stocks worth less than R$2.00 on the last sample date to avoid extreme abnormal returns.

2.5.2 Definition of news polarity

Using terms from the Sentiment Dictionary, we calculate the total number of positive and negative terms for each ESG news article. For this, we only consider the paragraphs that contain ESG terms, such as Schmidt (2019), or mentions of the company in question. We do this to capture the sentiment surrounding the ESG topic and the specific company, while excluding any irrelevant noise from other parts of the news article that may not be related to the ESG topic.

We calculate the following sentiment index for each news story j, where $Total\ Terms_j$ refers to the total number of terms present in paragraphs that mention either the ESG topic or the company:

$$S_j = \frac{Positive\ Terms_j - Negative\ Terms_j}{Total\ Terms_j} \tag{2.1}$$

The Table 1 provides descriptive statistics for the number of positive, negative, and total number of terms in paragraphs containing ESG terms or company mentions in each news item j, as well S_j.

As shown in the Eq. (2.2), we standardize S_j by subtracting it from its mean and dividing it by its standard deviation.

$$s_j = \frac{S_j - \mu_{S_j}}{\sigma_{S_j}} \tag{2.2}$$

Similar to Tetlock, Saar-Tsechansky and Macskassy (2008), we define positive news as news in the top quartile of the s_j distribution and negative news as in the bottom quartile.

We know the exact minute of publication for each news item. We update what we call "news reference date" in accordance with stock exchange closing times. If the stock exchange closes at 5:00 p.m. on day t, then we consider news published after that time as news of day $t+1$. We use the same method on weekends and holidays. If news comes out after the stock market closes on a Friday or the day before a holiday, the reference date will be the next business day of the stock exchange.

We then group the news by the reference date. We consider a date to be a positive (or negative) news date if it contains one or more positive (or negative) news stories. When there is at least one positive and one negative news item for the same company on the same date, we exclude the news (since the expected direction for stock returns would be ambiguous). This gives us the final set of news considered positive and negative for each SASB dimension, as shown in Figure 5. We can see that the Business Model & Innovation dimension has a very small number of negative news items. This is due to the fact that news about design and product lifecycle management, business model resilience, supply chain management, material sourcing and efficiency tend to be overwhelmingly positive, while news about physical impact of climate change in the Brazilian context is still almost nonexistent. Concerning the Environment, Human Capital and Social Capital dimensions, we classify approximately 3/4 of the news items as positive. In contrast, the Leadership & Governance dimension contains 70% of negative news, given the number of corruption-related stories.

Table 2 presents the news set, positive and negative, by industrial group. We can see that the statistics are in line with our previous conclusion: the lowest percentage of positive news is in the Extractives & Mineral Processing group, given the negative news regarding the Lava-Jato operation related to Petrobras. Table 3 presents the descriptive statistics for the variables $CAR^i_{t-5:t-2}$, AR^i_{t-1}, AR^i_t, AR^i_{t+1}, $CAR^i_{t+2:t+5}$ relating to final news set.

2.5.3 Data Panel

To estimate the reaction of companies' stock prices to ESG news, we use the same methodology employed by Serafeim and Yoon (2023), through panel regression of the Eq. (2.3).

Figure 5 – ESG news items polarity by SASB dimension.

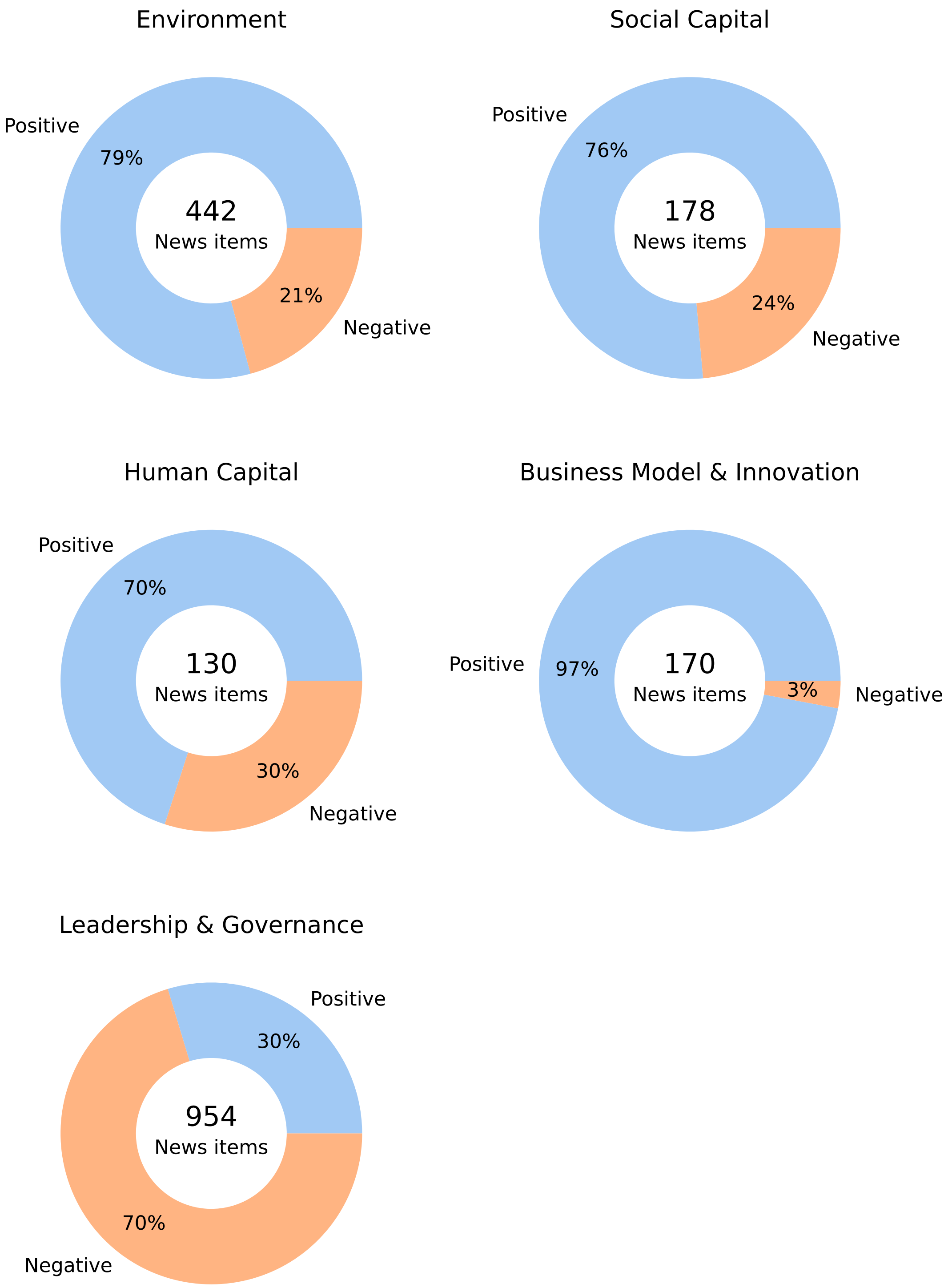


Table 2 – ESG news items polarity by SASB industry group.

SASB industry group	**Positive**	**Negative**	**% Positive**
Consumer Goods	62	21	74.7%
Extractives & Mineral Processing	376	490	43.4%
Financials	79	60	56.8%
Food & Beverage	49	47	51.0%
Health Care	28	15	65.1%
Infraestructure	176	93	65.4%
Renewable Resources & Alternative Energy	57	7	89.1%
Resource Transformation	84	63	57.1%
Services	20	10	66.7%
Technology & Communications	54	22	71.1%
Transportation	40	21	65.6%

Number of positive and negative news items by industrial group according to SASB classification.

Table 3 – *AR* and *CAR* summary statistics (ESG positive and negative news).

	N	**Min**	q_1	**Median**	**Mean**	q_3	**Max**	**St. Dev**	**#N/A**
$CAR^i_{t-5:t-2}$	1873	-0.2813	-0.0227	-0.0005	-0.0008	0.0210	0.4064	0.0438	1
AR^i_{t-1}	1873	-0.3605	-0.0131	-0.0011	-0.0006	0.0123	0.2323	0.0301	1
AR^i_t	1874	-0.3605	-0.0139	-0.0014	-0.0005	0.0118	0.3809	0.0310	0
AR^i_{t+1}	1874	-0.3605	-0.0127	-0.0007	0.0000	0.0128	0.1275	0.0255	0
$CAR^i_{t+2:t+5}$	1871	-0.3098	-0.0212	0.0000	0.0009	0.0236	0.2870	0.0441	3

AR^i_t is the abnormal return or market-adjusted return, that is, the difference between the (log) return on the stock's closing price and the (log) return on the B3 Ibovespa index. AR^i_{t-1} (AR^i_{t+1}) is the abnormal return for the day before (after) the release of a given news item. $CAR^i_{t-5:t-2}$ ($CAR^i_{t+2:t+5}$) is the cumulative abnormal return from $t-5$ to $t-2$ ($t+2$ to $t+5$) in relation to the date t of news item's publication.

$$\begin{aligned} AR^i_t = \beta_1\ NegativeNews^i_t + \beta_2\ PositiveNews^i_t \\ + Industry\ FE + Date\ FE + \varepsilon^i_t \end{aligned} \tag{2.3}$$

Our dependent variable, AR^i_t, is the abnormal return (market-adjusted return) for company i at t. $Negative\ News^i_t$ and $Positive\ News^i_t$ are our main independent variables, which indicate whether on day t company i had positive or negative ESG news. $Industry\ FE$ and $Date\ FE$ are fixed effects for industry and date. For industry fixed effects, we consider classification into SASB industry groups.

Consider that news about company i is published at time t. If the market has access to the information before the news publication, any stock reaction, if there is one, may be observed before t. Alternatively, even when the news is published timely, the market's reaction may occur after the news publication date at t. Therefore, to capture the full impact of ESG news on stock returns, we also evaluate market-adjusted returns at $t-1$ and $t+1$, and from $t-5$ to $t-2$ and $t+2$ to $t+5$, that is, we consider AR^i_{t-1}, AR^i_{t+1}, $CAR^i_{t-5:t-2}$ and $CAR^i_{t+2:t+5}$ replacing AR^i_t in the Eq. (2.3).

2.6 Results

Table 4 displays the regressions results of the Eq. (2.3) for the set of positive and negative news. Positive and negative price reactions are statistically significant in response to positive and negative ESG news, respectively. In other words, we found evidence that investors care about ESG issues, both positively and negatively. On average, the price reaction to positive news is 21 basis points on the release date, 53 basis points in the 2-day window between the day before the release of the news ($t-1$) and the day of the release (t), and 76 basis points in the window from $t-5$ to t. As for negative news, the average price reaction on the date of release is minus 25 basis points, and minus 70 basis points in the two-day window between the previous day and the date of release. This result differs from Krüger (2015), who observes a negative price reaction to positive news, and Capelle-Blancard and Petit (2019), who finds no price reaction to positive news. It also differs from Serafeim and Yoon (2022), as we observe a negative price reaction to negative news using only one news source.

Table 4 – Panel regressions for all ESG news.

	CAR	AR	AR	AR	CAR
	t-5:t-2	t-1	t	t+1	t+2:t+5
Positive News	0.0023*	0.0032**	0.0021*	-0.0004	0.0011
	(2.2170)	(2.8595)	(2.3854)	(-0.4463)	(0.8141)
Negative News	-0.0020	-0.0045***	-0.0025**	0.0014	0.0022
	(-1.2134)	(-3.6093)	(-2.8105)	(1.6558)	(1.1070)

The table presents the estimated coefficients in five panel regressions of the equation (2.3), where AR (or CAR) is the abnormal return (or cumulative abnormal return) of a given company (company stock log return minus the reference index log return), for the periods from $t-5$ to $t-2$, $t-1$, t, $t+1$, and from $t+2$ to $t+5$, where t is the day the news was published. Positive (Negative) News takes the value 1 when there is positive (negative) news item for the company in t. 1025 positive and 849 negative observations are used. All models consider fixed effects of time and at the industrial group level according to the SASB classification. Standard errors are clustered at the firm level and date. ***, **, and * denote significance at 0.1%, 1% and 5% levels respectively.

We explore one advantage of our Dictionary of ESG Terms: the ability to classify news into SASB's 26 categories and thus distinguish which news is related to topics considered financially material by SASB and which is not, for each industry group, and to determine whether there are differences in the stock returns of the companies. Then, similarly to Serafeim and Yoon (2022), we rerun the panel regression of the Eq. (2.3) for the news set containing material topics and for the news set excluding material topics. Table 5 shows the results. There are statistically significant price reactions to both positive and negative news in the news set containing only material topics. The average price reaction to positive news is 28 basis points on the release date and 69 basis points between $t-1$ and t. As for negative news, the average price reaction on news release date t is minus 31 basis points, minus 76 points over the two-day period from $t-1$ to t, and minus 110 basis points over the five-day period from $t-5$ to t. However, as shown in Panel B of Table 5, there is no statistically significant price reaction to news that does not contain topics deemed material by SASB, regardless of whether the news is positive or negative. This result is consistent with that obtained by Serafeim and Yoon (2022), indicating that the materiality of ESG issues, according to the SASB classification, matters in the investors' decision-making. Therefore, we find evidence that investors do not respond to all ESG news; that is, the results do not support the hypothesis that investors responded to ESG issues for reputational or non-pecuniary reasons.

Table 5 – Panel regressions for ESG news with Material Topics and for ESG news without Material Topics.

Panel A: Only ESG news with material topics

	CAR	AR	AR	AR	CAR
	t-5:t-2	t-1	t	t+1	t+2:t+5
Positive News	0.0026	0.0041*	0.0028*	-0.0012	0.0018
	(1.5057)	(2.2305)	(2.2509)	(-0.8421)	(0.7936)
Negative News	-0.0044***	-0.0045***	-0.0031***	0.0007	0.0006
	(-3.4306)	(-4.0116)	(-3.4464)	(1.2427)	(0.3561)

Panel B: ESG news excluding those with material topics

	CAR	AR	AR	AR	CAR
	t-5:t-2	t-1	t	t+1	t+2:t+5
Positive News	0.0017	0.0023	0.0018	0.0004	0.0011
	(1.2455)	(1.6807)	(1.3410)	(0.2738)	(0.5495)
Negative News	0.0015	-0.0041	-0.0015	0.0020	0.0058
	(0.5604)	(-1.4819)	(-0.7323)	(1.1722)	(1.7558)

Panel A only considers news items whose topics are classified as material by SASB, for the respective industry (520 positive events and 559 negative events). Panel B considers all news except news whose topics are classified as material by the SASB (524 positive events and 301 negative events). The table presents the estimated coefficients in five panel regressions of the equation (2.3), where AR (or CAR) is the abnormal return (or cumulative abnormal return) of a given company (company stock log return minus the reference index log return), for the periods from $t-5$ to $t-2$, $t-1$, t, $t+1$, and from $t+2$ to $t+5$, where t is the day the news was published. Positive (Negative) News takes the value 1 when there is positive (negative) news item for the company in t. All models consider fixed effects of time and at the industrial group level according to the SASB classification. Standard errors are clustered at the firm level and date. ***, **, and * denote significance at 0.1%, 1% and 5% levels respectively.

We perform another exercise in which we divide the news by SASB dimension and run separately a regression for each dimension. Panels A and B of Tables 6 and 7 present the results for the Environment and Leadership & Governance dimensions, taking into account all news of the dimension and only news with material topics. It is worth noting that the average price reaction for the Environment dimension is similar to the previous result: prices react significantly positively and negatively to positive and negative news, respectively. In absolute terms, however, the coefficients are on average greater than when we evaluate all dimensions collectively.

Table 6 – Panel regressions for ESG news by SASB dimension (all news).

Panel A: News from the Environment dimension

	CAR t-5:t-2	AR t-1	AR t	AR t+1	CAR t+2:t+5
Positive News	0.0060**	0.0046**	0.0032**	-0.0017	0.0007
	(2.7889)	(2.6876)	(2.6028)	(-1.1428)	(0.3663)
Negative News	0.0012	-0.0042*	-0.0077***	0.0029	0.0024
	(0.1399)	(-2.124)	(-3.3091)	(1.4425)	(1.2288)

Panel B: News from the Leadership & Governance dimension

	CAR t-5:t-2	AR t-1	AR t	AR t+1	CAR t+2:t+5
Positive News	0.0020	0.0025	0.0021	0.0000	0.0000
	(1.3362)	(1.3520)	(1.0094)	(-0.0281)	(-0.0417)
Negative News	-0.0041*	-0.0050**	-0.0023**	0.0016	0.0024
	(-2.3362)	(-3.2435)	(-2.7710)	(1.8467)	(0.8323)

Panel C: Social Capital, Business Model & Innovation and Human Capital

	CAR t-5:t-2	AR t-1	AR t	AR t+1	CAR t+2:t+5
Positive News	0.0003	0.0026	0.0022	0.0004	0.0027
	(0.1837)	(1.5221)	(1.7943)	(0.3344)	(1.2850)
Negative News	-0.0009	-0.0007	-0.0017	-0.0035	-0.0009
	(-0.1734)	(-0.3425)	(-0.6584)	(-1.8101)	(-0.1672)

News by SASB dimension. Panel A considers 366 positive events and 102 negative events, Panel B 289 positive events and 699 negative events, and Panel C 414 positive events and 95 negative events. The table presents the estimated coefficients in five panel regressions of the equation (2.3), where AR (or CAR) is the abnormal return (or cumulative abnormal return) of a given company (company stock log return minus the reference index log return), for the periods from $t-5$ to $t-2$, $t-1$, t, $t+1$, and from $t+2$ to $t+5$, where t is the day the news was published. Positive (Negative) News takes the value 1 when there is positive (negative) news item for the company in t. All models consider fixed effects of time and at the industrial group level according to the SASB classification. Standard errors are clustered at the firm level and date. ***, **, and * denote significance at 0.1%, 1% and 5% levels respectively.

As for the Leadership & Governance dimension, an interesting result emerges: only negative news causes a price reaction. Positive news does not result in a statistically significant price increase. As Panel B of the Table 7 demonstrates, this result holds true when considering only news with material topics. It is interesting to note that the Business Ethics category dominates this dimension, as shown in the Figure 3, which captured a significant number of corruption-related news stories, particularly those pertaining to the Lava-Jato operation. Thus, there is evidence that investors respond negatively to corruption-related news, but not positively to positive news in this dimension of Governance.

As shown in Figure 1, Social Capital, Business Model & Innovation, and Human

Table 7 – Panel regressions for ESG news with Material Topics, by SASB dimension.

Panel A: News from the Environment dimension

	CAR	AR	AR	AR	CAR
	t-5:t-2	t-1	t	t+1	t+2:t+5
Positive News	0.0070*	0.0056*	0.0035*	-0.0010	0.0026
	(2.3339)	(2.5180)	(2.4854)	(-0.5496)	(0.9650)
Negative News	-0.0033	-0.0046*	-0.0080***	0.0023	0.0012
	(-0.6607)	(-2.0503)	(-3.6343)	(1.5524)	(0.8675)

Panel B: News from the Leadership & Governance dimension

	CAR	AR	AR	AR	CAR
	t-5:t-2	t-1	t	t+1	t+2:t+5
Positive News	-0.0021	0.0025	-0.0007	0.0003	-0.0030
	(-0.7952)	(0.8731)	(-0.4066)	(0.1136)	(-0.9768)
Negative News	-0.0065***	-0.0048**	-0.0029***	0.0009	0.0002
	(-6.1394)	(-3.0520)	(-3.3646)	(1.4733)	(0.0770)

Panel C: Social Capital, Business Model & Innovation and Human Capital

	CAR	AR	AR	AR	CAR
	t-5:t-2	t-1	t	t+1	t+2:t+5
Positive News	0.0022	0.0034	0.0053*	-0.0026	0.0024
	(0.8200)	(1.3515)	(2.1655)	(-1.6668)	(0.5963)
Negative News	-0.0071	-0.0036	-0.0034	-0.0044*	-0.0044
	(-0.6139)	(-0.9545)	(-1.2620)	(-2.1701)	(-0.5895)

News with material topics, by SASB dimension. Panel A considers 244 positive events and 94 negative events, Panel B 121 positive events and 454 negative events, and Panel C 183 positive events and 50 negative events. The table presents the estimated coefficients in five panel regressions of the equation (2.3), where AR (or CAR) is the abnormal return (or cumulative abnormal return) of a given company (company stock log return minus the reference index log return), for the periods from $t-5$ to $t-2$, $t-1$, t, $t+1$, and from $t+2$ to $t+5$, where t is the day the news was published. Positive (Negative) News takes the value 1 when there is positive (negative) news item for the company in t. All models consider fixed effects of time and at the industrial group level according to the SASB classification. Standard errors are clustered at the firm level and date. ***, **, and * denote significance at 0.1%, 1% and 5% levels respectively.

Capital do not account for more than eight percent each of the total news. Due to the limited number of observations, we grouped news from these three dimensions for a new regression, which is presented in Panel C of Tables 6 for all news and Table 7 for news with material topics. We find statistically significant price reactions for this grouping only when considering the set of news with material topics: in this case, on the date t for positive news and on the date $t+1$ for negative news.

2.7 Discussion

We discuss in this section our methodological choices (Section 2.7.1) and the limitations of our work (Section 2.7.2).

2.7.1 Our methodological choices

We choose the word list (dictionary) approach to filter news due to its simplicity and transparency compared to other methods (although it is laborious to construct and verify). It can be easily complemented for further research if needed, and as our results indicate, this method is effective in classifying ESG news.

To construct the Sentiment Dictionary, we attach the Harvard-IV dictionary to our dictionary, given that it is a generic dictionary. We believe it is a more conservative approach than using a finance-specific dictionary, since the ESG context can be different from finance in general. We also addressed this issue by manually selecting words that we deemed most appropriate for our purpose.

We conduct this research on Brazil due to its global relevance in environmental issues and the wide range of unresolved social issues within the country. Additionally, in the Governance dimension, the country has faced probably the biggest corruption case ever reported, as widely documented by both local and international press.

We selected Valor Econômico as the news source for this study due to its leadership position as the largest economy, finance, and business newspaper in Brazil[23], extensively covering ESG topics in its reporting. Therefore, we believe that the information published in this news source is widely regarded and absorbed by the market. In fact, many other digital media sources in the country often replicate what they consider to be the most significant news from Valor Econômico. However, in our case, the inclusion of this type of media would not have an impact, as we use dummies in our regression for cases where there are positive or negative news on a given day, unless we filter for a minimum number of news sources that have published news about a particular company and ESG topic on the same day, as in Serafeim and Yoon (2022). This strategy could help filter news considered most relevant by the general media itself. Nevertheless, such approach may introduce some bias in news selection, potentially favoring certain companies or specific types of news, such as negative developments on a particular ESG topic.

2.7.2 Limitations of our Study

Our dictionary was originally created in Portuguese, based on the evaluation of news in Portuguese. We have made available an English version of the dictionary, however, it has not been tested in languages other than Portuguese. To adapt it to other languages, some terms may need to be changed or added, depending on the language-specific nuances of ESG terminology. Nonetheless, we believe that using our dictionary as a starting point can significantly reduce the cost and effort required to build a similar dictionary from scratch.

[23] <https://valorinternational.globo.com/about-valor/>

Another limitation of our study is that our Dictionary of ESG Terms is based on our understanding of each SASB category's generic description. Consequently, the dictionary is generic as opposed to industry-specific. Also, the decision of which terms should be related to which SASB classification depends on the researcher's evaluation and understanding of ESG news and SASB classification. A suggestion for future work is to compare our approach with another methodology of news classification, like a machine learning algorithm trained on a large corpus of news articles.

It is important to note that our analysis focuses on the impact on the market value of firms in the very short term, and we do not address the question of whether the market adequately rewards or punishes companies for their ESG practices over the long term (for a discussion of this topic, see Capelle-Blancard, Desroziers and Scholtens (2021), Cui and Docherty (2020), Glossner (2021)).

2.8 Conclusion

Recent research has sought to establish a causal link between ESG practices and companies' market value. While there are many sources of information on companies' ESG practices, news stands out for its timeliness and frequency. However, recent research in this area relies on commercial and non-transparent big data solutions. To address this gap, we propose a new ESG Dictionary that researchers can use to filter ESG news from regular ones, define news sentiment, and classify news based on SASB's dimensions and categories.

We evaluate the dictionary in the classification of news from 2014 to 2022 on the Valor Econômico portal. For the Brazilian case, we demonstrate that companies' stock prices react to ESG news, both to positive news (positive price reaction) and to negative news (negative price reaction). Since the news release date is t, the reaction occurs in t, or in some cases $t-1$ to t or $t-5$ to t. This indicates that a portion of the news is "old" or stale, meaning that, in certain instances, the market incorporates the information one or more days prior to its release.

We also reach the same conclusion as Serafeim and Yoon (2022): price reaction would only be significant when the news contains financially material information. In other words, we do not support the hypothesis that all ESG news is relevant to investors' decision-making processes: investors would not react for reputational or non-financial reasons alone. However, unlike Serafeim and Yoon (2022), we found a negative reaction to negative news using only one news source.

Some studies do not find a positive price reaction to positive news, as in Krüger (2015) (which finds a negative reaction and raises the hypothesis of an agency problem) and Capelle-Blancard and Petit (2019) (which finds no reaction). Our results contradict those works. When we investigate the source of our result by SASB dimension, we find that the Leadership & Governance dimension did not exhibit a statistically significant price response to positive news. In other words, the results may vary depending on the set of ESG topics considered in each study, which may explain a portion of the discrepancy between previous works.

3 Brazilian exchange rate forecasting in high frequency

3.1 Introduction

It is practically a consensus in the literature that foreign exchange rates are hard to predict. Meese and Rogoff (1983) showed that traditional macroeconomic models do not outperform a naive random walk in forecasting exercises for medium and short-term prediction horizons (one month to one year), even when using contemporary data for the covariates. Cheung, Chinn and Pascual (2005) and Rossi (2013) evaluated the models developed in the following decades and concluded that no model performed very well. The predictability of exchange rates, for some models, depends on the choice of forecast horizon, exchange rates, sample period, and the toughest benchmark to beat is the random walk without drift. In general, the macroeconomic models used in these studies make use of macroeconomic variables available only at low frequencies.

At higher frequencies, Evans and Lyons (2002) opened a new research area: how microstructure (order flows) affects exchange rates. Evans and Lyons (2005) showed that daily customer order flow, from one day to one month, can beat the Random Walk in a real out-of-sample setup, a conclusion similar to that of Rime, Sarno and Sojli (2010) with daily data. However, in contrast, Danielsson, Luo and Payne (2012) concludes that predictability is only valid at higher frequencies (intraday data).

In this paper, we go in a complementary direction to the micro-structure literature. Our contribution to the literature is to evaluate the predictability of the Brazilian exchange rate[1] at high frequencies (1, 5, and 15 minutes) using local and global economic variables, namely: short- and long-term Brazilian interest rates, the Brazilian stock market index, gold price, oil price, stock market option-based implied volatility (VIX), and exchange rates of 17 other countries. In addition to the ordinary least squares (OLS) method, we also evaluate the use of Machine Learning (ML) algorithms such as LASSO, Ridge, Elastic Net, Random Forest, and Gradient Boosting in the prediction exercises.

We chose the variables based on their availability at the desired frequency (maximum frequency of 1 minute) and their potential relationship with the Brazilian exchange rate. For example, short- and long-term interest rates on Interbank Deposit (DI) contracts

[1] The Brazilian case is interesting since its exchange rate is floating according to the IMF classification. Brazil is one of the largest emerging economies, and the Brazilian Real has shown high volatility and moments of great depreciation in recent years, with significant impacts on the macroeconomic environment. However, exchange rate forecasting papers for the Brazilian currency focus on monthly and daily frequencies, as in Gaglianone and Marins (2017) and Moura, Lima and Mendonça (2008).

reflect expectations regarding monetary and fiscal policy, both of which affect the exchange rate. The relationship between the stock exchange and the exchange rate is also a recurrent theme in research, as demonstrated by Tabak (2006). Ferraro, Rogoff and Rossi (2015) conjectured that, in small open commodity-exporting economies, the exchange rate is expected to reflect the movement of commodity prices. Beckmann, Czudaj and Arora (2017) also shows that oil price is a potential predictor of the exchange rate in the short run. VIX can be seen as a measure of uncertainty (Bekaert; Hoerova; Duca, 2013), and it can explain part of the daily variation of the nominal exchange rate (Kohlscheen; Avalos; Schrimpf, 2017). Gold is also investigated as a possible predictor, due to a potential bi-directional causal relationship with exchange rates in emerging countries (Güriş; Kiran, 2014; Nair; Choudhary; Purohit, 2015). Finally, motivated by Felício and Júnior (2014), who extract common factors from a set of floating exchange rates and assess their predictive capacity, we consider 17 foreign exchange rates as possible predictors, as well as a global factor extracted from these exchange rates through PCA.

We perform two types of forecasting exercises. The first, called out-of-sample fit, uses contemporary data as predictors (realized values of the predictor variables). This type of analysis captures correlations or co-movements. As Ferraro, Rogoff and Rossi (2015) put it in the forecasting literature, this type of prediction can be useful when we are interested in evaluating the predictive capacity of a model given a trajectory for some unmodeled set of variables. In other words, if it is possible to obtain a good model to predict this variable, then this model can be exploited to predict the exchange rate. Important examples of its use are the studies by Meese and Rogoff (1983) and Cheung, Chinn and Pascual (2005), which showed that, even using realized values of the predictor variables, traditional models were unable to beat the Random Walk in exchange rate prediction. The second exercise is the out-of-sample forecast, in which we seek to forecast the exchange rate at $t+h$ using information available only up to time t.

We verified that, in the out-of-sample fit exercise, for each variable, it is possible to predict the Brazilian exchange rate at high frequency with less error than the Random Walk for the 1, 5, and 15-minute frequencies. We also found that the local economic variables present a lower mean squared error than the global variables. The augmented model, which considers all variables, outperforms all individual models. Machine Learning models further reduce the mean squared error (MSE) when applied to the augmented model. Some ML models perform better than others, such as decision tree-based models or simpler models that utilize a strategy of determining hyperparameters at each advancement of a training window on the data, as we will detail in section 3.3.2. In the out-of-sample forecasting exercise, a simple autoregressive model beats the Random Walk over a horizon of up to 2 minutes, at a frequency of one minute. Moreover, Machine Learning models can beat the Random Walk for horizons of up to 4 minutes.

When applying Gradient Boosting, we also estimate the relative importance of each predictor for both exercises. We observed that the predictors considered most important are those that carry local information, such as short- and long-term interest rates and the index of the Brazilian stock exchange, along with the exchange rate of the Mexican economy, and the exchange rates of the BRICS countries present in our dataset. We also conducted an out-of-sample forecasting exercise using the rates of the Real/U.S. dollar futures contracts as predictors. As stated by Ventura and Garcia (2012), in Brazil, the exchange rate is first determined in the exchange rate futures market (at the next maturity) and then transmitted by arbitrage to the spot market. Therefore, another contribution of this paper is to show how this relationship translates into out-of-sample prediction ability at high frequency.

In relation to the use of Machine Learning techniques, our paper differs from works like those of Colombo and Pelagatti (2020), Zhang and Hamori (2020), Amat, Michalski and Stoltz (2018) in terms of frequency and set of economic variables. At high frequency, most works use technical trading strategies or univariate strategies for prediction, such as those by Manahov, Hudson and Gebka (2014), Palikuca and Seidl (2016), or combine microstructure with ML, as in the work of Choudhry et al. (2012). We differ from these by evaluating a different set of predictors and benchmarking against the Random Walk.

The paper is structured as follows: In Section 3.2, we describe the database. In Section 3.3, we present the models used for prediction and the methodology used to determine the parameters and hyperparameters. In Section 3.4, we present and discuss the results of the out-of-sample fit and out-of-sample prediction exercises. Section 3.5 concludes.

3.2 Data

We use intraday data from 2021-05-05 to 2021-11-12 (128 business days), with a 1-minute frequency (closing prices), obtained from Bloomberg, resulting in a total of 46,080 observations. For exercises on frequencies of 5 and 15 minutes, we re-sample the data. Each day, the interval considered is from 10:00 am to 4:00 pm. Although the Brazilian foreign exchange market operates from 9:00 am to 6:00 pm, the stock exchange opens at 10:00 am, and the trading of future interest rate contracts is halted at 4:00 pm.

Our interest lies in forecasting the Brazilian Real/U.S. dollar spot exchange rate. As candidates for predictor variables, we consider, for short- and long-term interest rates, *DI Futures* contracts maturing in January 2023 (DI23) and January 2029 (DI29). The underlying asset of the DI futures is the average daily interest rate of interbank deposits (DI). The notional value of one DI future contract is R$100 thousand, and the value on the trade date is equivalent to this amount discounted at the negotiated rate. This rate reflects the expected evolution of the DI, that is, the expectations regarding the future interest rate (Vartanian et al., 2021). As stated by Jeanneau, Araujo and Amante (2007), the Brazilian futures market is one of the main indicators of interest rate expectations, with the yield curve implicit in DI futures being the main benchmark for fixed-income investment in Brazil. As a representative of the back end of the yield curve, we chose the contracts maturing in 2029 because they are the most liquid of the longer-term futures contracts.

For the oil price, we use the West Texas Intermediate (WTI) crude oil price. The stock market index used was the Ibovespa from the B3 (Brasil, Bolsa, Balcão - Brazilian Stock Exchange). We also include the gold price and the Chicago Board Options Exchange Volatility Index (VIX).

We selected 17 other nominal exchange rates as predictors based on the following criteria: they are floating exchange rates; they are traded at the same times as the Brazilian exchange rate; and they are from countries (and economic alliances) that have a GDP of at least 10% of the Brazilian GDP. The countries and economic alliances that met these criteria were: Australia (USDAUD), Canada (USDCAD), Czech Republic (USDCZK), euro area (USDEUR), Israel (USDILS), Japan (USDJPY), Mexico (USDMXN), New Zealand (USDNZD), Norway (USDNOK), Poland (USDPLN), Russian Federation (USDRUB), South Africa (USDZAR), Sweden (USDSEK), Switzerland (USDCHF), Thailand (USDTHB), Turkey (USDTRY), and England (USDGBP). All exchange rates, including the Real/U.S. dollar rate, were obtained from the Bloomberg Generic Composite rate (BGN) pricing algorithm. In the spirit of Medeiros, Schütte and Soussi (2022) and Felício and Júnior (2014), we extract a common factor from the set of all foreign exchange rates, through PCA, and also use it as a predictive variable in the out-of-sample forecasting exercise. Finally, we also conducted an out-of-sample forecasting exercise using the Brazilian Real/U.S. dollar futures contracts rate (next maturity) as a predictor.

To achieve stationarity, all variables are in (log) first differences. Only intraday differences are considered; that is, day-to-day differences are discarded.

3.3 Methods

In the next subsection, we present the models considered in this paper, followed by the strategy for determining the parameters and hyperparameters, as well as the performance evaluation methods used.

3.3.1 Exchange Rate Forecasting Model

Let s_t be the logarithm of the exchange rate. Our interest is in predicting the change in the logarithm of the exchange rate h steps ahead, that is, $\Delta s_{t+h} = s_{t+h} - s_t$. Let

$\mathbf{x}_{t+1}$ be a set of predictors, with information up to $t+1$. Then our general **out-of-sample fit** prediction model is defined by:

$$\Delta s_{t+1} = f(\mathbf{x}_{t+1}) + \varepsilon_{t+1} \tag{3.1}$$

where $f()$ is a mapping to be estimated, and ε_{t+1} is the prediction error. Note that the out-of-sample fit exercise uses contemporary data (realized values of the predictor variables)[2].

The model of our second exercise, **out-of-sample** forecasting, is defined by:

$$\Delta s_{t+h} = f(\mathbf{x}_{t-p+1:t}) + \varepsilon_{t+h} \tag{3.2}$$

where $f()$ is also a mapping to be estimated, $\mathbf{x}_{t-p+1:t}$ represents a set of predictors with information only up to t, and p the number of lags considered. We also use the lags of the Real/U.S. dollar exchange rate as a predictor. We set $p = 3$, seeking a balance between the amount of past information used in the forecast and computational cost. Preliminary tests showed that using a larger number of lags does not significantly alter the forecasting results.

Note that we are working with direct forecasting, meaning for each $h > 1$, we aim to predict the entire exchange rate variation between t and $t+h$ at once. This implies that we will have a different model for each forecast horizon in the out-of-sample exercise.

For both exercises, we initially evaluate the predictive power of the variables using $f()$ as a Linear Regression Model, estimated by OLS on a rolling window, as presented in Subsection 3.3.2. Thus, the forecasting model is defined for the out-of-sample fit exercise by:

$$\Delta s^f_{t+1} = \mathbf{x}'_{t+1}\hat{\beta} \tag{3.3}$$

where $\hat{\beta}$ is a $k \times 1$ vector of parameters to be estimated, and $\mathbf{x}_{t+1}$ a $k \times 1$ vector of the predictors, including a constant term. For the out-of-sample forecasting exercise, the Linear Regression Model is defined by:

$$\Delta s^f_{t+h} = \mathbf{x}'_{t-p+1:t}\hat{\beta} \tag{3.4}$$

For both the out-of-sample fit and out-of-sample forecasting exercises, we initially examine the individual predictive capabilities of the variables considered in this paper. Subsequently, we group some of these variables, testing what we term ‘augmented models.’

[2] Note that this is not an in-sample prediction exercise, as $\mathbf{x}_{t+1}$ will not be used for parameter prediction.

When incorporating all available predictors, we consider Machine Learning methods such as LASSO, Ridge, ElasticNet, Random Forest, and Extreme Gradient Boosting for the mapping $f()$, as described below. Equations (3.3) and (3.4) are also applicable to regularized linear ML Models. However, the method of estimating $\hat{\beta}$ differs for each model, as will be detailed in the subsequent subsections.

3.3.1.1 LASSO

LASSO (Least Absolute Shrinkage and Selection Operator), as proposed by Tibshirani (1996), penalizes the regression with the L_1 norm of the parameter vector $\hat{\beta}$. That is, the $\hat{\beta}$ in Equations (3.3) and (3.4) is determined as follows[3]:

$$\hat{\beta} = \underset{\beta_1,\dots,\beta_k}{\arg\min} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\Delta s_{t+h} - \sum_{j=1}^{k} x_{j,t}\beta_j \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right] \tag{3.5}$$

Given the penalty imposed by the L_1 norm, in addition to performing shrinkage, LASSO also selects variables by zeroing out the coefficients deemed irrelevant. As such, it is one of the most popular regularization methods in data-rich environments (Masini; Medeiros; Mendes, 2020). When $\lambda = 0$, it defaults to OLS. The determination of λ, also referred to as a hyperparameter in Machine Learning literature, can be achieved through cross-validation or by employing criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Specifically for LASSO, we employed four different strategies to determine this hyperparameter, as detailed in Section 3.3.2.

3.3.1.2 Ridge

According to Masini, Medeiros and Mendes (2020), Ridge regression, proposed by Hoerl and Kennard (1970), seeks to combat problems generated by multicollinearity in Linear Regression, stabilizing the solution of the problem by introducing a small bias in exchange for reducing the variance of the estimator. Ridge penalizes the regression with the L_2 norm of the parameter vector. Therefore, $\hat{\beta}$ in Equations (3.3) and (3.4) is determined by:

$$\hat{\beta} = \underset{\beta_1,\dots,\beta_k}{\arg\min} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\Delta s_{t+h} - \sum_{j=1}^{k} x_{j,t}\beta_j \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right] \tag{3.6}$$

We determine λ by cross-validation using two different strategies, as presented in Subsection 3.3.2. The greater the λ, the more significant the shrinkage of the coefficients. For $\lambda = 0$, the model reduces to an OLS, similar to LASSO.

[3] In the descriptions of LASSO, Ridge, and Elastic Net, we use the notation adopted by Costa et al. (2021).

According to Zou and Hastie (2005), it has been empirically observed that Ridge performs better than LASSO in the presence of significant multicollinearity among the predictors. However, Ridge does not generate parsimonious models: the least relevant predictors have their coefficients shrunk towards zero, but they never become exactly zero.

3.3.1.3 Elastic Net

Proposed by Zou and Hastie (2005), the idea of Elastic Net is to combine the advantages of LASSO and Ridge. The $\hat{\beta}$ is determined by:

$$\hat{\beta} = \underset{\beta_1,\ldots,\beta_k}{\arg\min} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\Delta s_{t+h} - \sum_{j=1}^{k} x_{j,t}\beta_j \right)^2 + \lambda \left(\alpha \sum_{j=1}^{k} \beta_j^2 + (1-\alpha) \sum_{j=1}^{k} |\beta_j| \right) \right] \tag{3.7}$$

with $\alpha \in [0,1]$. The Elastic Net penalty is a convex combination of the L_1 penalties, which perform variable selection, and the L_2 penalty, which stabilizes the solution of the problem (Masini; Medeiros; Mendes, 2020). λ, in turn, determines the overall strength of the penalties. Note that LASSO and Ridge regressions are special cases of Elastic Net, for $\alpha = 0$ and $\alpha = 1$, respectively. We use two different strategies to determine α and λ, as presented in the Subsection 3.3.2.

3.3.1.4 Random Forest

Proposed by Breiman (2001), Random Forest is essentially an ensemble of decision trees. These decision trees recursively partition the domain of the variables into non-overlapping rectangular regions. Each region, denoted as R_i, is associated with a constant value in the context of a regression problem. Visually, these regions can be represented as step functions:

$$f(\mathbf{x}) = \sum_i c_i I(\mathbf{x} \in R_i) \tag{3.8}$$

In Random Forest, each tree is built based on a bootstrap sample of the training data. For each tree, the following steps are repeated for each terminal node, until a certain stopping criterion is met (for example, controlling the maximum depth or the number of leaves) (Friedman, 2017):

(i) Randomly select l variables from the available k variables;

(ii) Select the best variable/split-point among the l variables (the combination that minimizes the mean squared error) and split the node into two child nodes.

Let B be the number of bootstrap samples. The ensemble of trees $\{T_b\}_1^B$ is stored, and the prediction for a new point $\mathbf{x}$ is given by the average of the predictions from each tree:

$$\Delta s_{t+h}^f = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}_t) \tag{3.9}$$

where $T_b(\mathbf{x})$ is the prediction of the b-th regression tree. Thus, the overarching principle of Random Forest is to generate an average from several unbiased yet noisy models (trees), similar to bagging (bootstrap aggregation). This approach reduces the correlation between the trees, thereby decreasing the variance of the average. This reduction in variance is achieved by randomly selecting sets of variables for each division within the tree (Friedman, 2017).

3.3.1.5 Gradient Boosting

Proposed by Friedman (2001), Gradient Boosting is based on 'weak learners', which are trees with low predictive power, utilized in an additive, step-by-step model. At each step, a weak learner is trained to address the shortcomings of previous weak learners. This process can be illustrated by the algorithm presented in Friedman (2017), Section 10.10. For a loss function $L(y, f(x))$, with N observations in the training data and M weak learners, the operation is as follows:

1. Initialize the model with an optimal constant, which is simply a single terminal node of a tree: $f_0(\mathbf{x}) = \arg\min_\gamma = \sum_{i=1}^{N} L(y_i, \gamma)$;

2. For $m = 1, 2, ..., M$:

   a) For $i = 1, 2, ..., N$ compute:

   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

   b) Fit a regression tree to the targets r_{im}, giving terminal regions R_{jm}, $j = 1, 2, ..., J_m$.

   c) For $j = 1, 2, ..., J_m$ compute:

   $$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

3. The final forecast is given by $f_M(\mathbf{x})$

In this work, we use the Extreme Gradient Boosting (XGB) version developed by Chen and Guestrin (2016), which is an improved version of Gradient Boosting in terms of both predictive and computational performance[4].

3.3.2 Parameters and Hyper-parameters determination

For both Linear Regression and Machine Learning models, we estimate the model parameters using a rolling window as follows: We divide the data into two parts - training data, comprising the first half of the observations, and test data, comprising the final half. We estimate the model parameters on the training data and then make the prediction for the next observation. Subsequently, we incorporate the next observation from the test data into the training data, discard the original first observation of the training data, re-estimate the parameters over this updated rolling window, and then perform a new prediction. This process is repeated until the rolling window has cycled through all of the test data.

In the case of ML models, we determine the hyperparameters in three different ways[5]. Firstly, for LASSO, Ridge, and Elastic Net, when each rolling window advances one step, a grid search algorithm tests possible hyperparameter combinations as follows: for each combination, we determine the model parameters using 62.5% of the initial rolling window observations and use them to predict the next 12.5% of observations. We compute and store the MSE and repeat this process, first using 75% of the initial rolling window data for parameter estimation and then performing the prediction on the subsequent 12.5% of the data. We compute and store the MSE again. This process is repeated once more, using 87.5% of the rolling window data for parameter estimation and the last 12.5% for prediction, ensuring that all rolling window observations are utilized. We select the

[4] For example, regression trees can include a regularization parameter that influences the tree pruning mechanism, acting against overfitting and reducing the sensitivity of predictions to individual observations. The algorithm is also known for optimizations in handling large datasets. For instance, it can employ an Approximate Greedy Algorithm, which does not test every possible threshold at each tree split, but instead uses quantiles as candidate thresholds. Furthermore, XGB utilizes Parallel Learning, allowing the dataset to be split across multiple computers simultaneously. It innovates with a native mechanism for building trees in the presence of missing values (Sparsity-Aware Split Finding). Additionally, it seeks to optimize the use of computer hardware, for example, by storing gradients in the processor's cache to quickly calculate the scores used in defining the tree split points (Cache-Aware Access). XGB is also capable of employing techniques such as Sharding, which divides data between multiple storage units to access them in parallel (Blocks for Out-of-Core Computation). Due to these features – some of which are implementation-specific and not directly related to statistics – XGB is one of the favored algorithms in ML competitions (Costa et al., 2021).

[5] We aim to find the best methods to determine hyperparameters given the constraints of computational costs.

combination of hyperparameters that yields the smallest mean squared error for the final estimation of parameters over the entire rolling window/training data. Algorithms using this cross-validation method are indicated in the results tables with the suffix (CV) (e.g., LASSO (CV)).

Secondly, for algorithms with higher computational costs, such as Random Forest and Extreme Gradient Boosting, we perform the process described in the previous paragraph only once, using the initial training data. In other words, the hyperparameters are not redetermined at each advance of the rolling window; instead, they are determined at the beginning of the process and the same hyperparameters are used for every subsequent advance of the rolling window. This strategy is also applied to Ridge, LASSO, and Elastic Net for comparison purposes. Algorithms using this technique are indicated in the results tables with the suffix (GS) (for example, Ridge (GS)). Appendix B presents the range of hyperparameter values considered in the cross-validation processes[6].

Third, in the case of LASSO, we use the Least Angle Regression (LARS) algorithm, developed by Efron et al. (2004), in conjunction with either the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to determine the hyperparameter, instead of using cross-validation. This method is advantageous due to its low computational cost compared to other methods, and thus it is applied at each step of the rolling window. In the results tables, we refer to this method as LASSO (AIC) or LASSO (BIC), depending on the information criterion used.

3.3.3 Model Valuation

The Random Walk without drift is used as the benchmark, the model to beat. It implies a forecast of no change in the exchange rate h steps ahead, which is represented as:

$$\Delta s_{t+h}^{f} = 0 \tag{3.10}$$

For the Random Walk and for the estimated models, the mean squared errors are computed, and the ratios of the MSEs of the models in relation to the Random Walk are reported. That is, values lower than 1 for this ratio suggest a better predictive capacity of the model compared to the Random Walk.

In the out-of-sample exercise, we also use an autoregressive (AR) specification with one lag as a benchmark[7].

[6] The range or number of possible values for each hyperparameter may vary according to the cross-validation strategy used. We aimed to balance the number of possible values with the computational cost of each cross-validation strategy in determining these ranges.

[7] The one-lag specification was selected using the Bayesian Information Criterion (BIC).

We apply the Diebold and Mariano test (Diebold; Mariano, 1995) to determine whether the forecasting of the model is statistically different from that of the Random Walk.

We also apply the Model Confidence Sets (MCS) procedure as described by Hansen, Lunde and Nason (2011). The MCS involves a sequence of tests applied to a group of evaluated models to determine a subset considered 'superior' compared to the others. This procedure begins by assessing all models and calculating a statistic ($t_{i\cdot}$) for each model i, where this statistic represents the loss of that model relative to the average losses across models in the set, based on a given loss function (such as the mean squared error in our case). We set the confidence level $(1 - \alpha)$ at 80% as in Garcia, Medeiros and Vasconcelos (2017), with α representing the significance level of the test.

Under this framework, the model with the highest $t_{i\cdot}$ statistic, indicating the worst performance relative to the others, is excluded if its statistic exceeds a critical value derived from a bootstrap distribution, corresponding to the set α level. This exclusion implies that the model does not fall within the desired $(1 - \alpha)$ confidence interval for the superior set of models. The process continues iteratively, excluding one model at a time, and concludes when no further models can be statistically distinguished for exclusion based on the α level. This iterative exclusion ultimately results in the final Model Confidence Set (Bernardi; Catania, 2018).

3.4 Results

Initially, we present the results for the out-of-sample fit exercise, followed by the results for the out-of-sample forecasting exercise. Our forecasting exercises were implemented in Python, primarily using the Scikit-learn libraries (Pedregosa et al., 2011) for the ML algorithms, with the exception of XGBoost, for which we used the specifically developed library by Chen and Guestrin (2016). The Model Confidence Sets alone were determined using R, utilizing the MCS implementation from Bernardi and Catania (2018).

3.4.1 Out-of-sample-fit results

Table 8 presents the results of the out-of-sample fit prediction exercise for the Brazilian exchange rate, across frequencies of 1, 5, and 15 minutes. These results utilize Ordinary Least Squares and various Machine Learning algorithms for the augmented model. The values shown are the ratios of the models' mean squared errors to the MSE of the Random Walk. Therefore, values less than 1 indicate the models' superiority over the Random Walk, with statistical significance determined by the test of Diebold and Mariano (1995).

Table 8 – Brazilian Real/U.S. dollar exchange rate forecasting exercise, out-of-sample fit results.

		MSE ratio		
Predictors	**Method**	**freq. 1 min** *h=1*	**freq. 5 min** *h=1*	**freq. 15 min** *h=1*
-	RW	1.0000	1.0000	1.0000
DI23, DI29	OLS	0.8096***	0.7270***	0.7325***
Ibovespa	OLS	0.7934***	0.7625***	0.8127***
Oil price	OLS	0.9820***	0.9814***	0.9771**
VIX	OLS	0.9605***	0.9307***	0.9371***
Gold price	OLS	0.9727***	0.9740***	0.9779
17 currencies	OLS	0.8255***	0.7927***	0.8052
DI23, DI29, Ibovespa	OLS	0.7020***	0.6358***	0.6806***
DI23, DI29, Ibovespa, Oil price, VIX, Gold price	OLS	0.6760***	0.6078***	0.6463***
DI23, DI29, Ibovespa, Oil price, VIX, Gold price, 17 currencies	OLS	0.6237***	0.5578***	0.5886***
	LASSO (CV)	0.6817***	0.5262***	0.5505***
	LASSO (GS)	0.6227***	0.5577***	0.5777***
	LASSO (AIC)	0.6907***	0.5525***	0.5728***
	LASSO (BIC)	0.6896***	0.5446***	0.5647***
	Ridge (CV)	0.6178***	0.5297***	0.5523***
	Ridge (GS)	0.6237***	0.5578***	0.5855***
	ElasticNet (CV)	0.6818***	**0.5259*** **	**0.5499*** **
	ElasticNet (GS)	0.6237***	0.5570***	0.5803***
	RF (GS)	0.6038***	0.5443***	0.5550***
	XGB (GS)	**0.6014*** **	0.5357***	0.5517***

Models MSE ratios in relation to Random Walk without drift. In bold, the lowest MSE ratios for each frequency / horizon (h). ***, **, and * indicate rejection of the null hypothesis (model and Random Walk predictions are not different), at the 1%, 5% and 10% levels respectively. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window. DI23 and D29 are Interbank Deposit interest rates futures contracts maturing in January 2023 and January 2029. The 17 currencies correspond to the foreign exchange rates listed in the section 3.2.

The results for the OLS models, which incorporate as predictors future interest rates, Ibovespa, oil price, VIX, gold price, or some of the 17 exchange rates considered, suggest that if it is possible to accurately predict the values of any of these variables at $t+1$ (perfect foresight), then forecasting the Brazilian exchange rate at high frequency with less error than the Random Walk is also feasible for all tested frequencies. The only exception is when using gold as a predictor for the 15-minute frequency[8]. Our work finds the same co-movements between some of our predictors and the exchange rate at high frequencies as observed at lower frequencies, such as in Júnior (2014), Ferraro, Rogoff and Rossi (2015).

It is interesting to note that global variables - such as oil price, gold price, and VIX - exhibit predictive power in this out-of-sample fit exercise. However, as expected, their impact is to a lesser extent compared to local variables (such as short and long-term interest rate futures and Ibovespa), which absorb local news also impacting the exchange rate. Additionally, we observe that increasing the number of predictor variables in models using OLS as the method results in a decrease in the MSE. The linear regression specification with the lowest mean squared error is the one that utilizes all available predictor variables.

We also tested the performance of Machine Learning algorithms using all available variables as predictors. Generally, we observe that ML algorithms exhibit lower MSE ratios than OLS models, except for some models at the 1-minute frequency. Table 9 presents the sets of models deemed superior for each frequency by the Model Confidence Sets procedure. Notably, only the Machine Learning models are part of these superior model sets, outperforming the OLS-based specifications across all frequencies. Among the ML models, those that employed the cross-validation strategy at each step of the rolling training window (CV) and the decision tree-based models were considered superior for the 5 or 15-minute frequencies. For the 1-minute frequency, XGBoost and Random Forest were superior to the others. These results highlight the significance of the cross-validation strategy (whether determining the hyperparameters once or at each step of the training window). However, even with a simpler cross-validation strategy, the decision tree-based models proved competitive or even superior to models that redefined the hyperparameters for each point forecast, as was the case with the 1-minute frequency.

Algorithms based on regression trees, such as Random Forest or Extreme Gradient Boosting, can automatically provide estimates of the importance of each predictor. By 'importance,' we refer to the usefulness of each variable in reducing the mean squared error during the construction of trees in the algorithm's training stage[9]. We present in Figures 6 to 8 the average relative importance attributed by XGB to each predictor for the frequencies of 1, 5, and 15 minutes.

It is noteworthy that, across all frequencies, the short and long-term interest rates (DI23 and DI29), Ibovespa, and the Mexican exchange rate alternate as the four most important predictors. This indicates how local information embedded in key financial and

[8] In our out-of-sample fit exercise, we do not consider Brazilian Real/U.S. dollar futures, as their natural correlation with the spot exchange rate is well-established due to the non-arbitrage condition imposed by Covered Interest Rate Parity.

[9] For more details on how the importance of each variable is determined, see Section 10.13.1 of Friedman (2017)

Table 9 – Out-of-sample fit Superior Set of Models.

Frequency	Model	Rank	$t_{i\cdot}$	p-value
1 minute	XGB (GS)	1	-1.17	1.0000
	RF (GS)	2	1.17	0.2420
5 minutes	ElasticNet (CV)	1	-1.35	1.0000
	LASSO (CV)	2	-1.23	1.0000
	Ridge (CV)	3	0.11	0.9832
	XGB (GS)	4	0.92	0.4608
15 minutes	ElasticNet (CV)	1	-1.75	1.0000
	LASSO (CV)	2	-1.55	1.0000
	Ridge (CV)	3	-0.74	1.0000
	XGB (GS)	4	-0.45	1.0000
	RF (GS)	5	-0.17	1.0000
	LASSO (BIC)	6	1.07	0.5704
	LASSO (AIC)	7	1.58	0.2684

Sets created for an 80% confidence interval. 'Rank' corresponds to the ranking based on the $t_{i\cdot}$ statistic. In the context of the MCS procedure, for a confidence level of $1-\alpha$, models with p-values greater than α are not statistically worse than the best model and are therefore included in the set of superior models. Essentially, a higher p-value indicates a model's performance is closer to the top-performing models within the evaluated set. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window.

economic variables, along with the exchange rate of an economy similar to Brazil's, are reflected in the co-movements of high-frequency exchange rates. Additionally, significant importance is attributed to the South African and Russian exchange rates, the two BRICS countries available in our database.

3.4.2 Out-of-sample results

In the out-of-sample forecasting exercise, we seek to predict the Brazilian spot exchange rate in $t+h$ with information available only up to t. In our case, we use three lags of the predictor variables, as well as three lags of the log returns of the Real/U.S. dollar

Figure 6 – Importance of XGB predictors in out-of-sample fit forecasting exercise at 1-minute frequency.

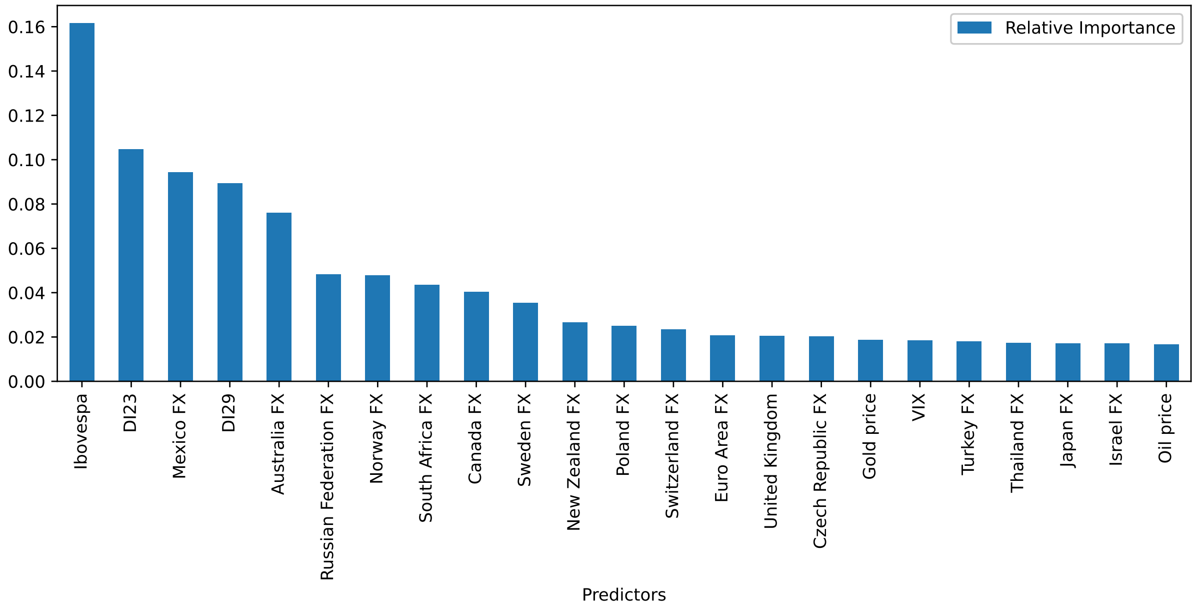
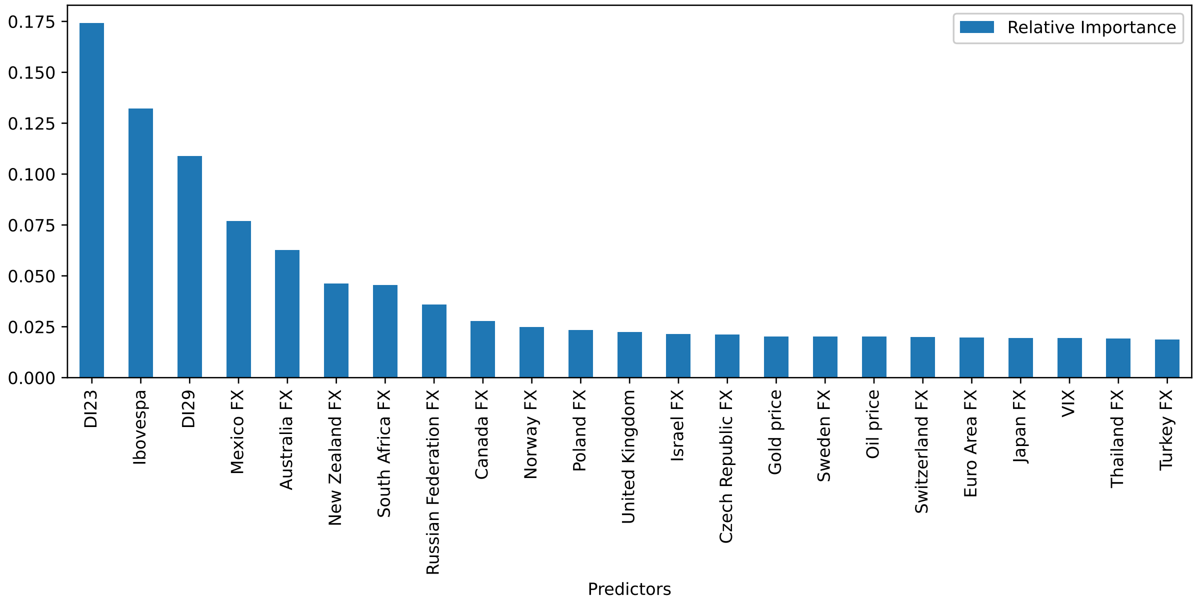


Figure 7 – Importance of XGB predictors in out-of-sample fit forecasting exercise at 5-minutes frequency.



exchange rate itself. We also use the same specifications explored in the out-of-sample fit exercise in terms of predictors and models, adding as a predictor variable a common factor of exchange rates extracted by PCA. Table 10 presents the results, where we can see how challenging this exercise is.

Figure 8 – Importance of XGB predictors in out-of-sample fit forecasting exercise at 15-minutes frequency.

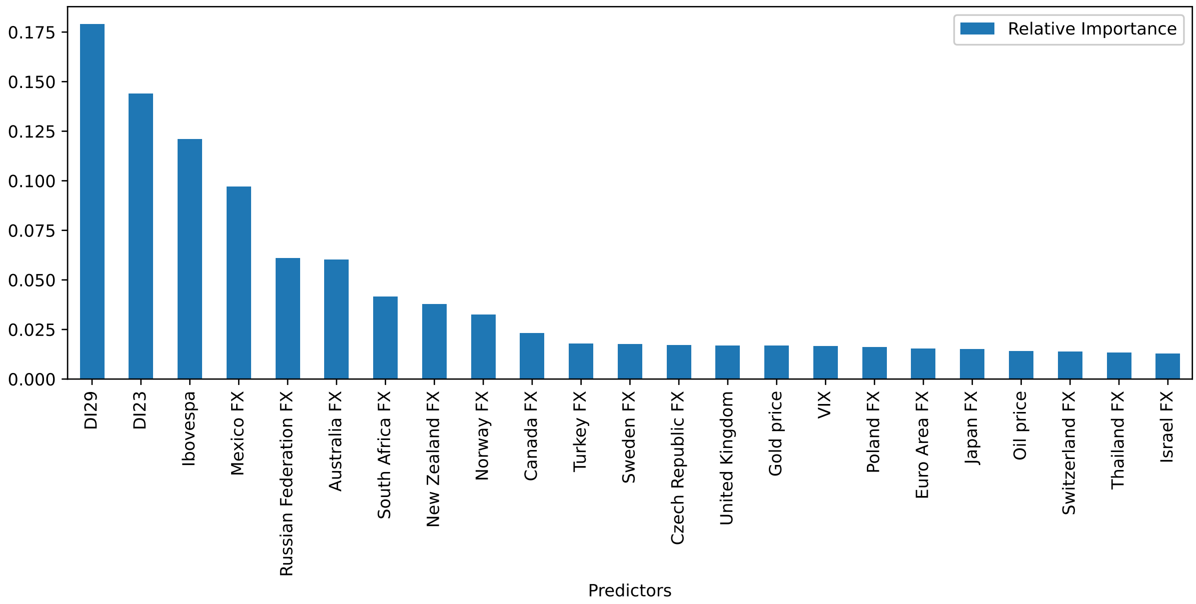


Table 10 – Brazilian Real/U.S. dollar exchange rate forecasting exercise, out-of-sample results.

		MSE ratio						
Predictors	**Method**	**frequency 1 min**					**f. 5 min**	**f. 15 min**
		h=1	*h=2*	*h=3*	*h=4*	*h=5*	*h=1*	*h=1*
-	RW	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
-	AR(1)	0.9960***	**0.9976****	0.9985	0.9984	0.9991	1.0014	1.0022
DI23, DI29	OLS	0.9962**	0.9980	0.9991	0.9992	1.0001	1.0078	1.0191
Ibovespa	OLS	0.9962**	0.9981	0.9992	0.9993	1.0000	1.0042	1.0078
Oil price	OLS	0.9963***	0.9981	0.9990	0.9991	0.9998	1.0018	1.0065
VIX	OLS	0.9961***	**0.9976***	0.9987	0.9990	0.9998	1.0048	1.0067
Gold price	OLS	0.9968**	0.9986	0.9994	0.9994	1.0002	1.0049	1.0073
17 currencies, FX common factor	OLS	0.9988	1.0011	1.0012	1.0008	1.0014	1.0210	1.0616
DI23, DI29, Ibovespa	OLS	0.9964**	0.9986	0.9997	0.9997	1.0005	1.0099	1.0232
DI23, DI29, Ibovespa, Oil price, VIX, Gold price	OLS	0.9972	0.9998	1.0009	1.0009	1.0017	1.0140	1.0336
DI23, DI29, Ibovespa, Oil price, VIX, Gold price, 17 currencies, FX common factor	OLS	0.9993	1.0023	1.0027	1.0024	1.0032	1.0312	1.0950
	LASSO (CV)	0.9953***	0.9984	0.9992	0.9987	0.9999	1.0004	1.0013
	LASSO (GS)	0.9950***	0.9986	0.9985**	**0.9980****	0.9991	1.0013	1.0009
	LASSO (AIC)	0.9956***	0.9990	1.0000	0.9992	1.0001	1.0029	1.0095
	LASSO (BIC)	**0.9947*****	**0.9976****	0.9984**	0.9985*	0.9997	1.0000	1.0012
	Ridge (CV)	0.9967*	0.9994	1.0001	1.0002	1.0009	1.0076	1.0040
	Ridge (GS)	0.9987	1.0018	1.0027	1.0024	1.0032	1.0020	1.0009
	ElasticNet (CV)	0.9953***	0.9984	0.9992	0.9987	0.9999	1.0004	1.0013
	ElasticNet (GS)	0.9950***	0.9984	0.9985	0.9983	0.9996	1.0008	1.0009
	RF (GS)	0.9965**	0.9977*	0.9995	0.9999	1.0027	1.0057	1.0041
	XGB (GS)	0.9949***	0.9985	**0.9977***	0.9985	0.9998	1.0032	1.0284

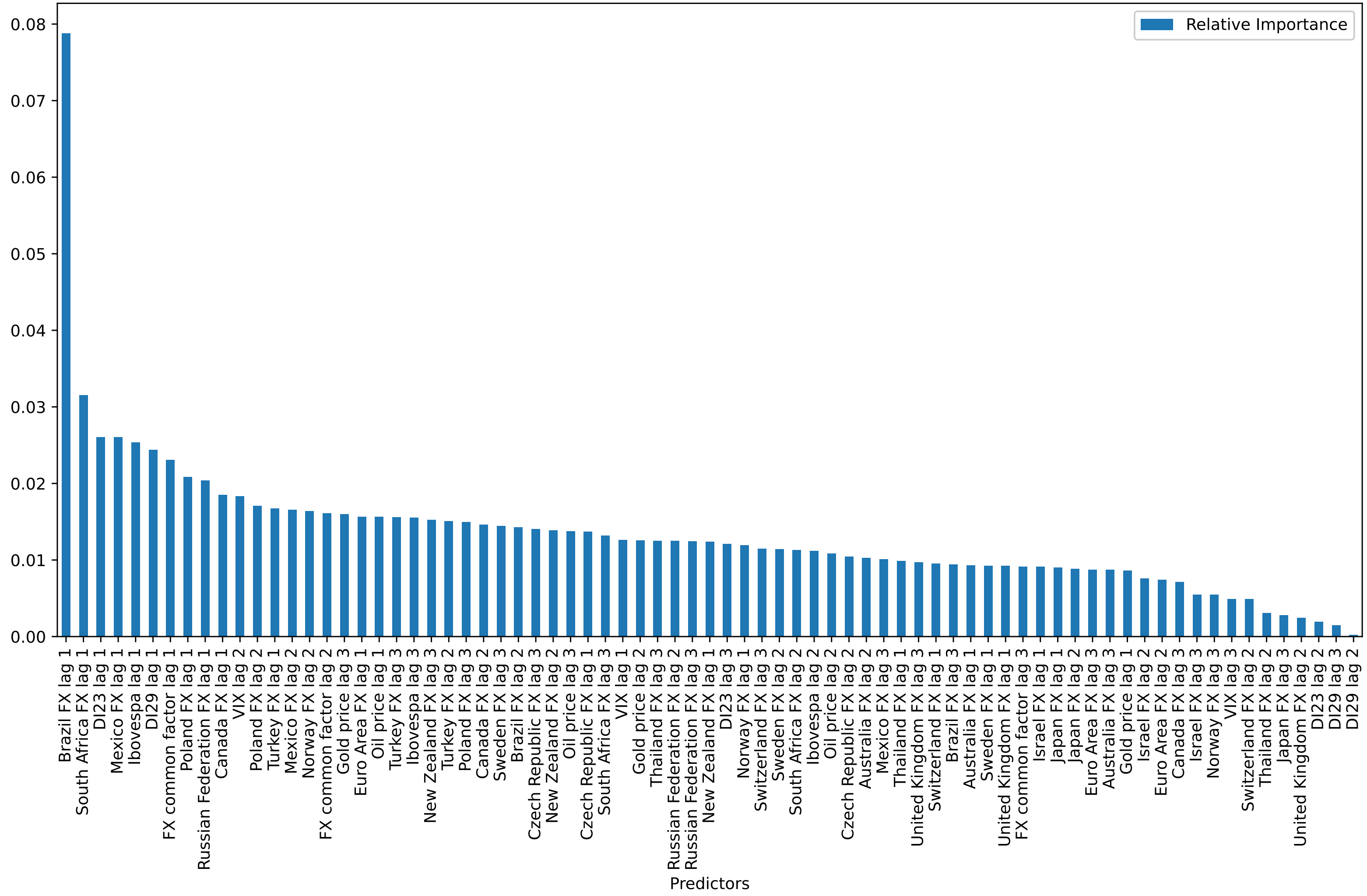
Models MSE ratios in relation to Random Walk without drift. In bold, the lowest MSE ratios for each frequency / horizon (h). All specifications also include the lags of the Brazilian real / dollar exchange rate itself. ***, **, and * indicate rejection of the null hypothesis (model and Random Walk predictions are not different), at the 1%, 5% and 10% levels respectively. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window. DI23 and D29 are Interbank Deposit interest rates futures contracts maturing in January 2023 and January 2029. The 17 currencies correspond to the foreign exchange rates listed in the section 3.2. FX common factor refers to the factor extracted from the set of foreign exchange rates by PCA.

For a frequency of 1 minute, a simple specification like an AR(1) model is capable of outperforming the Random Walk up to 2 minutes ahead. The OLS specifications, even with the addition of predictors, do not improve the result; on the contrary, they all show MSE ratios higher than that of the AR(1), indicating that some variables are adding more noise than signal to the models. However, both LASSO (GS) and LASSO (BIC) outperform the Random Walk up to 4 minutes ahead. XGBoost is able to surpass the Random Walk over a horizon of three minutes, considering a significance level of up to 10% in the Diebold-Mariano test. Thus, in the case of the out-of-sample exercise at a one-minute frequency, simpler Machine Learning models demonstrated the best results in terms of reducing the MSE. It is important to note that, in the case of LASSO (BIC), the determination of the hyperparameter was carried out at each step of the sliding training window, unlike with the decision tree models.

The margins, in comparison to a simple autoregressive model or even the Random Walk, are very narrow, especially for horizons greater than one step ahead. Therefore, when we apply the Model Confidence Set procedure to evaluate the set of superior models at a frequency of 1 minute, one step ahead, all the Machine Learning models, along with most of the OLS models, are classified in the same set as the AR(1). According to the MCS methodology, these models are considered similar in terms of performance within a 80% confidence interval. For horizons equal to or greater than 2 minutes, the set of superior models begins to include the Random Walk itself. Thus, no model is considered superior to the RW for these horizons according to the MCS methodology. The tables in Appendix C present these results.

Figure 9 presents the average importance assigned to each predictor by XGBoost in the out-of-sample forecasting exercise, conducted at a frequency of one minute for predicting one step ahead. As expected, given the performance of the autoregressive specification, the most important predictor is the first lag of the Brazilian exchange rate. The other significant predictors, similar to those observed in the out-of-sample fit exercise, include the first lags of the exchange rates of South Africa and Russia, both BRICS countries, ranking second and ninth respectively. Also notable are the first lags of short- and long-term interest rates (DI23 and DI29), and the Brazilian stock exchange index (Ibovespa), which are among the six most important predictors. This highlights the relevance of local information in predicting the exchange rate. Additionally, the first lag of the Mexican exchange rate holds the fourth position, similar result to the out-of-sample fit exercise.

Figure 9 – Importance of XGB predictors in out-of-sample forecasting exercise at 1-minute frequency, horizon=1 minute ahead.



For the frequencies of 5 and 15 minutes, we simply do not find predictability for one or more steps ahead (for simplicity, in Table 10 we only present the results for one step ahead for these frequencies). When we add variables to the linear regression model, we end up worsening the mean square error, indicated by MSE ratios greater than 1. The ML models do reduce the MSE ratios at these two frequencies, but without making the MSE of the models lower than that of the Random Walk. The LASSO (BIC) method even reaches a MSE ratio of 1 for the 5-minute frequency, but this result hides the real behavior of the algorithm: when analyzing the coefficients of the predictors, we find that all were zeroed in the vast majority of forecasts, except for the coefficient of the intercept. That is, the algorithm considers all predictors irrelevant for forecasting and makes its prediction based on the average of the values of the variable to be predicted, the log returns of the Brazilian exchange rate itself.

Inspired by Ventura and Garcia (2012), who demonstrated that in Brazil the exchange rate is firstly determined at the exchange rate future market and then transmitted by arbitrage to the spot market, we conducted a complementary exercise using as predictors the Brazilian Real/U.S. dollar futures contracts rates (USDBRL futures), along with the lagged value of the spot exchange rate itself in an OLS specification. The result is presented in Table 11. At the 1-minute frequency, considering a statistical significance level in the Diebold Mariano test of 1%, it is possible to beat the Random Walk with this simple specification up to the 5-minute horizon. Considering 10%, up to 6 minutes ahead. However, by the MCS procedure, the OLS model with the Real/U.S. dollar futures contracts rates as predictors is not considered superior to the other ML models from the previous exercise (those presented in Table 10). This is likely due to the narrow margin of error in which it is possible to beat the Random Walk.

Table 11 – Brazilian Real/U.S. dollar exchange rate forecasting exercise, out-of-sample results using the rates of Brazilian Real/U.S. dollar futures contracts as predictors.

		MSE ratio							
Predictors	**Method**	**frequency 1 min**						**f. 5 min**	**f. 15 min**
		h=1	*h=2*	*h=3*	*h=4*	*h=5*	*h=6*	*h=1*	*h=1*
-	RW	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
-	AR(1)	0.9960***	0.9976**	0.9985	0.9984	0.9991	0.9998	1.0014	1.0022
USDBRL futures	OLS	**0.9847*** **	**0.9911*** **	**0.9944*** **	**0.9957*** **	**0.9970*** **	**0.9983* **	1.0005	1.0025

Models MSE ratios in relation to Random Walk without drift. In bold, the lowest MSE ratios for each frequency / horizon (h). The OLS specification also includes the lags of the Brazilian real / dollar exchange rate itself. ***, **, and * indicate rejection of the null hypothesis (model and Random Walk predictions are not different), at the 1%, 5% and 10% levels respectively. USDBRL futures refer to the prices of Brazilian Real/U.S. dollar future contracts for the next maturity.

We also tested Machine Learning models with all the other variables as predictors, in addition to the Real/U.S. dollar futures contracts rates, but the results were not superior to the simple OLS model. This indicates that if we consider Brazilian Real/U.S. dollar futures contracts rates as predictors, the other variables are overshadowed. The Appendix D presents the importances assigned by the XGB algorithm in the forecasting exercise one minute ahead, where we can see that the average importance attributed to the first lag of the Real/U.S. dollar futures contracts rates is practically more than double the importance of the other predictors.

Finally, for the frequencies of 5 and 15 minutes, even with this new predictor, it was not possible to beat the Random Walk.

3.5 Conclusion

This article investigates the predictability of the Brazilian spot exchange rate at high frequencies (1, 5, and 15 minutes) using a set of predictor variables, including short and long-term Brazilian interest rates, the Brazilian stock market index, gold price, oil price, the VIX, and exchange rates of 17 other countries, along with a global factor extracted by PCA from these currencies. We conduct two types of forecasting exercises: out-of-sample fit, in which we use contemporaneous values of the predictors, and an out-of-sample forecasting, in which we attempt to predict the exchange rate at time $t + h$ using data available up to time t.

In the out-of-sample fit exercise, we demonstrate that if the variables used were predictable, it would be possible to forecast the exchange rate at all frequencies better than the Random Walk. Local variables exhibit better predictive power than global ones, as they incorporate local information that impacts the foreign exchange market. The lowest mean squared errors were obtained using Machine Learning techniques. In the out-of-sample forecasting exercise, we show that it is possible to beat the Random Walk within a horizon of up to 4 minutes for a 1-minute frequency, using ML algorithms, when considering the Diebold and Mariano test as a criterion to differentiate forecasts. However, the margins between the errors are much narrower than those observed in the out-of-sample fit exercise. For the frequencies of 5 and 15 minutes, it is not possible to beat the Random Walk for any considered horizon. This demonstrates how rapidly the exchange rate adjusts. When evaluating the importance of variables in the forecasting exercises, we found that local variables are among the most important, along with the exchange rates of BRICS countries or economies similar to Brazil, such as Mexico.

We are also able to outperform the Random Walk model in the out-of-sample exercise by incorporating Brazilian Real/U.S. dollar futures as predictors, specifically at a one-minute frequency and up to six minutes ahead. This phenomenon is likely attributed to the characteristics of the Brazilian foreign exchange market. The futures market exhibits much higher liquidity compared to the spot market, and as per Ventura and Garcia (2012), the exchange rate is primarily determined in the exchange rate futures market and subsequently transmitted to the spot market through arbitrage.

4 Measuring inequality using electronic payment data

4.1 Introduction

Economic inequality, which includes differences in income, wealth, and consumption, is a highly discussed issue in economics because of its significant impact on social welfare and political stability (Alesina; Tella; MacCulloch, 2004; Roe; Siegel, 2011). However, the task of measuring it presents significant challenges. Traditionally, inequality is quantified through sample surveys or census data, generally at national, regional, or state levels. These approaches, however, are not without problems: they are prone to errors, consume considerable public resources, and may underestimate inequality (Medeiros; Souza; Castro, 2015). Moreover, they are carried out infrequently. For instance, Brazil's demographic census, the sole survey capable of providing population data for municipal-level inequality calculations, is conducted only once every ten years, on average. This infrequent data collection can limit the effectiveness of proposing and evaluating public policies aimed at promoting equity[1]. It becomes impractical to assess the impact of a political cycle (4 years) on the inequality of Brazilian municipalities. Additionally, available data predominantly pertain to the income dimension. Data on consumption, which might offer a better measure of well-being (Hassett; Mathur, 2012; Meyer; Sullivan, 2009; Trapeznikova, 2019), are scarcer (Attanasio; Pistaferri, 2016).

In our research, we propose to measure municipal-level consumption inequality using a new database: the electronic payment methods from the Brazilian Central Bank's Payment System. This dataset includes credit card data and Pix transactions, the instant transfer and payment instrument that now exceeds more than 3.5 billion transactions per month. With this database, we were able to calculate the Gini index, a commonly used measure of inequality, which we consider a suitable metric for assessing consumption inequality in Brazilian municipalities. This approach allows for a timely analysis with up-to-date data[2]. The most recent available municipal inequality data, for example, are from the 2010 census. We can also mention as an advantage that the data is neither sample-based nor declarative, hence not subject to common survey errors[3]. For these

[1] Inequality is positively correlated with criminality and negatively correlated with income growth at the municipal level, and people living in more unequal municipalities classify themselves as less happy than those living in more egalitarian places (Glaeser; Resseger; Tobio, 2009).

[2] We chose the Gini index as a measure of inequality for comparative purposes with other studies and also because it is the most commonly used measure of inequality (Maio, 2007).

[3] Moore et al. (2000), Bee and Rothbaum (2019) review the literature on measurement issues in surveys, showing that, generally, survey respondents underreport income. Hokayem, Bollinger and Ziliak (2015) shows that populations with the lowest and highest incomes are those most likely not to respond

reasons, we believe that our measure of consumption inequality has the potential to support public policies, especially at the municipal level, that require more timely diagnoses and monitoring.

Despite the Electronic Payment data serving as a proxy for consumption, we compare the inequality calculated from our database with the income inequality calculated using data from the 2010 Brazilian Institute of Geography and Statistics (IBGE) census. Even with the substantial time lag, as inequality tends to exhibit some degree of persistence, we demonstrate a moderate correlation between our inequality index and income inequality measured from census data.

We also present an application for our inequality index: we explore the relationship between inequality and economic complexity, a concept that assesses the sophistication of economic activities in a given economy (Hidalgo; Hausmann, 2009), adapted to the municipal level in our case. This relationship, particularly concerning income inequality, has been the subject of recent debate when studied at the country level (Hidalgo, 2021; Hartmann et al., 2017; Lee; Vu, 2020; Chu; Hoang, 2020; Lee; Wang, 2021; Pham; Truong; Hoang, 2023; Amarante; Lanzilotta; Torres, 2023) and also at the regional level, albeit by a limited number of studies (Sbardella; Pugliese; Pietronero, 2017; Gao; Zhou, 2018; Török; Benedek; Gómez-Zaldívar, 2022; Morais; Swart; Jordaan, 2021). The literature's findings are still mixed, so we hope to contribute to the discussion. We calculated the Economic Complexity Index (ECI) for each Brazilian municipality and, through cross-sectional regressions, we show that ECI has a non-linear and negative relationship with consumption inequality: higher economic complexity is associated with lower inequality.

As contributions to the literature, to the best of our knowledge, we are the first to utilize an extensive electronic payment database to examine consumption inequality. We are also the first to measure consumption inequality at the municipal level in Brazil. Furthermore, while other works focus on investigating the relationship between income inequality and economic complexity, we are the first to investigate the relationship between consumption inequality and economic complexity. We are also the first to investigate the relationship between inequality-ECI at the municipal level in Brazil.

In addition to the availability of the database used, Brazil presents an intriguing case for such a study, as it represents a significant number of economies that can be classified as lower-middle to upper-middle income while simultaneously experiencing high levels of poverty and inequality (Morais; Swart; Jordaan, 2021). Additionally, Brazil is characterized by substantial regional economic disparities.

The paper is structured as follows. In Section 4.2, we provide a brief introduction

to income questionnaires. Burkhauser et al. (2018) uses administrative data (income tax data) to demonstrate that in the case of the United Kingdom, the increase in income inequality measured using survey data might be underestimated.

to the literature. In Section 4.3, we describe the data sources used for constructing the electronic payment Consumption Inequality Index, the Economic Complexity Index, and the other variables employed in our analysis. In Section 4.4, we present the methodologies for calculating the Gini index and the Economic Complexity Index. In Section 4.5, we explore the Consumption Inequality Index derived from electronic payment methods and its correlation with the Gini Income Inequality Index as reported by IBGE. In Section 4.6, we present our study of the Inequality-ECI relationship. Section 4.7 concludes.

4.2 Literature review

4.2.1 Electronic Payment Methods Data and Consumption Inequality

Our work aims to obtain a measure of inequality based on electronic payments made by individuals, which we consider as a proxy for consumption. While it is not a measure of income, which is the variable with better availability in advanced economies (Trapeznikova, 2019; Attanasio; Pistaferri, 2016), Aguiar and Bils (2015) show that consumption inequality has closely tracked income inequality in the United States. Consumption also can be considered a better measure of well-being than income, considering that savings and loans can be used to smooth consumption over time (Hassett; Mathur, 2012; Meyer; Sullivan, 2009; Trapeznikova, 2019). Furthermore, the joint assessment of income inequality and consumption inequality can be interesting, for example, in enabling the investigation of consumption smoothing mechanisms and the nature of income shocks (temporary or permanent) (Attanasio; Pistaferri, 2016).

In the United States, research on consumption inequality typically relies on data from the Consumer Expenditure Survey (CE), a microdata source that has been available since the 1980s, or the Panel Study of Income Dynamics (PSID), which since 1999 has covered approximately 70% to 90% of the expenditures collected by the CE (Attanasio; Pistaferri, 2014; Attanasio; Pistaferri, 2016). Another data source explored by researchers in the United States is the Residential Energy Consumption Survey (RECS), which enables the assessment of consumption inequality in durable goods (Hassett; Mathur, 2012). Similar data sources are available in other countries, such as the Chinese Residential Energy Consumption Survey (CRECS) used by Wu, Zheng and Wei (2017) to evaluate inequality in rural areas of China.

The literature using administrative data to study income inequality is well-established. Examples include Piketty and Saez (2003), Piketty, Saez and Zucman (2018), Larrimore, Mortenson and Splinter (2021). Regarding the use of administrative data to study consumption, studies are more scarce, and household consumption is generally determined indirectly. For example, Browning and Leth-Petersen (2003), Kolsrud, Landais and Spinnewijn (2017), Eika, Mogstad and Vestad (2020) use extensive administrative databases of income, taxes, and wealth and the following accounting identity to determine household consumption: the total household spending is equal to income plus capital gains minus the change in wealth over a certain period.

Regarding the use of electronic payment data specifically, studies have utilized credit card data (Gross; Souleles, 2002; Aydin, 2015) and financial aggregator data[4] (Gelman et al., 2014; Gelman et al., 2020; Baker; Yannelis, 2017; Baker, 2018; Olafsson; Pagel, 2018) to investigate consumption, although without assessing economic inequality. The database we used in our study has been explored by other authors in different contexts, such as by Gonçalves et al. (2022) in the area of nowcasting economic activity.

4.2.2 Economic Complexity Index

Since Kuznets (1955)[5], various works seek to establish a relationship between economic growth and inequality (Barro, 2008; Thomas, 2015; Galbraith, 2007; Palma, 2011; Deininger; Squire, 1996; Perera; Lee, 2013). However, conclusions appear to depend on theoretical preferences, the econometric methods employed, the economies under consideration, and the type of income distribution used (Dominicis; Florax; Groot, 2008). Furthermore, economic growth may only reflect a portion of economic development (Hartmann et al., 2017; Caous; Huarng, 2020), and the determinants of inequality are broader, encompassing a range of economic, social, institutional, historical trajectories, technological changes, and rates of return to capital (Chu; Hoang, 2020; Hartmann; Pinheiro, 2022). In this context, new measures of economic development are needed to capture some of these factors, and Hidalgo and Hausmann (2009)'s Economic Complexity Index may be one such measure (Hartmann et al., 2017; Hartmann; Pinheiro, 2022).

Hausmann et al. (2014) defines economic complexity based on the distribution and utilization of knowledge within a society. Products and services serve as means of transferring and integrating knowledge (Hidalgo; Hausmann, 2009). However, tacit knowledge, which is challenging to transfer, limits growth and development. The challenge of incorporating tacit knowledge leads to training in specific occupations and the specialization

[4] Web or mobile applications where users can link virtually any financial account, such as bank accounts and credit card accounts (Baker, 2018; Gelman et al., 2014; Gelman et al., 2020).

[5] Kuznets (1955) suggested that economic development, measured by the income level of an economy, is related to income inequality through an inverted U-shaped curve. The hypothesis is that in the early stages of development, there would be an increase in inequality as a transition from a rural to an industrial structure occurs. Urban-rural inequality increases in a scenario where the productivity of the agricultural sector is lower than that of the industrial sector, while entrepreneurial wages grow more rapidly than those of workers in urban centers, as these wages are pushed downward due to the injection of cheap labor from rural areas. At a certain stage of development, there is a significant movement of part of the workforce into new, higher-paying sectors, and there is also an increase in agricultural sector productivity, along with institutional transformations such as democratization, redistribution policies, and the establishment of a welfare state, which exert pressure for reduced inequality (Dominicis; Florax; Groot, 2008; Hartmann; Pinheiro, 2022; Sbardella; Pugliese; Pietronero, 2017; Soave; Gomes; Junior, 2019).

of organizations so that they can perform specific functions or tasks efficiently. Adapting to the expanding realm of knowledge involves distributing parts of that knowledge to individuals, and harnessing the diversity of this knowledge requires society to form organizations connected by intricate networks. In this way, the amount of productive knowledge utilized by an economy is mirrored in the diversity of firms, the range of necessary occupations, and the extent of interactions between them. The Economic Complexity Index (ECI) proposed by Hidalgo and Hausmann (2009) is a measure of how entwined this network of interactions is, i.e., how much productive knowledge is allocated by an economy (Hausmann et al., 2014).

The ECI measures the sophistication of a country, region, or municipality's productive structure by combining information about the diversity of products it exports or productive sectors it possesses and the ubiquity of these products or sectors (the number of countries, regions, or municipalities that export the product or possess a certain productive sector) (Hidalgo, 2021). Complex economies are those with high diversity and export products or possess productive sectors with low ubiquity, meaning they are more exclusive (only a few diverse economies are capable of producing these products or possessing sophisticated sectors (Hartmann et al., 2017)). Less complex economies are those capable of producing only a few products that are highly prevalent in the market. The Economic Complexity Index explores the interaction between entity diversity and product/sector ubiquity to measure the productive structure of an economy, incorporating information about the sophistication of its products/sectors[6].

4.2.3 Economic Complexity Index and Inequality

How would economic complexity affect inequality? Advocates of a negative relationship between the ECI and inequality argue that economies with a greater variety and sophistication of products or sectors tend to offer better occupational opportunities and upward mobility in social stratification, more inclusive institutions, a more equitable distribution of political power, and greater bargaining power for workers — forces capable of reducing economic inequality (Hartmann et al., 2017; Hartmann, 2014; Constantine; Khemraj, 2019). Arif (2021) demonstrates that an economy's sophistication leads to a higher labor share, which serves as a mechanism for inequality reduction. Given the need for physical capital, human capital, and technological advancements, the production of complex products results in increased demand for skilled workers, labor productivity, and wages proportional to labor efficiency. As labor inherently embodies tacit productive

[6] Products are distinguished by the amount of resources required for their production. The more diverse capabilities needed to manufacture a good, the more sophisticated it is considered. This complexity also applies to the economy as a whole, which becomes more advanced as it produces a greater variety of sophisticated goods. In other words, product sophistication and economic complexity are driven by the variety and quantity of capabilities available in a given locality or country (Hausmann; Hidalgo, 2011).

knowledge, workers' bargaining power increases, subsequently allowing them to secure better wages and thereby increase their share of total product participation (Arif, 2021).

Furthermore, as an economy becomes more sophisticated, a broader spectrum of densely interconnected products promotes an increase in productive interactions across various sectors. This demands a more diversified labor force with broader skills and multiple levels of expertise. From a flatter occupational structure characterized by a higher number of job positions and learning opportunities, less specialized workers may gain advantages over more specialized workers, contributing to a reduction in inequality (Pham; Truong; Hoang, 2023). More diversified economies would also ensure better long-term business sustainability in the face of volatility or crises, maintaining employability and wages at all levels, thus preventing an increase in inequality (Chu; Hoang, 2020).

On the other hand, in less sophisticated and diversified economies that heavily rely on natural resources, the income of the majority of workers depends on economic activities with diminishing returns to scale and low productivity. These individuals also face learning constraints and occupational limitations. Only a small portion of the population ends up enjoying higher income from more productive (yet limited) activities, as well as the knowledge and skills that remain confined within these groups (Lee; Vu, 2020).

Hartmann et al. (2017) are the first to study the relationship between economic complexity and inequality, providing support for a negative relationship between ECI and income inequality. They utilize economic complexity indices from the Observatory of Economic Complexity[7] at MIT to explain inequality, measured by the Gini index, for over 70 countries in a cross-sectional regression. They control for per capita GDP (and its square), education, population, and variables representing country institutions. In all models tested, the Economic Complexity Index was a negative and significant predictor of inequality. This result holds when they perform a fixed-effects panel estimation using data from 1962 to 2012: economic complexity reduces inequality. Lee and Vu (2020) arrives at a similar result when conducting cross-sectional OLS regression estimates for 96 countries, using data averages from 1980 to 2014 and similar controls to those used by Hartmann et al. (2017).

There are also hypotheses on how increased complexity could lead to greater inequality. Greater complexity creates a higher demand for skilled workers as new sectors emerge, replacing or rendering traditional sectors obsolete. While retraining is possible for low-income or low-skilled workers, it is easier and less costly for skilled workers to advance, as they have a greater capacity to adapt to changes, thereby widening income inequality (Lee; Vu, 2020; Chu; Hoang, 2020; Violante, 2008; Pham; Truong; Hoang, 2023). Automation can also play a significant role, making medium or low-skilled jobs obsolete (Sebastian; Biagi, 2018). A process of "deindustrialization" may also occur as

[7] <atlas.media.mit.edu>

the economy becomes more complex: it specializes in sophisticated products and replaces in-house manufacturing with imports for resource-intensive or medium- to low-skilled labor-intensive products. Thus, with part of the workforce unable to qualify for higher-skilled jobs, they end up being reallocated to lower-income positions in other sectors, thereby increasing inequality (Pham; Truong; Hoang, 2023; Violante, 2008).

Lee and Vu (2020) finds a positive relationship between economic complexity and inequality when estimating the relationship using system-GMM. Chu and Hoang (2020) also finds a positive relationship between economic complexity and inequality using 2SLS and system-GMM for 88 countries from 2002 to 2017. However, when interacting the ECI with socioeconomic variables, the authors show that at certain levels of education, more efficient government spending, and trade openness, increased complexity can act to reduce inequality. Lee and Wang (2021) also finds a positive relationship between economic complexity and inequality using fixed-effects panel strategies for their complete sample of 43 countries from 1991 to 2016. However, when dividing the sample into two subgroups, high-income countries and others, they demonstrate that economic complexity reduces inequality in the former group and increases it in the latter.

Given the favorable and unfavorable hypotheses about the impact of economic complexity on reducing inequality, other studies find nonlinear relationships between these two variables. Pham, Truong and Hoang (2023) identifies a U-shaped relationship between inequality and complexity in a system-GMM estimation for 99 countries from 2002 to 2016: initially, an increase in complexity would reduce inequality, but with an inversion beyond a certain level of complexity. The authors' hypothesis for the increase in inequality beyond a certain complexity threshold is the effect of deindustrialization (destruction of low and medium complexity jobs in manufacturing). This conclusion contradicts that of Amarante, Lanzilotta and Torres (2023), who, in a fixed-effects panel estimation for 126 countries from 1995 to 2018, finds an inverted U-shaped relationship: when economic complexity is low, increases in the sophistication of an economy's productive structure would increase inequality, and beyond a certain threshold, an increase in complexity would reduce inequality. The authors note that it is possible that the negative relationship between complexity and inequality evidenced in previous studies may be driven by the group of high-income countries that have already reached this threshold.

In general, disparities among the results of empirical studies may depend on the country sample, the periods considered, and the estimation methods employed. Thus, at the country level, the question of the relationship between economic complexity and inequality remains open.

4.2.4 Economic Complexity Index and Regional Income Inequality

Does the relationship between economic complexity and inequality at the regional level follow the same dynamics as observed at the country level? Sbardella, Pugliese and Pietronero (2017) assess the relationship between economic complexity and wage inequality among countries and also among counties within the United States. They use an alternative measure of Economic Complexity, called Fitness, developed by Tacchella et al. (2012). To construct the economic complexity index at the county level, they use employment data by economic activity sector instead of exported products. They find a positive relationship between economic complexity and wage inequality for U.S. counties in a cross-sectional assessment, in contrast to an inverted U-shaped curve when assessing countries, showing that the relationship is not scale-invariant. According to the authors, the fact that institutions are relatively homogeneous in the United States explains the difference in the complexity-inequality relationship between the two approaches.

In their study, Gao and Zhou (2018) used data from 2690 firms to calculate the ECI for 31 Chinese provinces, based on a "Province-Industry" network[8], where the number of firms in each of 70 categories for each province is considered instead of exported products. In a bivariate analysis, they found a negative relationship between ECI and income inequality for the provinces analyzed. Török, Benedek and Gómez-Zaldívar (2022), in cross-sectional and fixed-effects panel estimations for counties in Romania, using data from 2008 to 2018, also found a negative relationship between ECI and inequality.

Morais, Swart and Jordaan (2021) analyze the relationship between ECI and inequality using panel data for Brazilian states, employing Pooled OLS and Random Effects, with data spanning from 2002 to 2014. They find an inverted U-shaped relationship. One hypothesis for this relationship is that in the early stages, an increase in economic complexity benefits capital owners and high-skilled workers, leading to an increase in inequality. Beyond a certain level of economic complexity, other components of economic complexity become more important, such as institutions, labor unions, job opportunities, among others, which act to reduce inequality.

4.3 Data

4.3.1 Consumption Inequality

To measure consumption inequality, we utilized electronic payment data from the Brazilian Central Bank's Payment System[9], specifically data from the Pix and Credit Card

[8] As will be seen in section 4.4.2, for regional ECI calculation, each element of the matrix $\mathbf{M}$, $M_{p,i}$, receives the value of 1 if province p has revealed comparative advantage in industry i and 0 otherwise.

[9] A Payment System is a set of instruments, rules, procedures, and technologies used to settle money transfers between economic agents (Aprigliano; Ardizzi; Monteforte, 2019).

payment instruments. The Instant Payment System, launched in November 2020, enables 24/7 settlement of instant payments (Pix) and has spurred the creation of numerous new financial applications, such as QR code-payable invoices. The credit card data refers to the outstanding balance of individuals over the month, which includes spot purchases made during the month, plus installments of on-credit purchases due in the period[10].

Credit card payments are already used as a proxy for consumption, while Pix has become a standard for spot purchases in Brazil. For this reason, we believe that by using payments made through both instruments, we construct a good proxy for consumption[11].

We extracted payment data over the year 2021 for each payment instrument at the individual level. We excluded all payments sent by companies since we want to measure the inequality of people (and not companies). Additionally, we excluded transactions made from and to the same person since they represent a simple fund transfer between financial institutions. We also excluded Pix payments made to institutions belonging to activity groups that we consider unlikely to receive transactions related to consumption. The list of CNAE codes of these institutions is in Appendix I. Next, we aggregated payments sent to other individuals (outgoing payments) since it reflects a consumption intent, as opposed to payments received from other individuals (income payments). Finally, we divided the values by 12 to obtain a consumption monthly average for each individual. Using data from the Brazilian Federal Revenue to determine the municipality of residence for each, we calculated the Gini inequality index of consumption as described in section 4.4.1.

4.3.2 Economic Complexity Index

To calculate the municipal Economic Complexity Index (ECI), we utilized employment data by municipality, classified according to the National Classification of Economic Activities (CNAE) 2.0, as presented in the Annual Social Information List (RAIS). The data were obtained from the Ministry of Labor and Employment[12].

[10] It's important to note that these figures may not exactly match an individual's monthly credit card statement because the due date for the statement may differ, and payments may occur in the same month or the following month. As a result, within the population, some captured values will be higher than the actual monthly statement amount, while others will be lower. On average, though, we consider this data a reasonable representation of credit card consumption.

[11] In Appendix F, we show the correlation between the average annual consumption calculated based on electronic payment methods, base year 2021, and the average annual income estimated by Neri and Hecksher (2023), base year 2020, by municipality. We found a correlation that can be considered moderate to high (0.75), which would be expected if we have a good proxy for consumption. It is important to note, however, that the income calculated by Neri and Hecksher (2023) only considers tax return data, meaning it does not capture information from those who do not file tax returns, which is a considerable portion of the population.

[12] Available at <http://pdet.mte.gov.br/acesso-online-as-bases-de-dados>.

4.3.3 Control Variables of the Application

As regression controls for assessing the relationship between consumption inequality and the ECI, we obtained the GDP per capita (base year 2020) and the population of each municipality (base year 2021) from the IBGE, as well as the distance from each municipality to the nearest hub city[13]. As a measure of human capital, we used a proxy for the quality of education based on the age-grade distortion rate, i.e., the percentage of students in the correct grade for their age in high school. These data are obtained from the Institute of Applied Economic Research (IPEA), base year 2021. As a measure related to institutions, we created a variable with the percentage of coverage of legislation and planning instruments for each municipality[14], based on information provided by the IBGE for Brazilian municipalities, for the year 2021[15].

4.4 Methods

4.4.1 Calculation of the Inequality Index

To measure consumption inequality in each municipality using electronic payment methods, we employ the Gini index[16]. Let the population percentages be ordered from the poorest (or lowest consumption) to the richest (or highest consumption) on the horizontal axis of the graph in Figure 10, and on the vertical axis, the cumulative proportion of income (or consumption) held by the population. Let the line of perfect equality be the diagonal line where each percentage of the population has an equal share of income or consumption, and the Lorenz curve be the actual income or consumption accumulation curve. The Gini index is given by the ratio of the area between the line of perfect equality and the Lorenz curve (area A in Figure 10) to the total area below the line of perfect equality (area $A + B$ in Figure 10).

To compute the Gini index in a direct and efficient manner (Cowell, 2011), we order all incomes or consumptions from lowest to highest, $y_{(1)}, y_{(2)}, ..., y_{(n)}$ (meaning $y_{(1)}$ is the lowest income/consumption, $y_{(2)}$ is the next, and so on, up to person n), and we calculate:

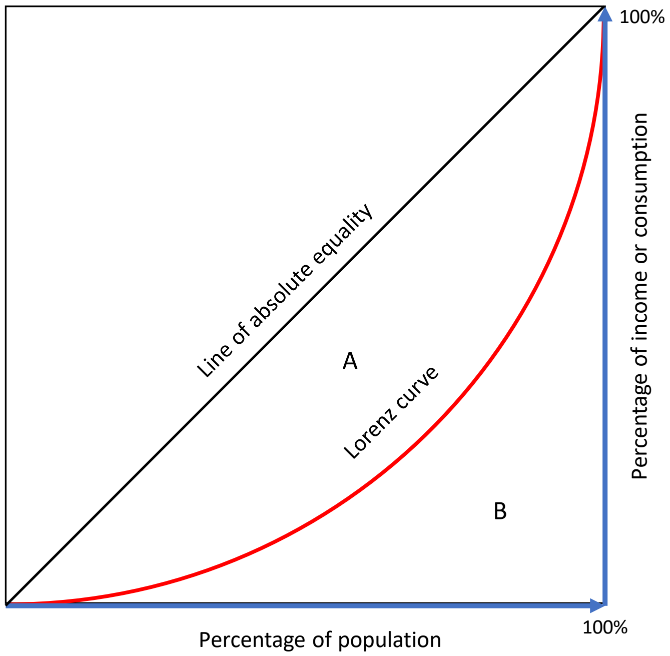
[13] For hub city, we refer to municipalities classified as metropolises or regional capitals by the IBGE. The list of these municipalities is available at <https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15798-regioes-de-influencia-das-cidades.html>.

[14] Quantity of legislation and planning instruments existing in relation to the total expected by the IBGE Basic Municipal Information Survey, listed in A.

[15] Given the limitation of available data, we faced difficulties in creating measures of institutions for Brazilian municipalities. The measure we propose here is an attempt to partially measure Formal Institutions based on the concept from Alston et al. (2016). However, our measure presents a number of limitations, as we have outlined in the section 4.6.3.

[16] For the origin of the Gini index, see Ceriani and Verme (2012).

Figure 10 – Lorenz curve.


100%
Percentage of income or consumption
Line of absolute equality
A
Lorenz curve
B
Percentage of population
100%

$$G = \frac{2}{n^2\overline{y}}\left[y_{(1)} + 2y_{(2)} + 3y_{(3)} + ... + ny_{(n)}\right] - \frac{(n+1)}{n} \tag{4.1}$$

Where:

$$\overline{y} = \sum_{i=1}^{n} \frac{y_{(i)}}{n} \tag{4.2}$$

This Gini index calculation method was implemented in SQL, directly in the database query, due to efficiency and computational cost considerations. The SQL script can be consulted in H.

4.4.2 Calculation of the Economic Complexity Index

Following Hidalgo and Hausmann (2009), Hausmann et al. (2014), Kemp-Benedict (2014), we define the Revealed Comparative Advantage (RCA) of a country c in a product p as:

$$RCA_{cp} = \frac{X_{cp} / \sum_{p'} X_{cp'}}{\sum_{c'} X_{c'p} / \sum_{c'p'} X_{c'p'}} \tag{4.3}$$

In which X_{cp} represents the total exports of product p from country c, $\sum_{p'} X_{cp'}$ denotes the total export portfolio of country c, $\sum_{c'} X_{c'p}$ signifies the total exports of product p by all countries, and $\sum_{c'p'} X_{c'p'}$ represents the total exports of all products by all countries. RCA_{cp} will be greater than 1 if the export of product p is higher than expected given the size of country c's export economy and the global market for the product, indicating that the country has a comparative advantage in the product (Hartmann et al., 2017).

Let $\mathbf{M}$ be a country-product matrix with elements M_{cp} indexed by country c and product p. We define each element of M_{cp} as:

$$M_{cp} = 1 \; if \; RCA_{cp} >= 1 \tag{4.4}$$

$$M_{cp} = 0 \; if \; RCA_{cp} < 1 \tag{4.5}$$

We then define the diversity of a country c and the ubiquity of a product p as:

$$Diversity = k_{c,0} = \sum_p M_{cp} \tag{4.6}$$

$$Ubiquity = k_{p,0} = \sum_c M_{cp} \tag{4.7}$$

That is, we measure diversity and ubiquity by summing over the rows or columns of the matrix $\mathbf{M}$ while fixing the country or product, respectively. However, these indicators separately are imprecise measures of economic complexity. For countries, we need to calculate the average ubiquity of the products they export, the average diversity of the countries producing these products, and so on. For products, we must calculate the average diversity of the countries producing them, the average ubiquity of the other products these countries produce, and so forth[17]. Therefore, the economic complexity proposed by Hidalgo and Hausmann (2009) jointly and interactively computes the average values of diversity and ubiquity using the so-called method of reflections, where:

$$k_{c,N} = \frac{1}{k_{c,0}} \sum_p M_{cp} \cdot k_{p,N-1} \tag{4.8}$$

$$k_{p,N} = \frac{1}{k_{p,0}} \sum_c M_{cp} \cdot k_{c,N-1} \tag{4.9}$$

Substituting 4.9 into 4.8, we arrive at:

[17] For example, diamonds have low ubiquity but are generally produced by countries with low diversification, indicating a low requirement for productive knowledge. On the other hand, medical imaging devices have low ubiquity and are generally produced by countries with high diversification, indicating a high requirement for productive knowledge. Thus, a correction is necessary to ensure that uncommon products are considered truly complex only if they are produced by diversified countries. Similarly, relatively common products can also be considered complex if their production is limited to a group of diversified countries (Sousa, 2018).

$$k_{c,N} = \sum_{c'} \left(\frac{1}{k_{c,0}} \sum_{p} M_{cp} \frac{1}{k_{p,0}} M_{c'p} \right) k_{c',N-2} \tag{4.10}$$

Which can be written in matrix form as:

$$\vec{k}_N = \mathbf{W}\vec{k}_{N-2} \tag{4.11}$$

Where $\vec{k}_N$ represents the set of values for countries $k_{c,N}$, and the matrix $\mathbf{W}$ has elements:

$$W_{cc'} = \frac{1}{k_{c,0}} \sum_{p} M_{cp} \frac{1}{k_{p,0}} M_{c'p} \tag{4.12}$$

The Economic Complexity Index for the set of countries is then defined as the eigenvector of **W** associated with the second largest eigenvalue (Kemp-Benedict, 2014; Hausmann et al., 2014).

In our work, we adopt the methodology adapted by Brandão (2023) and Sbardella, Pugliese and Pietronero (2017) from Hidalgo and Hausmann (2009) to analyze municipalities instead of countries. Given the geographical focus of our study, we chose to use the number of employment links as a proxy for the size of each of the 285 activity groups in the CNAE 2.0, rather than relying on product export data. In this context, the number of employment links in each sector within a municipality serves a role analogous to the export value of products for each country in Hidalgo and Hausmann (2009)'s original model. Therefore, the economic complexity of a municipality is assessed based on the diversity and ubiquity of its productive sectors[18]. We used the R package *economiccomplexity* to calculate the ECI for each municipality.

4.5 Consumption Inequality Gini Index

Figure 11 shows the distribution of the Gini inequality index calculated using electronic payment methods across Brazilian municipalities. Table 12 displays the descriptive statistics. Figure 19 illustrates the heat map of municipalities according to their consumption inequality Gini index.

[18] According to Brandão (2023), the use of labor market data for regional ECI analysis is more appropriate because:

1. Labor market data encompass a broader range of economic activities than a municipality's export portfolio, including the services sector;
2. International trade data cover only external trade and do not account for the trade of products within municipalities themselves;
3. Due to bureaucratic or administrative reasons, a product's origin may be in one city but is often recorded in international trade data under another city.

Figure 11 – Consumption inequality Gini coefficients histogram.

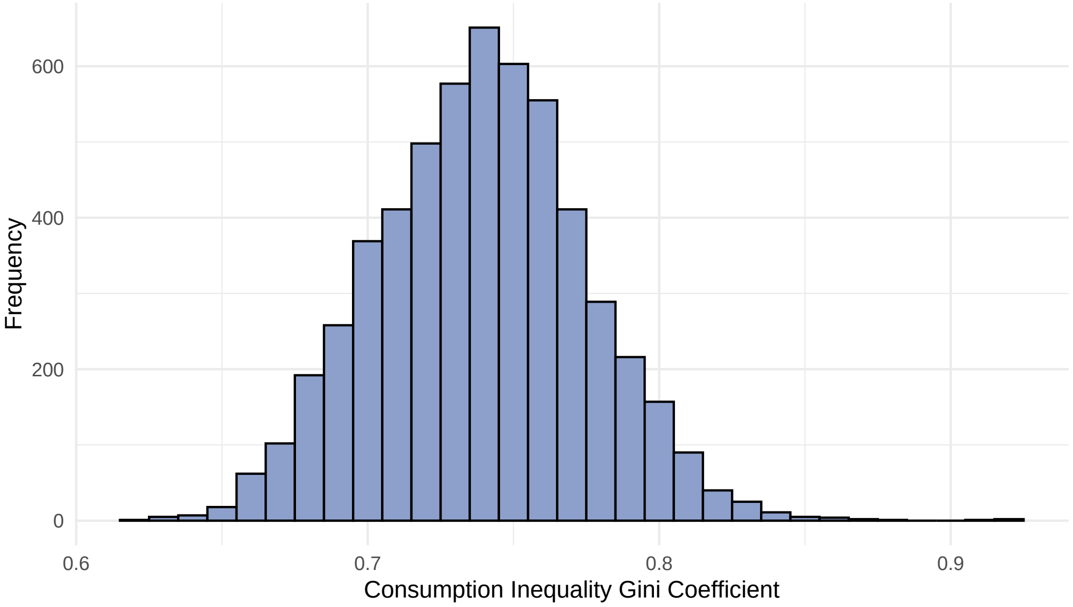


Table 12 – Summary statistics of municipal consumption inequality Gini coefficients.

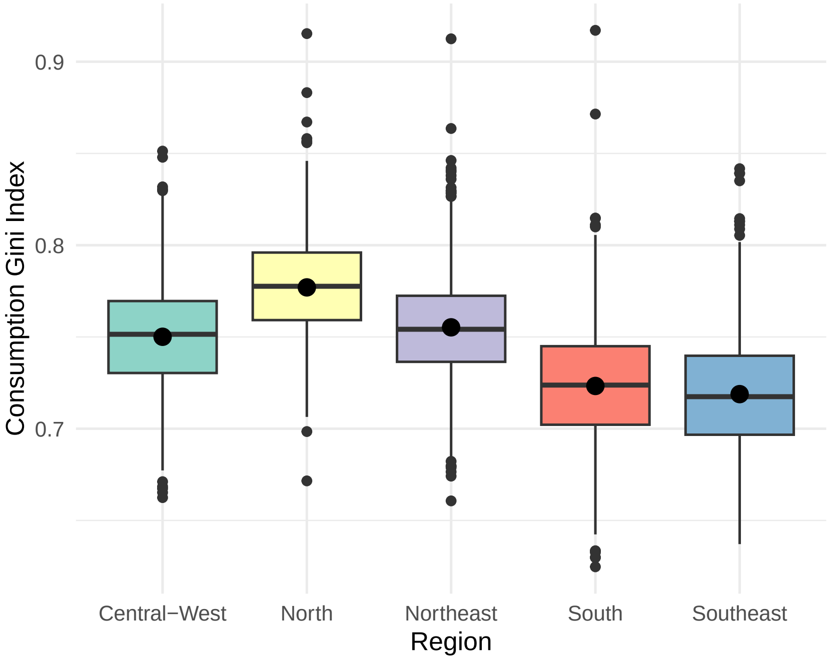
	N	**Min**	q_1	**Median**	**Mean**	q_3	**Max**	**St. Dev**
Gini Coefficients	5563	0.625	0.714	0.739	0.739	0.762	0.917	0.036

Figure 12 displays the regional boxplots: the northern region has the highest municipal average of the consumption Gini index, while the South and Southeast regions are similar and have the lowest averages. This ranking is similar to the one found in the evaluation of the Gini index boxplots calculated from income according to the 2010 IBGE census, shown in Figure 13. In Appendix G, we present the boxplots by state, both for our consumption inequality index and for the one calculated based on data from the 2010 IBGE census.

Although our index is calculated based on data we consider a proxy for consumption, we expect it to be correlated with the income inequality measured by the 2010 IBGE census data, even with a lag of over a decade between the databases, given that inequality tends to show a certain persistence. Figure 14 shows the Pearson correlation between both inequality indices. There is also a distinction of municipalities by region, where we can see the patterns established in Figures 12 and 13: municipalities from the southern and southeastern regions are concentrated in the lower left corner of the chart, indicating lower inequality on average, while municipalities from the northern and northeastern regions are mostly in the upper right corner, indicating higher inequality.

We might find slightly higher correlations between the two indices if we exclude very small municipalities. Figure 15 shows the correlations for cases where we filter the

Figure 12 – Boxplots of municipal consumption Gini coefficients highlighting variability among regions.



municipalities by a minimum population threshold. For instance, if we consider only municipalities with more than 25,000 inhabitants, the Pearson correlation between the indices could be higher than 0.55. However, as most Brazilian municipalities are small, if we exclude municipalities with less than 25,000 inhabitants, we reduce the number of

Figure 13 – Boxplots of municipal IBGE 2010 income Gini coefficients highlighting variability among regions.

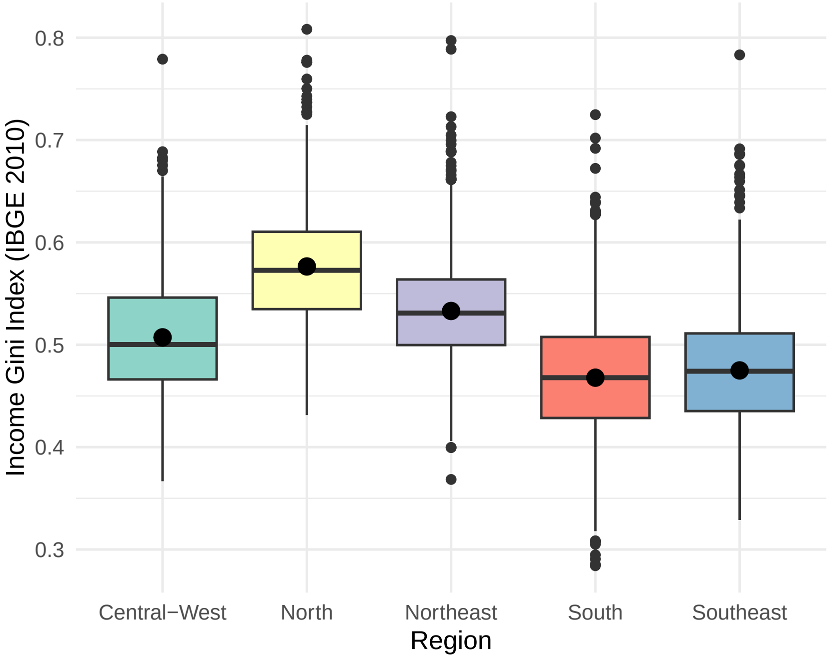
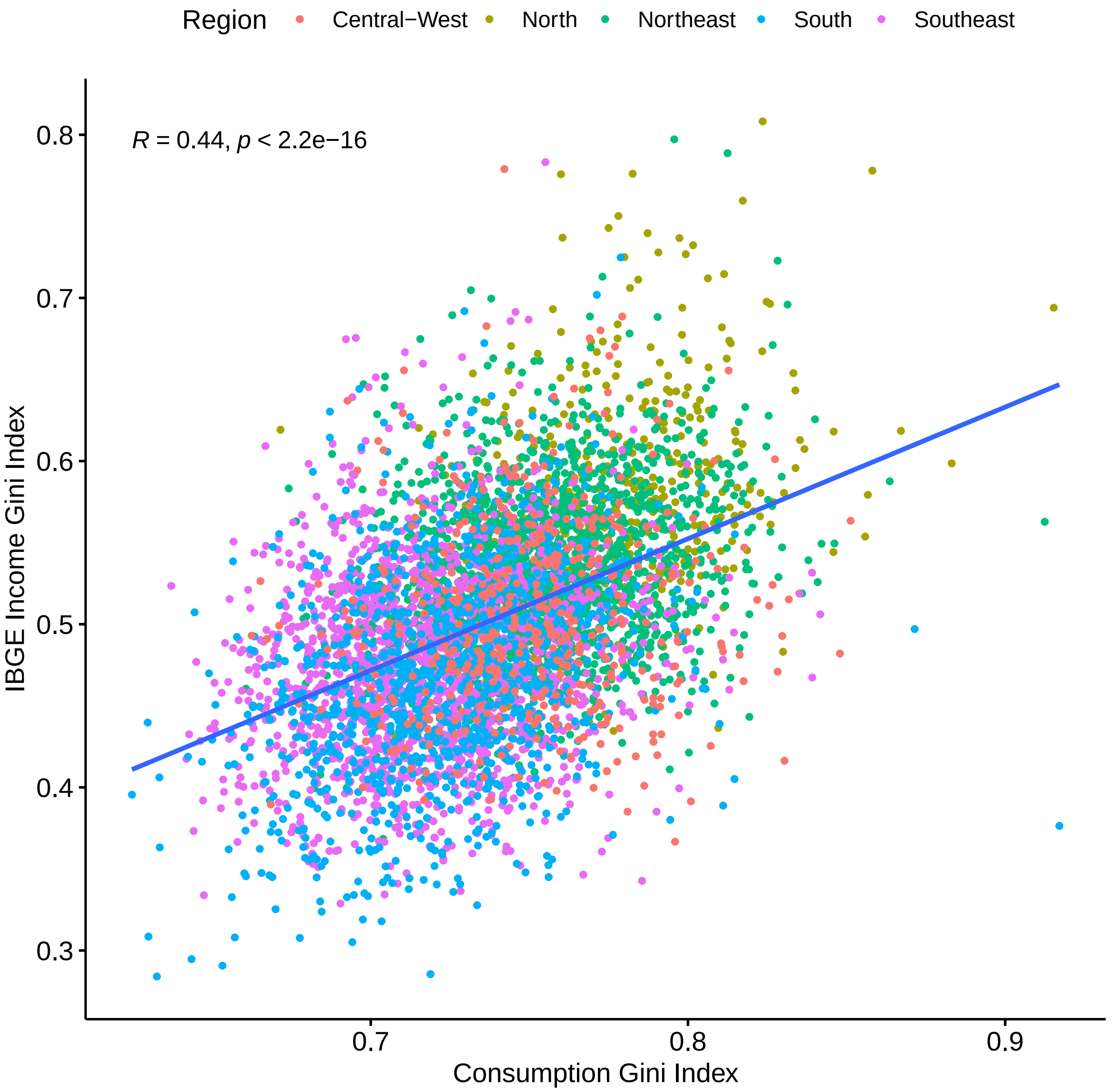


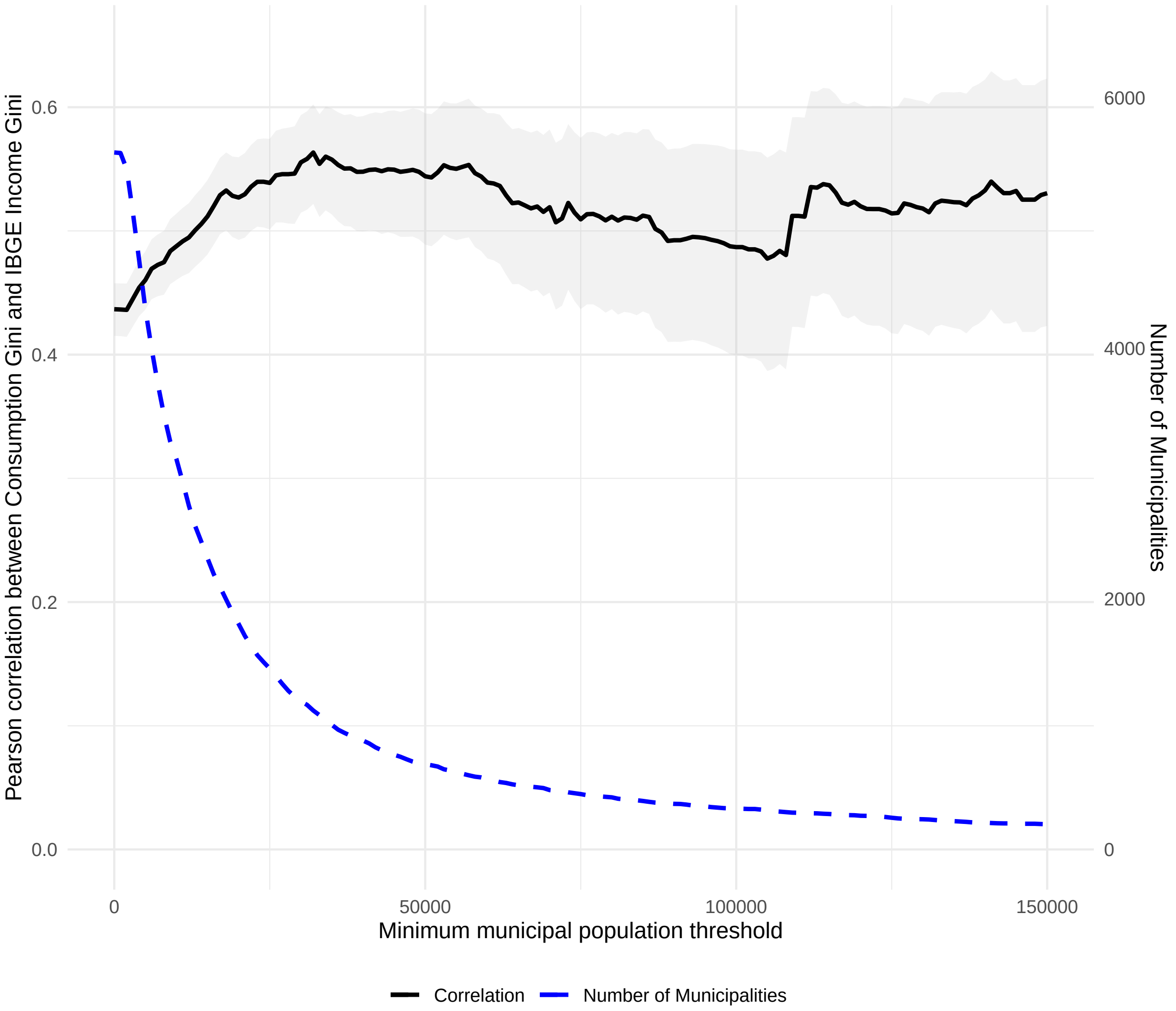
Figure 14 – Scatter plot comparing consumption and IBGE income Gini coefficients with Pearson Correlation (R), differentiated by region.



municipalities considered by more than 2/3.

We also note that the IBGE Gini index generally indicates less inequality for Brazilian municipalities. Our hypothesis regarding the difference in inequality is that, in the distribution of Electronic Payment transactions, the sum of transactions per person at the tail end of the poorest is lower, on average, than the household incomes of the poorest captured by the IBGE. It is possible that this population conducts more cash transactions than the higher-income population, which we do not capture in our database. Additionally, it is likely that higher incomes are under-reported in household surveys, but do not escape electronic transactions. Another factor is that the IBGE's Gini index is calculated based on per capita household income, which sums and then divides the income of all residents

Figure 15 – Evolution of the Pearson Correlation between consumption and IBGE income Gini coefficients as a function of an increasing sample's minimum municipal population threshold, with confidence intervals. The dashed blue line represents the number of municipalities considered in the sample for a given minimum municipal population threshold, as indicated by the secondary y-axis.

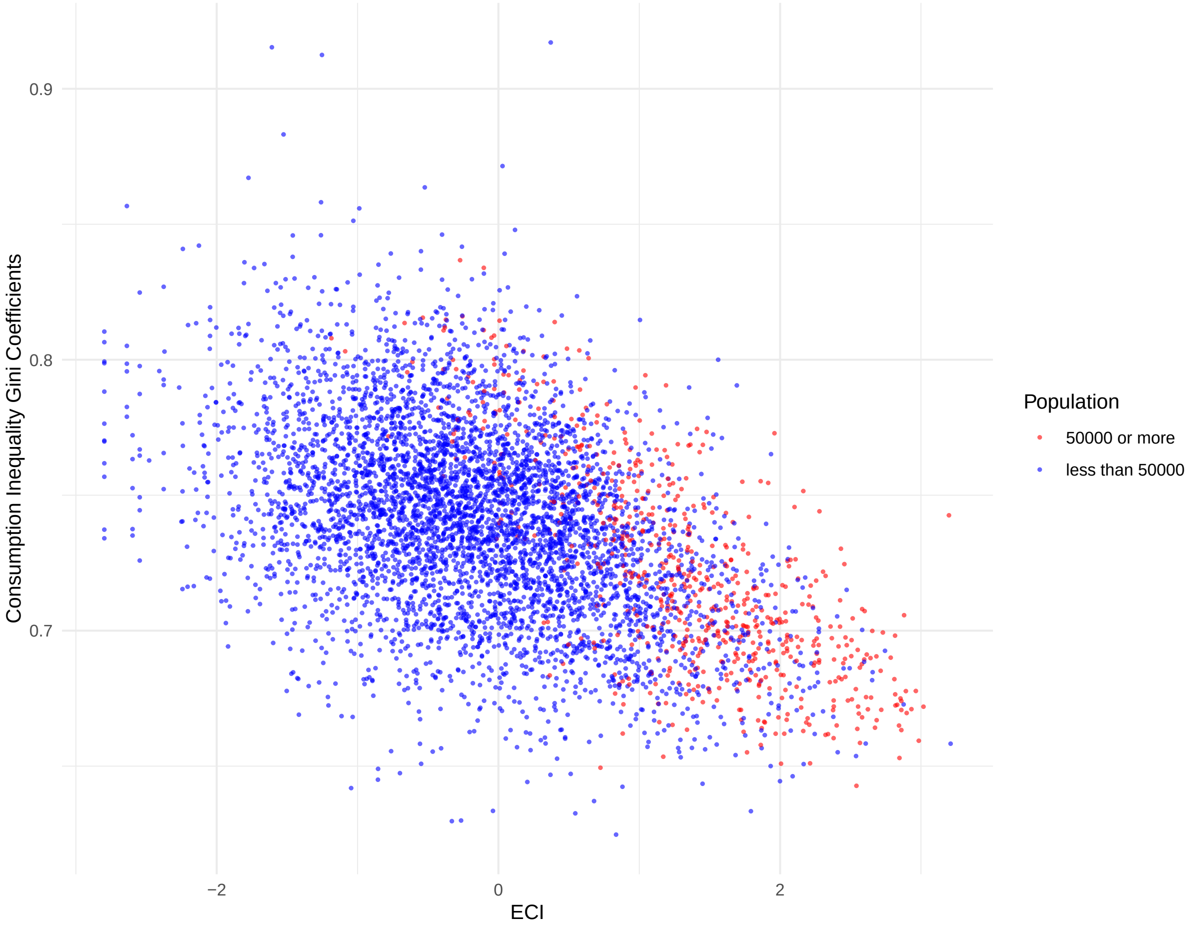


by the number of residents in each household. This likely eliminates some extremes in the income distribution, compared to our data distribution, where we consider the payments made by each individual, regardless of their residence. Given these factors, our index tends to show greater inequality than the Gini index calculated by the IBGE. Conversely, we have an inequality index that can be calculated promptly and as frequently as desired.

4.6 Application - Inequality vs. Economic Complexity Index

In this section, we explore an application for our Gini index of consumption inequality based on electronic payment methods: similarly to Hartmann et al. (2017), we assess the relationship between inequality (in our case, consumption inequality) and the

Figure 16 – Scatter plot of the relationship between the consumption inequality Gini index and the Economic Complexity Index for Brazilian municipalities.



Economic Complexity Index at the municipal level.

4.6.1 Model

Figure 16 shows the scatter plot depicting the relationship between our Gini index of consumption inequality and the Economic Complexity Index for Brazilian municipalities. The relationship does not appear to be completely linear at first glance, with an initial indication that higher economic complexity might be associated with lower inequality. We can also observe that larger municipalities tend to have a higher ECI, on average.

We then proceed to a multivariate analysis using cross-sectional data for Brazilian municipalities, where we estimate the coefficients of Equation 4.13 using Ordinary Least Squares (OLS):

$$Gini_i = \beta_0 + \beta_1 ECI_i + \beta_2 ECI_i^2 + \boldsymbol{\beta}_3 \boldsymbol{X}_i + \varepsilon_i \tag{4.13}$$

For each municipality i, $Gini_i$ is the Gini index of consumption inequality calculated based on electronic payment methods; ECI_i the Economic Complexity Index and ECI_i^2 its square, aiming to capture a possible nonlinear relationship, as in Pham, Truong and Hoang (2023), Amarante, Lanzilotta and Torres (2023), Morais, Swart and Jordaan (2021); $\boldsymbol{X}_i$ a vector of control variables, including the natural logarithm of per capita GDP and its square; the natural logarithm of the population; an education measure which is the rate of students in the correct grade for their age; the natural logarithm of the distance from the municipality to the nearest hub city[19]; as an institutional measure the percentage of legislation implemented relative to that planned in the IBGE survey on Legislation and Planning Instruments.

We estimate Equation 4.13 with and without the squared term ECI_i for all Brazilian municipalities included in our database[20]. We also estimate Equation 4.13 for each Brazilian region, with the aim of verifying whether the results remain consistent in each region.

4.6.2 Results

Table 13 shows the regressions results for specifications with and without the squared ECI term. We can observe that the ECI coefficient is negative and statistically significant at 1% in both the first and second specifications, and its squared term is also negative and statistically significant at 1% in the second specification. This indicates a negative and nonlinear relationship between consumption inequality and the ECI, meaning that higher economic complexity is associated with lower inequality, and the greater the economic complexity, the larger the impact of ECI changes on inequality.

The negative relationship between inequality and the ECI is in line with the findings of Hartmann et al. (2017) and Lee and Vu (2020) (in OLS regressions) at the country level, and at the regional level, it aligns with the results of Gao and Zhou (2018) and Török, Benedek and Gómez-Zaldívar (2022), and is contrary to that of Sbardella, Pugliese and Pietronero (2017).

If the relationship between economic inequality and the Economic Complexity Index is the same for both income inequality and consumption inequality, our result does not support the hypothesis of Hartmann and Pinheiro (2022) regarding the reversal of the relationship, from negative to positive, when scaling down the assessment to the regional level. Sbardella, Pugliese and Pietronero (2017) posits that the difference in results when assessing the relationship at the country level versus at the municipal level is due to the fact that institutions are relatively homogeneous within the United States. It might be assumed that within Brazil, institutions are not as homogeneous as in the United States. However, this does not seem to explain the difference in the sign of the ECI coefficient

[19] For municipalities that are hubs, i.e., with zero distance, a distance of "1" Km was imputed to enable the use of the natural logarithm.

[20] We exclude municipalities that have missing data for any of the considered variables.

Table 13 – OLS regressions for all municipalities.

	Dependent variable:	
	Consumption Inequality Gini Index	
	(1)	(2)
ECI	-0.012^{***}	-0.013^{***}
	(0.001)	(0.001)
ECI^2		-0.004^{***}
		(0.0003)
ln(GDP pc)	-0.015^{***}	-0.022^{***}
	(0.004)	(0.005)
ln(GDP pc)^2	0.001^{*}	0.002^{***}
	(0.001)	(0.001)
ln(population)	0.004^{***}	0.006^{***}
	(0.001)	(0.001)
ln(hub distance)	0.015^{***}	0.014^{***}
	(0.001)	(0.001)
adr high school	-0.026^{***}	-0.024^{***}
	(0.003)	(0.003)
n legis	0.004^{**}	0.002
	(0.001)	(0.001)
Constant	0.859^{***}	0.899^{***}
	(0.037)	(0.039)
Observations	5,554	5,554
R^2	0.396	0.413
Adjusted R^2	0.396	0.412
F Statistic	520.503^{***} (df = 7; 5546)	487.572^{***} (df = 8; 5545)

ECI refers to the Economic Complexity Index, *GDP pc* to the municipal Gross Domestic Product per capita, *hub distance* to the distance to the nearest hub municipality, *adr high school* to the age-grade distortion rate (the percentage of students in the correct grade for their age in high school), *n legis* to an institutional measure - the quantity of legislation and planning instruments existing in relation to the total expected by the IBGE Basic Municipal Information Survey. ***, ** and * denote significance at 1%, 5% and 10% levels respectively.

between the studies.

We also do not support the finding of Morais, Swart and Jordaan (2021): although we find a nonlinear relationship between inequality and ECI, it is not in an inverted U-shape. Thus, we do not support the hypothesis that municipalities with low economic complexity would experience an increase in inequality at the initial levels of increased complexity.

We can assume that some of the hypotheses applicable to a negative relationship between ECI and inequality at the country level might also be applied at the municipal level: greater variety and sophistication of sectors would tend to offer better occupational opportunities, more inclusive institutions, more equitable distribution of power, greater bargaining power for workers, or a higher share of labor in the total product.

Regarding the control variables, as we can see in the Table 13, we do not find empirical support for the Kuznets curve when evaluating the coefficients related to GDP per capita. Contrary to Lee and Vu (2020), Hartmann et al. (2017), the coefficient of per capita GDP is negative and statistically significant at 1%, and the squared term is positive and statistically significant at 1%, indicating a U-shaped relationship between Inequality and per capita GDP.

Table 13 also shows that a larger population is associated with greater inequality, as found in Hartmann et al. (2017), and a greater distance to a hub city is also related to higher inequality. Our education measure is negatively related to inequality, while our institutional measure is positively related only in the first specification, losing statistical significance when we include the squared term of the ECI in the regression.

Table 14 displays specific regressions for each Brazilian region. The coefficients of ECI and its square are negative and statistically significant at the 1% level across all regressions, indicating a consistent relationship where higher complexity is associated with lower consumption inequality in all Brazilian regions.

The coefficients of control variables seem to depend on the regions analyzed. For instance, there is support for Kuznets' inverted U-shaped curve in the relationship between per capita GDP and inequality in the Southern region, but not in the Central-West and North regions, with an opposite pattern observed in the Southeast and Northeast regions. Population is a predictor of inequality in the South, Northeast, and North regions, but not in the Southeast and Central-West. Distance to the nearest hub city relates to inequality in all regions. Better education is linked to lower inequality in the North, but the opposite is true in the Central-West and Northeast regions. Our institutional measure correlates negatively with inequality only in the Northeast, showing no relation in the other regions.

Table 14 – OLS regressions for all municipalities by region.

	Dependent variable:				
	Gini				
	South	Southeast	Central-West	Northeast	North
ECI	-0.010^{***}	-0.006^{***}	-0.007^{**}	-0.011^{***}	-0.012^{***}
	(0.002)	(0.001)	(0.003)	(0.001)	(0.004)
ECI^2	-0.006^{**}	-0.002^{***}	-0.007^{***}	-0.003^{***}	-0.004^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln(GDP pc)	0.036^{***}	-0.022^{***}	0.011	-0.054^{***}	-0.001
	(0.023)	(0.008)	(0.021)	(0.011)	(0.011)
ln(GDP pc)^2	-0.005	0.002^{***}	-0.003	0.008^{***}	-0.001
	(0.003)	(0.001)	(0.003)	(0.002)	(0.001)
ln(population)	0.004^{***}	-0.0001	0.001	0.010^{***}	0.007^{***}
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
ln(hub distance)	0.011^{***}	0.015^{***}	0.005^{**}	0.011^{***}	0.007^{***}
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
adr high school	0.001	-0.001	0.058^{***}	0.010^{**}	-0.042^{***}
	(0.009)	(0.007)	(0.017)	(0.005)	(0.013)
n legis	0.001	-0.002	-0.003	-0.005^{**}	0.001
	(0.004)	(0.003)	(0.006)	(0.002)	(0.004)
Constant	0.384^{***}	0.933^{***}	0.621^{***}	1.107^{***}	0.747^{***}
	(0.197)	(0.069)	(0.180)	(0.092)	(0.106)
Observations	1,187	1,667	466	1,787	447
R^2	0.261	0.334	0.133	0.251	0.308
Adjusted R^2	0.256	0.331	0.117	0.248	0.295
F Statistic	52.108^{***}	103.938^{***}	8.733^{***}	74.619^{***}	24.345^{***}

ECI refers to the Economic Complexity Index, *GDP pc* to the municipal Gross Domestic Product per capita, *hub distance* to the distance to the nearest hub municipality, *adr high school* to the age-grade distortion rate (the percentage of students in the correct grade for their age in high school), *n legis* to an institutional measure - the quantity of legislation and planning instruments existing in relation to the total expected by the IBGE Basic Municipal Information Survey. ***, ** and * denote significance at 1%, 5% and 10% levels respectively.

4.6.3 Limitations

Due to data limitations, we do not have long enough time series to perform fixed effects panel estimations. Payment method data is recent, with the introduction of Pix, for example, occurring only in 2020. Its adoption continues to grow: around 1.4 billion transactions were made in December 2021, while over 3 billion were conducted in July 2023[21]. This trend is likely changing the distribution of payment method data over the past years and is expected to continue doing so in the future.

Another limitation is the scarcity of municipal data which, when available, is generally outdated. For example, municipal GDP is computed with a two-year lag. Therefore, as more data becomes available, future research can explore more robust estimation methods. The limitation of municipal data can also impact the quality of the controls used in our regressions. Our measure of institutions, for example, does not capture the quality of the laws developed for each municipality, nor whether these laws are actually being applied or adopted. Additionally, we lack a measure for Informal Institutions.

4.7 Conclusion

Municipal-level inequality is positively correlated with crime and negatively with income growth, and people living in more unequal municipalities declare themselves more unhappy (Glaeser; Resseger; Tobio, 2009). However, measuring inequality at the municipal level depends on census surveys, which in the case of Brazil, are conducted approximately once a decade. This compromises the proposition and evaluation of local public policies aimed at reducing inequality and its impacts.

In this context, our work proposes to calculate consumption inequality at the municipal level using electronic payment data, such as credit card and Pix payments, as a proxy. In this way, we can timely calculate a measure of municipal-level inequality at any desired frequency. We show that our Gini index of consumption inequality is moderately correlated with the income inequality Gini index calculated using the 2010 IBGE Census data for Brazilian municipalities. Additionally, our consumption inequality index presents a similar regional distribution ordering to the income inequality index, although it generally indicates higher inequality, given the nature of the data used (electronic payments).

As an application of our index, we evaluate the relationship between consumption inequality and Economic Complexity at the municipal level. In a cross-sectional analysis, we show that Economic Complexity is negatively related to consumption inequality, meaning municipalities with more complex productive structures tend to have lower consumption inequality. This relationship is non-linear, indicating that the greater the complexity, the

[21] For more details, see <https://www.bcb.gov.br/estabilidadefinanceira/estatisticaspix>.

same variation in economic complexity is associated with a greater variation in inequality. This result remains quantitatively consistent when evaluated region by region.

5 Conclusions

In the first article, we assess the influence of ESG (Environmental, Social, and Governance) news on the stock returns of leading Brazilian companies listed on the Stock Exchange. Unlike previous work that uses proprietary and closed big data solutions, we created our own dictionary of ESG terms and a sentiment dictionary. This allowed for the selection, classification, and sentiment attribution to ESG news across the five dimensions and 26 categories defined by SASB. Using a firm-day panel where stock returns are the dependent variable and the principal independent variables are the instances of positive and negative company news, we found that positive (negative) news leads to corresponding positive (negative) returns. Significantly, the financial materiality of the news is a crucial factor, with investors resonating only with financially material news, disregarding mere reputational or non-pecuniary issues. Moreover, we found that the impact varies based on the SASB dimension under consideration. For instance, news within the Environmental sector influenced stock prices both positively and negatively. Conversely, within the Leadership and Governance spectrum, which encompasses news on corruption, only negative news showed potential to affect stock returns. Additionally, often the reaction to the news occurs within the 5-day or one-day prior window of the news release, indicating that the market incorporated the information one or more days before the announcement.

In the second article, we investigate the predictability of the Brazilian exchange rate using high-frequency intraday data (at 1, 5, and 15-minute frequencies) with local and global predictive variables, along with exchange rates from 17 other countries. We conducted two types of exercises: an out-of-sample fit, which uses contemporary data as predictors, and an out-of-sample forecast, where we attempt to predict the exchange rate at time $t + h$ using data available only up to time t. We found that the out-of-sample fit approach can indeed predict the Brazilian high-frequency exchange rate more accurately than the Random Walk approach across all methodologies employed. However, Machine Learning algorithms outperformed linear models estimated by OLS. In the out-of-sample forecasting exercise, it was possible to outperform the Random Walk with ML algorithms for horizons of up to four minutes at the 1-minute frequency, although the margins are narrow and predictability vanishes at the 5 and 15-minute frequencies. Using the rates of the futures contracts of the Brazilian Real/U.S. dollar as predictors, we were able to beat the Random Walk for a horizon of up to 6 minutes at the 1-minute frequency. This success is attributed to the initial determination of the exchange rate in Brazil's foreign exchange futures market, which then influences the spot market through arbitrage. We also evaluated the importance of each predictor when using Gradient Boosting in both exercises, finding that variables carrying local information, such as the Ibovespa and the short- and long-term interest rates, along with the Mexican exchange rate and the exchange rates of the BRICS countries included in the sample, are among the most important predictors.

In the third article, we propose calculating a Gini consumption inequality index using electronic payment data, specifically credit card and Pix. We show that our measure is moderately correlated with income inequality calculated using 2010 demographic census data at the municipal level, and that the ordering of the distributions of the inequality indices at the regional level are similar. However, the consumption inequality index is on average higher than the income inequality index, primarily due to the nature of the data used. Nonetheless, we believe we have a way to assess inequality at any desired frequency, thus enabling the evaluation of public policies aimed at reducing inequality. As an application, we investigate the relationship between the consumption inequality index and the Economic Complexity Index at the municipal level. We find a negative relationship, indicating that higher economic complexity is associated with lower consumption inequality. Additionally, the relationship is non-linear: with increasing ECI, the influence on consumption inequality becomes more significant. We also verify that this relationship holds when we evaluate the Brazilian regions individually.

Bibliography

Aguiar, M.; Bils, M. Has consumption inequality mirrored income inequality? *American Economic Review*, American Economic Association, v. 105, n. 9, p. 2725–2756, 2015.

Alesina, A.; Tella, R. D.; MacCulloch, R. Inequality and happiness: are europeans and americans different? *Journal of public economics*, Elsevier, v. 88, n. 9-10, p. 2009–2042, 2004.

Alston, L. J.; Melo, M. A.; Mueller, B.; Pereira, C. *Brazil in transition: Beliefs, leadership, and institutional change.* : Princeton University Press, 2016.

Amarante, V.; Lanzilotta, B.; Torres, J. Inequality and productive structure. *United nations University*, 2023.

Amat, C.; Michalski, T.; Stoltz, G. Fundamentals and exchange rate forecastability with simple machine learning methods. *Journal of International Money and Finance*, Elsevier, v. 88, p. 1–24, 2018.

Amel-Zadeh, A.; Serafeim, G. Why and how investors use esg information: Evidence from a global survey. *Financial Analysts Journal*, Taylor & Francis, v. 74, n. 3, p. 87–103, 2018.

Aprigliano, V.; Ardizzi, G.; Monteforte, L. Using payment system data to forecast economic activity. *60th issue (October 2019) of the International Journal of Central Banking*, 2019.

Arif, I. Productive knowledge, economic sophistication, and labor share. *World Development*, Elsevier, v. 139, p. 105303, 2021.

Attanasio, O.; Pistaferri, L. Consumption inequality over the last half century: some evidence using the new psid consumption measure. *American Economic Review*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, v. 104, n. 5, p. 122–126, 2014.

Attanasio, O. P.; Pistaferri, L. Consumption inequality. *Journal of Economic Perspectives*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203-2418, v. 30, n. 2, p. 3–28, 2016.

Aydin, D. *The marginal propensity to consume out of liquidity: Evidence from a randomized controlled trial.* 2015.

Baier, P.; Berninger, M.; Kiesel, F. Environmental, social and governance reporting in annual reports: A textual analysis. *Financial Markets, Institutions & Instruments*, Wiley Online Library, v. 29, n. 3, p. 93–118, 2020.

Baker, M.; Bergstresser, D.; Serafeim, G.; Wurgler, J. *Financing the response to climate change: The pricing and ownership of US green bonds.* 2018.

Baker, S. R. Debt and the response to household income shocks: Validation and application of linked financial account data. *Journal of Political Economy*, University of Chicago Press Chicago, IL, v. 126, n. 4, p. 1504–1557, 2018.

Baker, S. R.; Yannelis, C. Income changes and consumption: Evidence from the 2013 federal government shutdown. *Review of Economic Dynamics*, Elsevier, v. 23, p. 99–124, 2017.

Barro, R. J. *Inequality and growth revisited.* 2008.

Beckmann, J.; Czudaj, R.; Arora, V. The relationship between oil prices and exchange rates: theory and evidence. *US Energy Information Administration working paper series*, p. 1–62, 2017.

Bee, A.; Rothbaum, J. The administrative income statistics (ais) project: Research on the use of administrative records to improve income and resource estimates. *US Census Bureau SEHSD working paper*, v. 36, n. 2019, p. 2017–39, 2019.

Bekaert, G.; Hoerova, M.; Duca, M. L. Risk, uncertainty and monetary policy. *Journal of Monetary Economics*, Elsevier, v. 60, n. 7, p. 771–788, 2013.

Berg, F.; Koelbel, J. F.; Rigobon, R. Aggregate confusion: The divergence of esg ratings. *Review of Finance*, Oxford University Press, v. 26, n. 6, p. 1315–1344, 2022.

Bernardi, M.; Catania, L. The model confidence set package for r. *International Journal of Computational Economics and Econometrics*, Inderscience Publishers (IEL), v. 8, n. 2, p. 144–158, 2018.

Brandão, L. S. *Complexidade Econômica: uma análise de como as desigualdades se materializam no território.* Tese (Doutorado) — Universidade de Brasília, 2023.

Breiman, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

Browning, M.; Leth-Petersen, S. Imputing consumption from income and wealth information. *The Economic Journal*, Oxford University Press Oxford, UK, v. 113, n. 488, p. F282–F301, 2003.

Burkhauser, R. V.; Hérault, N.; Jenkins, S. P.; Wilkins, R. Top incomes and inequality in the uk: reconciling estimates from household survey and tax return data. *Oxford Economic Papers*, Oxford University Press, v. 70, n. 2, p. 301–326, 2018.

Busco, C.; Consolandi, C.; Eccles, R. G.; Sofra, E. A preliminary analysis of sasb reporting: Disclosure topics, financial relevance, and the financial intensity of esg materiality. *Journal of Applied Corporate Finance*, Wiley Online Library, v. 32, n. 2, p. 117–125, 2020.

Caous, E. L.; Huarng, F. Economic complexity and the mediating effects of income inequality: Reaching sustainable development in developing countries. *Sustainability*, MDPI, v. 12, n. 5, p. 2089, 2020.

Capelle-Blancard, G.; Desroziers, A.; Scholtens, B. Shareholders and the environment: a review of four decades of academic research. *Environmental Research Letters*, IOP Publishing, v. 16, n. 12, p. 123005, 2021.

Capelle-Blancard, G.; Petit, A. Every little helps? esg news and stock market reaction. *Journal of Business Ethics*, Springer, v. 157, n. 2, p. 543–565, 2019.

Ceriani, L.; Verme, P. The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, Springer, v. 10, p. 421–443, 2012.

Chatterji, A. K.; Durand, R.; Levine, D. I.; Touboul, S. Do ratings of firms converge? implications for managers, investors and strategy researchers. *Strategic Management Journal*, Wiley Online Library, v. 37, n. 8, p. 1597–1614, 2016.

Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. p. 785–794.

Cheung, Y.-W.; Chinn, M. D.; Pascual, A. G. Empirical exchange rate models of the nineties: Are any fit to survive? *Journal of international money and finance*, Elsevier, v. 24, n. 7, p. 1150–1175, 2005.

Choudhry, T.; McGroarty, F.; Peng, K.; Wang, S. High-frequency exchange-rate prediction with an artificial neural network. *Intelligent Systems in Accounting, Finance and Management*, Wiley Online Library, v. 19, n. 3, p. 170–178, 2012.

Chu, L. K.; Hoang, D. P. How does economic complexity influence income inequality? new evidence from international data. *Economic Analysis and Policy*, Elsevier, v. 68, p. 44–57, 2020.

Colombo, E.; Pelagatti, M. Statistical learning and exchange rate forecasting. *International Journal of Forecasting*, Elsevier, v. 36, n. 4, p. 1260–1289, 2020.

Constantine, C.; Khemraj, T. Geography, economic structures and institutions: A synthesis. *Structural Change and Economic Dynamics*, Elsevier, v. 51, p. 371–379, 2019.

Costa, A. et al. *Machine Learning and Oil Price Point and Density Forecasting*. 2021.

Cowell, F. A. *Measuring inequality*. : Oxford University Press, 2011.

Cui, B.; Docherty, P. Stock price overreaction to esg controversies. SSRN Working Paper 3559915. 2020.

Danielsson, J.; Luo, J.; Payne, R. Exchange rate determination and inter-market order flow effects. *The European Journal of Finance*, Taylor & Francis, v. 18, n. 9, p. 823–840, 2012.

Dasgupta, S.; Laplante, B.; Mamingi, N. Pollution and capital markets in developing countries. *Journal of Environmental Economics and management*, Elsevier, v. 42, n. 3, p. 310–335, 2001.

Deininger, K.; Squire, L. A new data set measuring income inequality. *The World Bank Economic Review*, Oxford University Press, v. 10, n. 3, p. 565–591, 1996.

Diebold, F. X.; Mariano, R. S. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, v. 13, n. 3, p. 253–263, 1995.

Dominicis, L. D.; Florax, R. J.; Groot, H. L. D. A meta-analysis on the relationship between income inequality and economic growth. *Scottish Journal of Political Economy*, Wiley Online Library, v. 55, n. 5, p. 654–682, 2008.

Dorfleitner, G.; Kreuzer, C.; Sparrer, C. Esg controversies and controversial esg: about silent saints and small sinners. *Journal of Asset Management*, Springer, v. 21, n. 5, p. 393–412, 2020.

Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *The Annals of Statistics*, v. 32, n. 1, p. 407–499, 2004.

Eika, L.; Mogstad, M.; Vestad, O. L. What can we learn about household consumption expenditure from data on income and assets? *Journal of Public Economics*, Elsevier, v. 189, p. 104163, 2020.

Ender, M.; Brinckmann, F. Impact of csr-relevant news on stock prices of companies listed in the austrian traded index (atx). *International Journal of Financial Studies*, MDPI, v. 7, n. 3, p. 36, 2019.

Evans, M. D.; Lyons, R. K. Order flow and exchange rate dynamics. *Journal of political economy*, The University of Chicago Press, v. 110, n. 1, p. 170–180, 2002.

Evans, M. D.; Lyons, R. K. Meese-rogoff redux: Micro-based exchange-rate forecasting. *American Economic Review*, v. 95, n. 2, p. 405–414, 2005.

Felício, W. R. d. O.; Júnior, J. L. R. Common factors and the exchange rate: results from the brazilian case. *Revista Brasileira de Economia*, SciELO Brasil, v. 68, p. 49–71, 2014.

Ferraro, D.; Rogoff, K.; Rossi, B. Can oil prices forecast exchange rates? an empirical analysis of the relationship between commodity prices and exchange rates. *Journal of International Money and Finance*, Elsevier, v. 54, p. 116–141, 2015.

Flammer, C. Corporate social responsibility and shareholder reaction: The environmental awareness of investors. *Academy of Management Journal*, Academy of Management Briarcliff Manor, NY, v. 56, n. 3, p. 758–781, 2013.

Franco, C. D. Esg controversies and their impact on performance. *The Journal of Investing*, Institutional Investor Journals Umbrella, v. 29, n. 2, p. 33–45, 2020.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

Friedman, J. H. *The elements of statistical learning: Data mining, inference, and prediction.* : springer open, 2017.

Gaglianone, W. P.; Marins, J. T. M. Evaluation of exchange rate point and density forecasts: an application to brazil. *International Journal of Forecasting*, Elsevier, v. 33, n. 3, p. 707–728, 2017.

Galbraith, J. K. Global inequality and global macroeconomics. *Journal of Policy modeling*, Elsevier, v. 29, n. 4, p. 587–607, 2007.

Gao, J.; Zhou, T. Quantifying china's regional economic complexity. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 492, p. 1591–1603, 2018.

Garcia, M. G.; Medeiros, M. C.; Vasconcelos, G. F. Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, Elsevier, v. 33, n. 3, p. 679–693, 2017.

Gelman, M.; Kariv, S.; Shapiro, M. D.; Silverman, D.; Tadelis, S. Harnessing naturally occurring data to measure the response of spending to income. *Science*, American Association for the Advancement of Science, v. 345, n. 6193, p. 212–215, 2014.

Gelman, M.; Kariv, S.; Shapiro, M. D.; Silverman, D.; Tadelis, S. How individuals respond to a liquidity shock: Evidence from the 2013 government shutdown. *Journal of Public Economics*, Elsevier, v. 189, p. 103917, 2020.

Gillan, S. L.; Koch, A.; Starks, L. T. Firms and social responsibility: A review of esg and csr research in corporate finance. *Journal of Corporate Finance*, Elsevier, v. 66, p. 101889, 2021.

Glaeser, E. L.; Resseger, M.; Tobio, K. Inequality in cities. *Journal of Regional Science*, Wiley Online Library, v. 49, n. 4, p. 617–646, 2009.

Glossner, S. Repeat offenders: Esg incident recidivism and investor underreaction. SSRN Working Paper 3004689. 2021.

Glück, M.; Hübel, B.; Scholz, H. Esg rating events and stock market reactions. SSRN Working Paper 3803254. 2021.

Gonçalves, R. N. C. et al. *Nowcasting Brazilian GDP with Eletronic Payments Data.* : Banco Central do Brasil, 2022.

Grewal, J.; Hauptmann, C.; Serafeim, G. Material sustainability information and stock price informativeness. *Journal of Business Ethics*, Springer, v. 171, p. 513–544, 2021.

Gross, D. B.; Souleles, N. S. Do liquidity constraints and interest rates matter for consumer behavior? evidence from credit card data. *The Quarterly journal of economics*, MIT Press, v. 117, n. 1, p. 149–185, 2002.

Guo, T.; Jamet, N.; Betrix, V.; Piquet, L.-A.; Hauptmann, E. *ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction.* 2020.

Güriş, B.; Kiran, B. The price of gold and the exchange rate: Evidence from threshold cointegration and threshold granger causality analyses for turkey. *Acta Oeconomica*, Akadémiai Kiadó, v. 64, n. 1, p. 91–101, 2014.

Hansen, P. R.; Lunde, A.; Nason, J. M. The model confidence set. *Econometrica*, Wiley Online Library, v. 79, n. 2, p. 453–497, 2011.

Hartmann, D. *Economic complexity and human development: How economic diversification and social networks affect human agency and welfare.* : Taylor & Francis, 2014.

Hartmann, D.; Guevara, M. R.; Jara-Figueroa, C.; Aristarán, M.; Hidalgo, C. A. Linking economic complexity, institutions, and income inequality. *World development*, Elsevier, v. 93, p. 75–93, 2017.

Hartmann, D.; Pinheiro, F. L. Economic complexity and inequality at the national and regional level. *arXiv preprint arXiv:2206.00818*, 2022.

Hassett, K. A.; Mathur, A. A new measure of consumption inequality. *AEI Economic Studies*, n. 2, 2012.

Hausmann, R.; Hidalgo, C. A. The network structure of economic output. *Journal of economic growth*, Springer, v. 16, p. 309–342, 2011.

Hausmann, R.; Hidalgo, C. A.; Bustos, S.; Coscia, M.; Simoes, A. *The atlas of economic complexity: Mapping paths to prosperity.* : Mit Press, 2014.

Hidalgo, C. A. Economic complexity theory and applications. *Nature Reviews Physics*, Nature Publishing Group UK London, v. 3, n. 2, p. 92–113, 2021.

Hidalgo, C. A.; Hausmann, R. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 106, n. 26, p. 10570–10575, 2009.

Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

Hokayem, C.; Bollinger, C.; Ziliak, J. P. The role of cps nonresponse in the measurement of poverty. *Journal of the American Statistical Association*, Taylor & Francis, v. 110, n. 511, p. 935–945, 2015.

Jeanneau, S.; Araujo, M.; Amante, A. The search for liquidity in the brazilian domestic government bond market. *BIS Quarterly Review, June*, 2007.

Jebe, R. The convergence of financial and esg materiality: taking sustainability mainstream. *American Business Law Journal*, Wiley Online Library, v. 56, n. 3, p. 645–702, 2019.

Júnior, J. L. R. The usefulness of financial variables in predicting exchange rate movements. *Insper Working Paper WPE: 332/2014*, 2014.

Kaspereit, T.; Lopatta, K. The value relevance of sam's corporate sustainability ranking and gri sustainability reporting in the e uropean stock markets. *Business Ethics: A european review*, Wiley Online Library, v. 25, n. 1, p. 1–24, 2016.

Kearney, C.; Liu, S. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, Elsevier, v. 33, p. 171–185, 2014.

Kemp-Benedict, E. An interpretation and critique of the method of reflections. *Munich Personal RePEc Archive*, 2014.

Khan, M.; Serafeim, G.; Yoon, A. Corporate sustainability: First evidence on materiality. *The accounting review*, American Accounting Association, v. 91, n. 6, p. 1697–1724, 2016.

Klassen, R. D.; McLaughlin, C. P. The impact of environmental management on firm performance. *Management science*, INFORMS, v. 42, n. 8, p. 1199–1214, 1996.

Kohlscheen, E.; Avalos, F. H.; Schrimpf, A. When the walk is not random: commodity prices and exchange rates. *International Journal of Central Banking*, v. 13, n. 2, p. 121–158, 2017.

Kolsrud, J.; Landais, C.; Spinnewijn, J. *Studying consumption patterns using registry data: lessons from Swedish administrative data.* 2017.

Krüger, P. Corporate goodness and shareholder wealth. *Journal of financial economics*, Elsevier, v. 115, n. 2, p. 304–329, 2015.

Kuznets, S. Economic growth and income inequality. *The American Economic Review*, n. 45, p. 1—-28, 1955.

Laplante, B.; Lanoie, P. The market response to environmental incidents in canada: a theoretical and empirical analysis. *Southern Economic Journal*, John Wiley & Sons, Incorporated, v. 60, n. 3, p. 657–672, 1994.

Larrimore, J.; Mortenson, J.; Splinter, D. Household incomes in tax data: Using addresses to move from tax-unit to household income distributions. *Journal of Human Resources*, University of Wisconsin Press, v. 56, n. 2, p. 600–631, 2021.

Lee, C.-C.; Wang, E.-Z. Economic complexity and income inequality: Does country risk matter? *Social Indicators Research*, Springer, v. 154, p. 35–60, 2021.

Lee, K.-K.; Vu, T. V. Economic complexity, human capital and income inequality: a cross-country analysis. *The Japanese Economic Review*, Springer, v. 71, p. 695–718, 2020.

Liang, H.; Renneboog, L. Corporate social responsibility and sustainable finance: A review of the literature. European Corporate Governance Institute–Finance Working Paper. 2020.

Maio, F. G. D. Income inequality measures. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 61, n. 10, p. 849–852, 2007.

Manahov, V.; Hudson, R.; Gebka, B. Does high frequency trading affect technical analysis and market efficiency? and if so, how? *Journal of International Financial Markets, Institutions and Money*, Elsevier, v. 28, p. 131–157, 2014.

Masini, R. P.; Medeiros, M. C.; Mendes, E. F. Machine learning advances for time series forecasting. *arXiv preprint arXiv:2012.12802*, 2020.

Medeiros, M.; Souza, P. H. G. F. d.; Castro, F. Á. d. A estabilidade da desigualdade de renda no brasil, 2006 a 2012: estimativa com dados do imposto de renda e pesquisas domiciliares. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 20, p. 971–986, 2015.

Medeiros, M. C.; Schütte, E. C. M.; Soussi, T. S. Global inflation: Implications for forecasting and monetary policy. Available at SSRN: https://ssrn.com/abstract=4145665. 2022.

Meese, R.; Rogoff, K. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, v. 14, p. 3–24, 1983.

Mervelskemper, L.; Streit, D. Enhancing market valuation of esg performance: is integrated reporting keeping its promise? *Business Strategy and the Environment*, Wiley Online Library, v. 26, n. 4, p. 536–549, 2017.

Meyer, B. D.; Sullivan, J. X. *Five decades of consumption and income poverty*. 2009.

Miralles-Quirós, M. M.; Miralles-Quirós, J. L.; Gonçalves, L. M. V. The value relevance of environmental, social, and governance performance: The brazilian case. *Sustainability*, MDPI, v. 10, n. 3, p. 574, 2018.

Moore, J. C.; Stinson, L. L.; Welniak, E. J. et al. Income measurement error in surveys: A review. *Journal of Official Statistics-Stockholm-*, ALMQVIST & WIKSELL INTERNATIONAL, v. 16, n. 4, p. 331–362, 2000.

Morais, M. B.; Swart, J.; Jordaan, J. A. Economic complexity and inequality: Does regional productive structure affect income inequality in brazilian states? *Sustainability*, MDPI, v. 13, n. 2, p. 1006, 2021.

Moura, M. L.; Lima, A. R.; Mendonça, R. M. Exchange rate and fundamentals: the case of brazil. *Economia Aplicada*, SciELO Brasil, v. 12, n. 3, p. 395–416, 2008.

Myšková, R.; Hájek, P. Sustainability and corporate social responsibility in the text of annual reports—the case of the it services industry. *Sustainability*, MDPI, v. 10, n. 11, p. 4119, 2018.

Nair, G. K.; Choudhary, N.; Purohit, H. The relationship between gold prices and exchange value of us dollar in india. *EMAJ: Emerging Markets Journal*, v. 5, n. 1, p. 17–25, 2015.

Neri, M.; Hecksher, M. *Mapa da Riqueza*. 2023. <https://cps.fgv.br/riqueza>. Accessed: 2023-11-27.

Olafsson, A.; Pagel, M. The liquid hand-to-mouth: Evidence from personal finance management software. *The Review of Financial Studies*, Oxford University Press, v. 31, n. 11, p. 4398–4446, 2018.

Palikuca, A.; Seidl, T. *Predicting High Frequency Exchange Rates Using Machine Learning*. 2016.

Palma, J. G. Homogeneous middles vs. heterogeneous tails, and the end of the 'inverted-u': It's all about the share of the rich. *development and Change*, Wiley Online Library, v. 42, n. 1, p. 87–153, 2011.

Pedregosa, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.

Perera, L. D. H.; Lee, G. H. Have economic growth and institutional quality contributed to poverty and inequality reduction in asia? *Journal of Asian Economics*, Elsevier, v. 27, p. 71–86, 2013.

Pham, M. H.; Truong, H. D. H.; Hoang, D. P. Economic complexity, shadow economy, and income inequality: fresh evidence from panel data. Preprint. 2023.

Piketty, T.; Saez, E. Income inequality in the united states, 1913–1998. *The Quarterly journal of economics*, MIT Press, v. 118, n. 1, p. 1–41, 2003.

Piketty, T.; Saez, E.; Zucman, G. Distributional national accounts: methods and estimates for the united states. *The Quarterly Journal of Economics*, Oxford University Press, v. 133, n. 2, p. 553–609, 2018.

Rime, D.; Sarno, L.; Sojli, E. Exchange rate forecasting, order flow and macroeconomic information. *Journal of International Economics*, Elsevier, v. 80, n. 1, p. 72–88, 2010.

Roe, M. J.; Siegel, J. I. Political instability: Effects on financial development, roots in the severity of economic inequality. *Journal of Comparative Economics*, Elsevier, v. 39, n. 3, p. 279–309, 2011.

Rossi, B. Exchange rate predictability. *Journal of economic literature*, v. 51, n. 4, p. 1063–1119, 2013.

Sbardella, A.; Pugliese, E.; Pietronero, L. Economic development and wage inequality: A complex system analysis. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 9, p. e0182774, 2017.

Schmidt, A. Sustainable news–a sentiment analysis of the effect of esg information on stock prices. Available at SSRN: https://ssrn.com/abstract=3809657. 2019.

Sebastian, R.; Biagi, F. *The routine biased technical change hypothesis: a critical review*. 2018.

Serafeim, G.; Yoon, A. Which corporate esg news does the market react to? *Financial Analysts Journal*, Taylor & Francis, v. 78, n. 1, p. 59–78, 2022.

Serafeim, G.; Yoon, A. Stock price reactions to esg news: The role of esg ratings and disagreement. *Review of accounting studies*, Springer, v. 28, p. 1500–1530, 2023.

Soave, G. P.; Gomes, F. A. R.; Junior, F. B. Desigualdade e desenvolvimento: revisitando a hipótese de kuznets após a redução da desigualdade nos municípios brasileiros. *Revista Brasileira de Estudos Regionais e Urbanos*, v. 13, n. 4, p. 581–605, 2019.

Sousa, R. A. d. *A Teoria da Complexidade reencontra o desenvolvimento econômico: uma análise de insumo-produto*. Dissertação (Mestrado) — Universidade de Brasília, 2018.

Stone, P. J. *General Inquirer Harvard-IV Dictionary*. 2002. <https://inquirer.sites.fas.harvard.edu/homecat.htm>. Accessed: 2022-12-31.

Tabak, B. M. The dynamic relationship between stock prices and exchange rates: Evidence for brazil. *International Journal of Theoretical and Applied Finance*, World Scientific, v. 9, n. 08, p. 1377–1396, 2006.

Tacchella, A.; Cristelli, M.; Caldarelli, G.; Gabrielli, A.; Pietronero, L. A new metrics for countries' fitness and products' complexity. *Scientific reports*, Nature Publishing Group UK London, v. 2, n. 1, p. 723, 2012.

Tetlock, P. C.; Saar-Tsechansky, M.; Macskassy, S. More than words: Quantifying language to measure firms' fundamentals. *The journal of finance*, Wiley Online Library, v. 63, n. 3, p. 1437–1467, 2008.

Thomas, C. Income inequality and economic development in latin america: A test for the kuznets inverted-u curve. *Indian Journal of Economics and Business*, Dr. Kishore G. Kulkarni, Indian Journal of Economics and Business, v. 14, n. 1, 2015.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

Török, I.; Benedek, J.; Gómez-Zaldívar, M. Quantifying subnational economic complexity: Evidence from romania. *Sustainability*, MDPI, v. 14, n. 17, p. 10586, 2022.

Trapeznikova, I. Measuring income inequality. *IZA World of Labor*, 2019.

Vartanian, P. R.; Citro, S. G.; Scarano, P. R. et al. Determinants of spot and forward interest rates in brazil in an international liquidity scenario: An econometric analysis for the period 2007-2019. *International Journal of Applied Economics, Finance and Accounting*, Online Academic Press, v. 10, n. 1, p. 19–31, 2021.

Ventura, A.; Garcia, M. Mercados futuro e à vista de câmbio no brasil: o rabo abana o cachorro. *Revista Brasileira de Economia*, SciELO Brasil, v. 66, p. 21–48, 2012.

Vincentiis, P. de. Do international investors care about esg news? *Qualitative Research in Financial Markets*, Emerald Publishing Limited, v. 15, n. 4, p. 572–588, 2023.

Violante, G. L. Skill-biased technical change. *The new Palgrave dictionary of economics*, Palgrave Macmillan Basingstoke, v. 2, p. 1–6, 2008.

Wu, S.; Zheng, X.; Wei, C. Measurement of inequality using household energy consumption data in rural china. *Nature Energy*, Nature Publishing Group UK London, v. 2, n. 10, p. 795–803, 2017.

Yoon, B.; Lee, J. H.; Byun, R. Does esg performance enhance firm value? evidence from korea. *Sustainability*, MDPI, v. 10, n. 10, p. 3635, 2018.

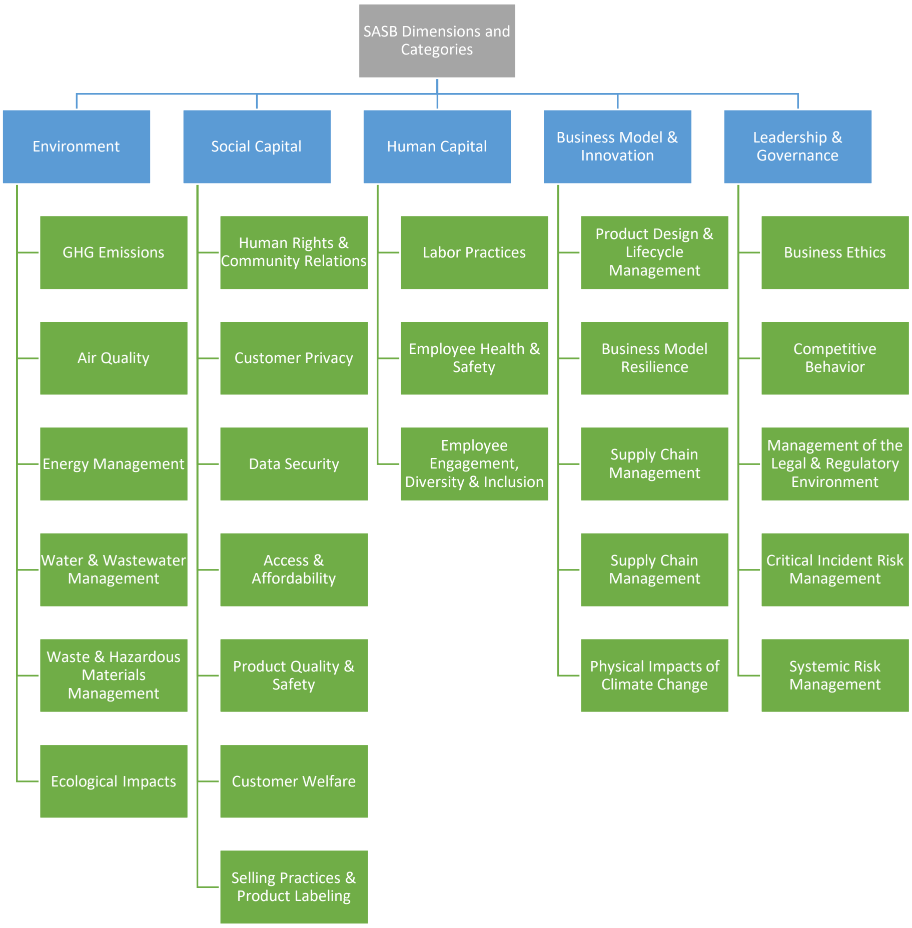
Zhang, Y.; Hamori, S. The predictability of the exchange rate when combining machine learning and fundamental models. *Journal of Risk and Financial Management*, Multidisciplinary Digital Publishing Institute, v. 13, n. 3, p. 48, 2020.

Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, Wiley Online Library, v. 67, n. 2, p. 301–320, 2005.

Appendix

APPENDIX A – SASB Dimensions and Categories

Figure 17 – The 5 dimensions (in blue) and the 26 categories (in green) in which ESG topics are organized according to the SASB standard. For a description of the categories, see <https://www.sasb.org/standards/materiality-finder/.>.


SASB Dimensions and Categories
Environment
Social Capital
Human Capital
Business Model & Innovation
Leadership & Governance
GHG Emissions
Air Quality
Energy Management
Water & Wastewater Management
Waste & Hazardous Materials Management
Ecological Impacts
Human Rights & Community Relations
Customer Privacy
Data Security
Access & Affordability
Product Quality & Safety
Customer Welfare
Selling Practices & Product Labeling
Labor Practices
Employee Health & Safety
Employee Engagement, Diversity & Inclusion
Product Design & Lifecycle Management
Business Model Resilience
Supply Chain Management
Supply Chain Management
Physical Impacts of Climate Change
Business Ethics
Competitive Behavior
Management of the Legal & Regulatory Environment
Critical Incident Risk Management
Systemic Risk Management

APPENDIX B – Machine Learning Hyperparameters Range

Table 15 – Range of values assessed for each model during the cross-validation process.

Method	**Hyperparameter (sklearn class: LassoCV)**	**Range**
LASSO (CV)	‘alphas’	None (auto)
	‘n_alphas’	100

Method	**Hyperparameter (sklearn class: Lasso)**	**Value or Range**
LASSO (GS)	‘alpha’	$10^{-8} : 10$

Method	**Hyperparameter (sklearn class: RidgeCV)**	**Value or Range**
Ridge (CV)	‘alphas’	$10^{-4} : 10^{4}$

Method	**Hyperparameter (sklearn class: Ridge)**	**Value or Range**
Ridge (GS)	‘alpha’	$10^{-8} : 10^{8}$

Method	**Hyperparameter (sklearn class: ElasticNetCV)**	**Range**
ElasticNet (CV)	‘l1_ratio’	[0.01, 0.05, .1, .3, .5, .7, .9, .95, 1]
	‘n_alphas’	100
	‘alphas’	None (auto)

Method	**Hyperparameter (sklearn class: ElasticNet)**	**Range**
ElasticNet (GS)	‘l1_ratio’	0.1 : 1
	‘alpha’	[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}]

Method	**Hyperparameter (sklearn class: RandomForestRegressor)**	**Range**
RF (GS)	‘n_estimators’	[100, 500]
	‘max_features’	[1/4, 1/3, 1/2, ‘auto’]
	‘max_depth’	[1, 2, 4, 6, 8, 10, 20, None]
	‘min_samples_split’	[2, 5, 10]
	‘min_samples_leaf’	[1, 2, 4]

Method	**Hyperparameter (class: XGBRegressor)**	**Range**
XGB (GS)	‘n_estimators’	[100, 500]
	‘learning_rate’	[.004, .01, .02, .05, 0.1]
	‘max_depth’	[3, 4, 6]
	‘min_child_weight’	[5, 50, 100, 500]
	‘subsample’	[.6, .8, 1]
	‘colsample_bytree’	[.5, .7, 1]
	‘gamma’	[0.001, 0]

APPENDIX C – Out-of-sample Model Confidence Sets (MCS)

Table 16 – Out-of-sample Superior Set of Models, 1-minute frequency, horizon=1 minute ahead.

Model	Rank	$t_{i\cdot}$	p-value
LASSO (BIC)	1	-2.47	1.0000
ElasticNet (GS)	2	-2.16	1.0000
LASSO (GS)	3	-2.13	1.0000
ElasticNet (CV)	4	-1.39	1.0000
LASSO (CV)	5	-1.39	1.0000
XGB (GS)	6	-1.32	1.0000
LASSO (AIC)	7	-0.44	1.0000
AR(1)	8	0.18	1.0000
RF (GS)	11	0.59	0.9942
Ridge (CV)	15	0.79	0.7918

Sets created for an 80% confidence interval. OLS models were omitted with the intention of simplifying the table. 'Rank' corresponds to the ranking based on the $t_{i\cdot}$ statistic. In the context of the MCS procedure, for a confidence level of $1 - \alpha$, models with p-values greater than α are not statistically worse than the best model and are therefore included in the set of superior models. Essentially, a higher p-value indicates a model's performance is closer to the top-performing models within the evaluated set. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window.

Table 17 – Out-of-sample Superior Set of Models, 1-minute frequency, horizon=2 minutes ahead.

Model	Rank	$t_{i\cdot}$	p-value
LASSO (BIC)	1	-2.06	1.0000
AR(1)	3	-1.62	1.0000
RF(GS)	4	-1.20	1.0000
ElasticNet (CV)	8	-0.43	1.0000
LASSO (CV)	9	-0.43	1.0000
ElasticNet (GS)	10	-0.32	1.0000
LASSO (GS)	14	-0.04	1.0000
LASSO (AIC)	15	0.61	0.9998
XGB (GS)	12	-0.08	1.0000
Ridge (CV)	16	1.06	0.8780
Random Walk	17	1.18	0.8078

Sets created for an 80% confidence interval. OLS models were omitted with the intention of simplifying the table. 'Rank' corresponds to the ranking based on the $t_{i\cdot}$ statistic. In the context of the MCS procedure, for a confidence level of $1 - \alpha$, models with p-values greater than α are not statistically worse than the best model and are therefore included in the set of superior models. Essentially, a higher p-value indicates a model's performance is closer to the top-performing models within the evaluated set. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window.

Table 18 – Out-of-sample Superior Set of Models, 1-minute frequency, horizon=3 minutes ahead.

Model	Rank	$t_{i\cdot}$	p-value
XGB (GS)	1	-2.45	1.0000
ElasticNet (GS)	2	-1.85	1.0000
AR(1)	3	-1.54	1.0000
LASSO (GS)	4	-1.45	1.0000
LASSO (BIC)	5	-1.43	1.0000
LASSO (CV)	9	-0.19	1.0000
ElasticNet (CV)	11	-0.15	1.0000
RF(GS)	12	0.12	1.0000
Random Walk	15	0.66	0.9942
Ridge (CV)	16	0.92	0.9464
LASSO (AIC)	17	0.94	0.9414

Sets created for an 80% confidence interval. OLS models were omitted with the intention of simplifying the table. 'Rank' corresponds to the ranking based on the $t_{i\cdot}$ statistic. In the context of the MCS procedure, for a confidence level of $1 - \alpha$, models with p-values greater than α are not statistically worse than the best model and are therefore included in the set of superior models. Essentially, a higher p-value indicates a model's performance is closer to the top-performing models within the evaluated set. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window.

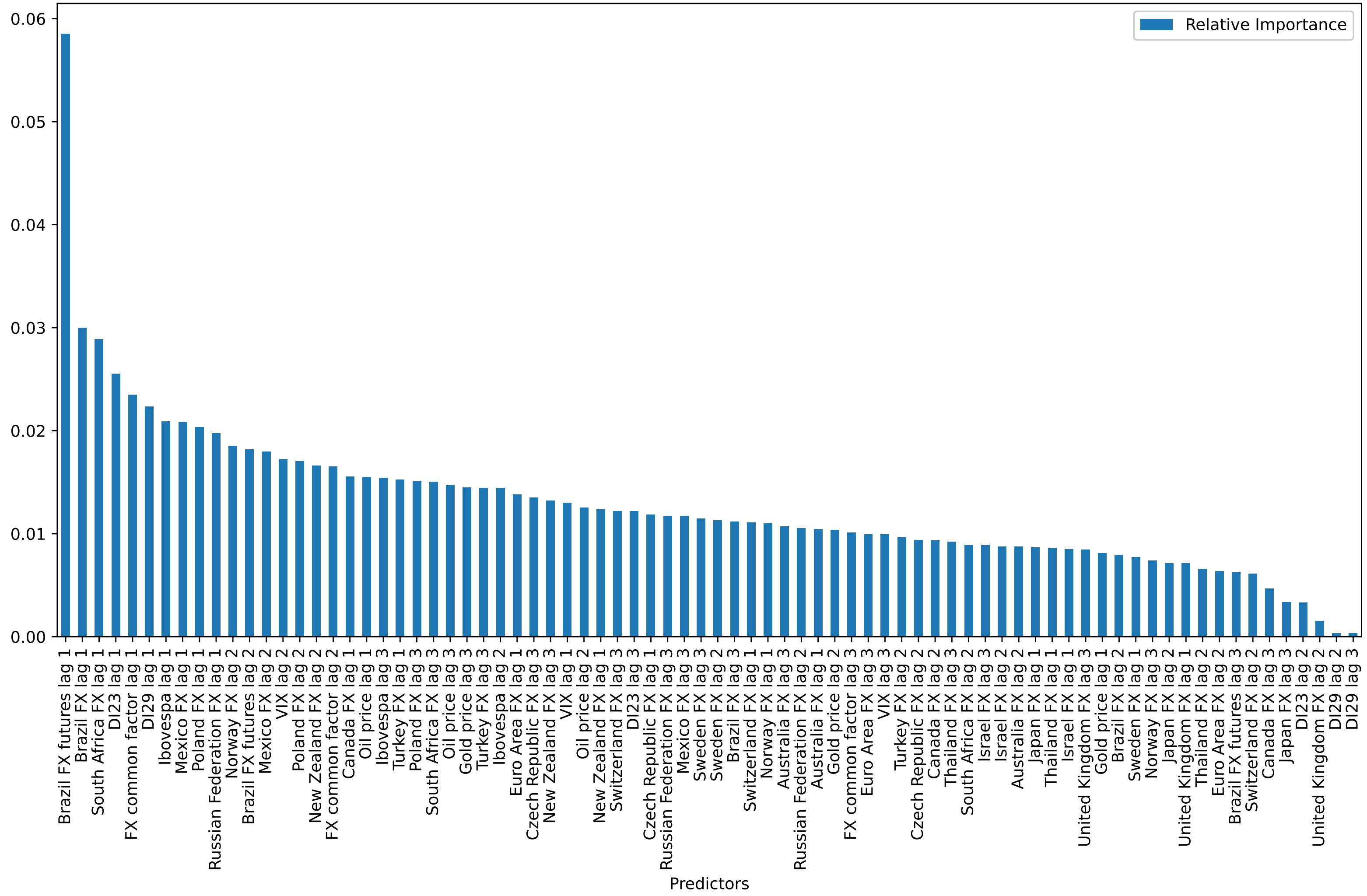
Table 19 – Out-of-sample Superior Set of Models, 1-minute frequency, horizon=4 minutes ahead.

Model	Rank	$t_{i\cdot}$	p-value
LASSO (GS)	1	-2.66	1.0000
ElasticNet (GS)	2	-2.26	1.0000
AR(1)	3	-1.80	1.0000
LASSO (BIC)	4	-1.64	1.0000
XGB (GS)	5	-1.18	1.0000
ElasticNet (CV)	6	-1.08	1.0000
LASSO (CV)	7	-1.06	1.0000
LASSO (AIC)	11	-0.05	1.0000
RF(GS)	14	0.56	0.9984
Random Walk	16	0.76	0.9844
Ridge (CV)	17	1.17	0.8402

Sets created for an 80% confidence interval. OLS models were omitted with the intention of simplifying the table. 'Rank' corresponds to the ranking based on the $t_{i\cdot}$ statistic. In the context of the MCS procedure, for a confidence level of $1-\alpha$, models with p-values greater than α are not statistically worse than the best model and are therefore included in the set of superior models. Essentially, a higher p-value indicates a model's performance is closer to the top-performing models within the evaluated set. The suffix of each Machine Learning method indicates the method for determining the hyperparameters as described in Section 3.3.2: (CV) indicates that the hyperparameters are determined, by cross-validation, at each advance of the rolling training window; (GS) indicates that the hyperparameters are defined once, on the initial training data; (AIC) and (BIC) indicate the respective informational criteria used for determining the hyperparameters when using the LARS algorithm at each advance of the rolling window.

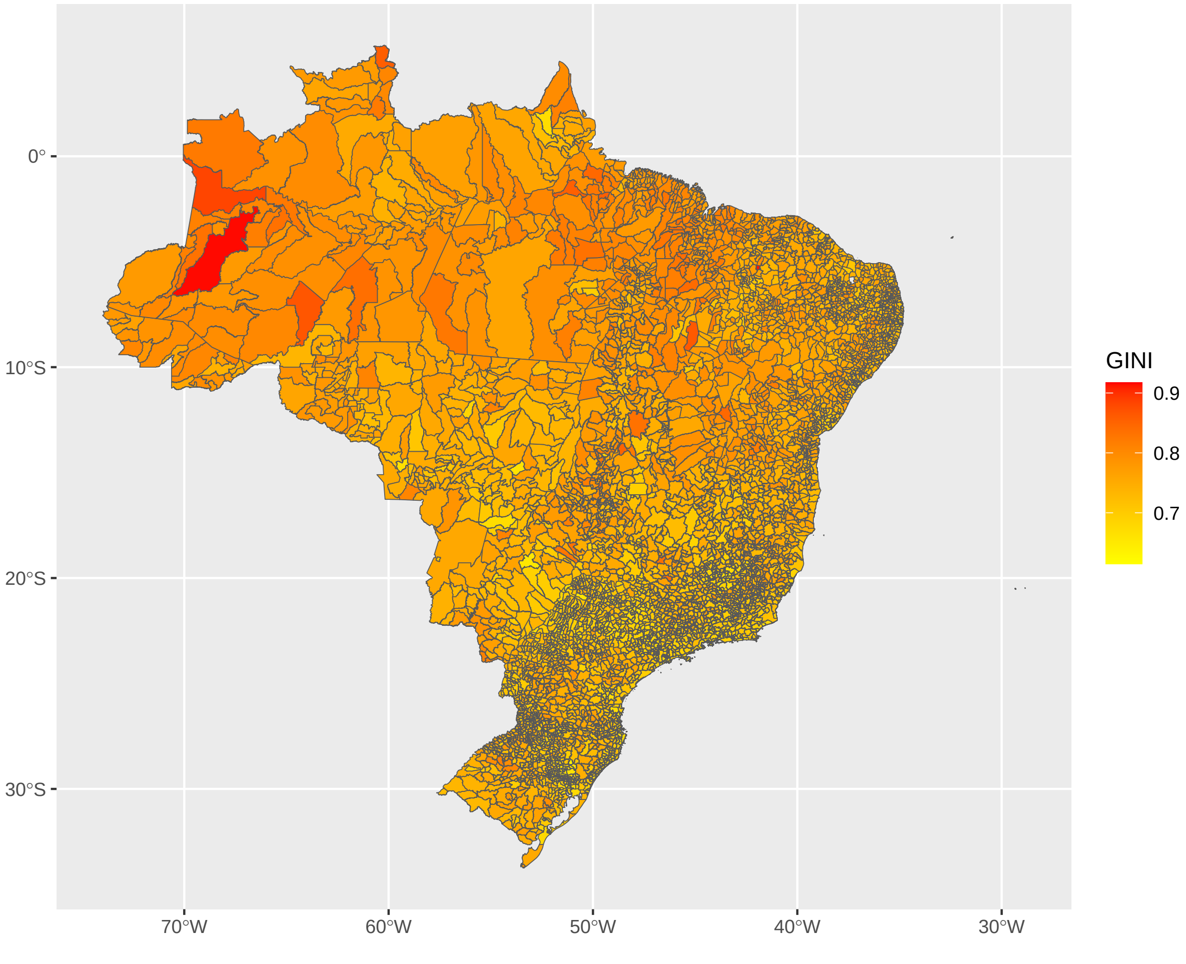
APPENDIX D – Out-of-sample XGB Importances with Real/U.S. dollar futures contracts rate

Figure 18 – Importance of XGB predictors in out-of-sample forecasting exercise at 1-minute frequency, horizon=1 minute ahead, with Real/U.S. dollar futures contracts rate as predictor.



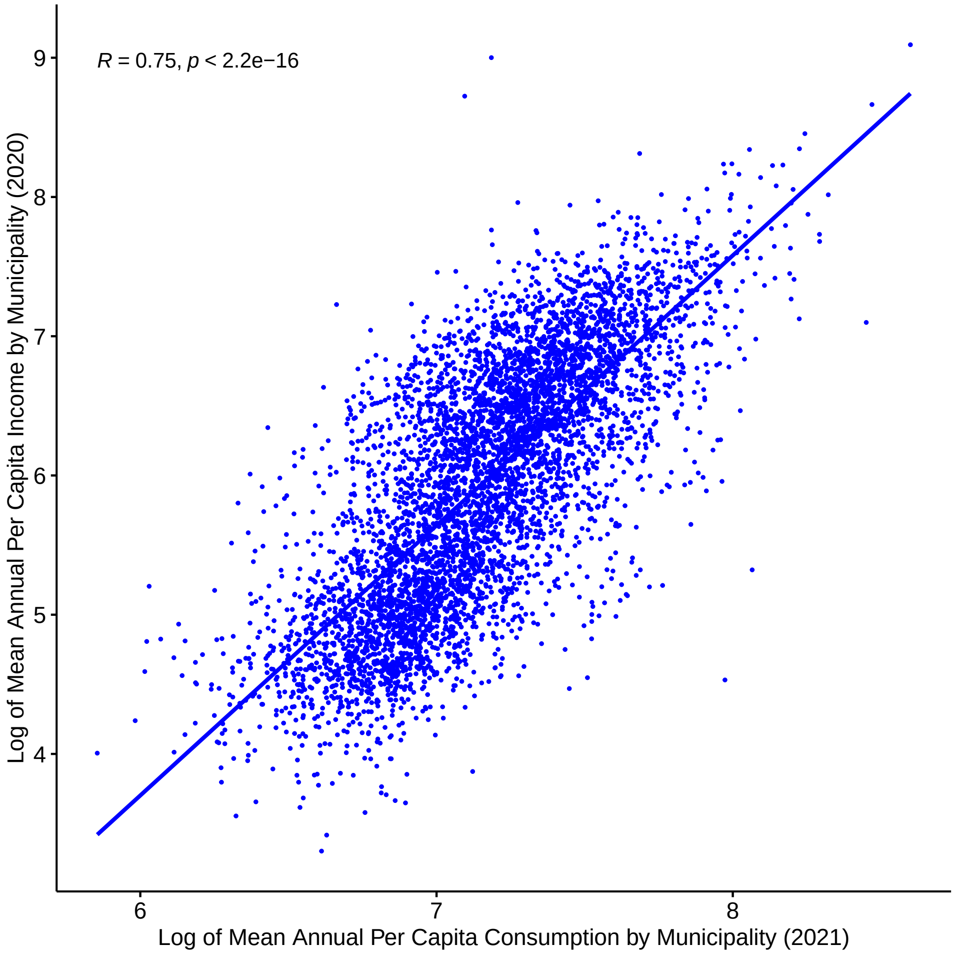
APPENDIX E – Geographic heat map of consumption inequality Gini coefficients

Figure 19 – Geographic heat map of consumption inequality Gini coefficients, highlighting regional disparities.



APPENDIX F – Correlation between consumption and income data

Figure 20 – Correlation between the average annual per capita consumption of each municipality based on electronic payment data from the base year 2021, and the average annual per capita income estimated by Neri and Hecksher (2023) for the base year 2020.



APPENDIX G – Boxplots of municipal consumption Gini coefficients, by state

Figure 21 – Boxplots of municipal consumption Gini coefficients highlighting variability among states.

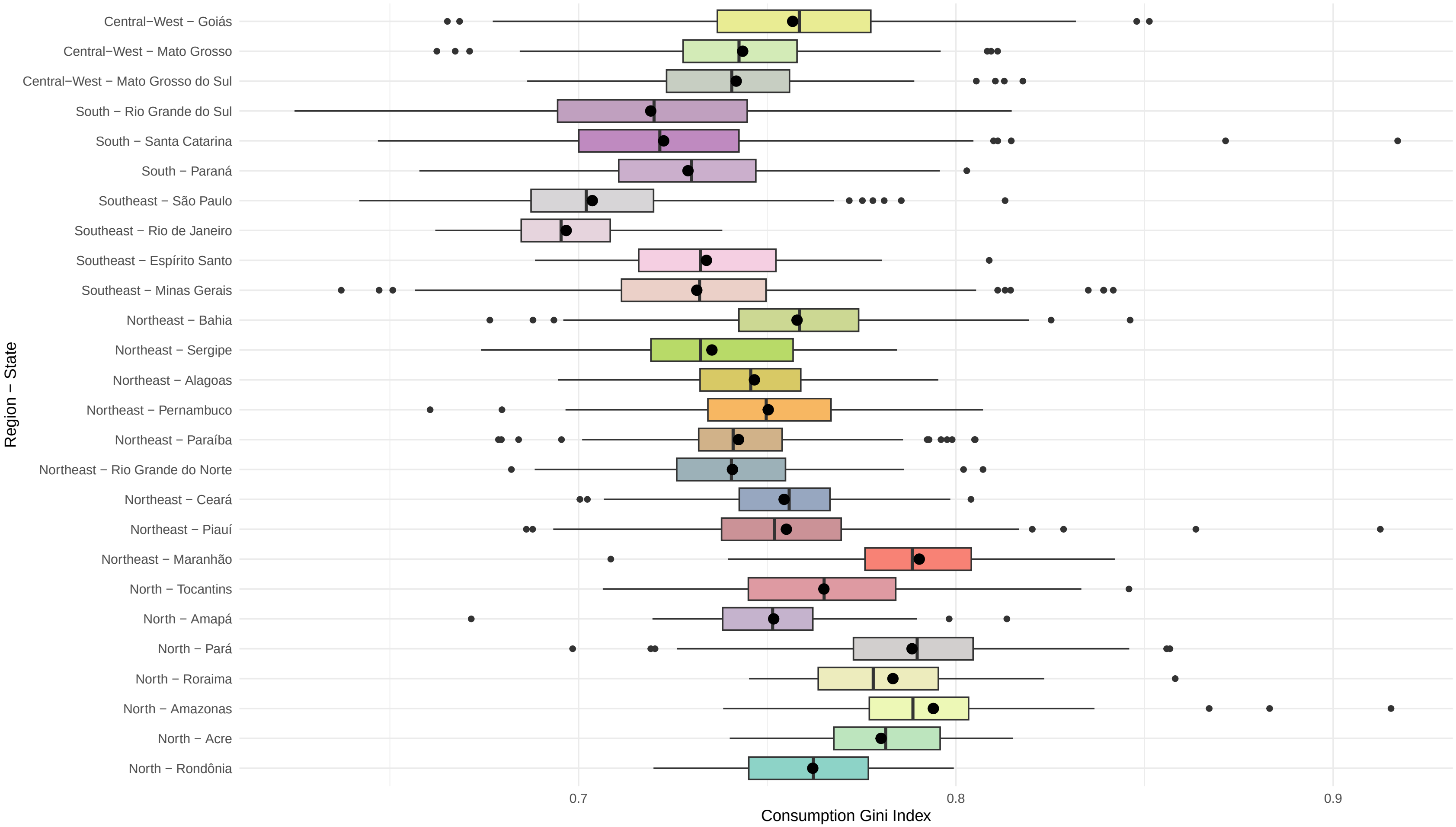
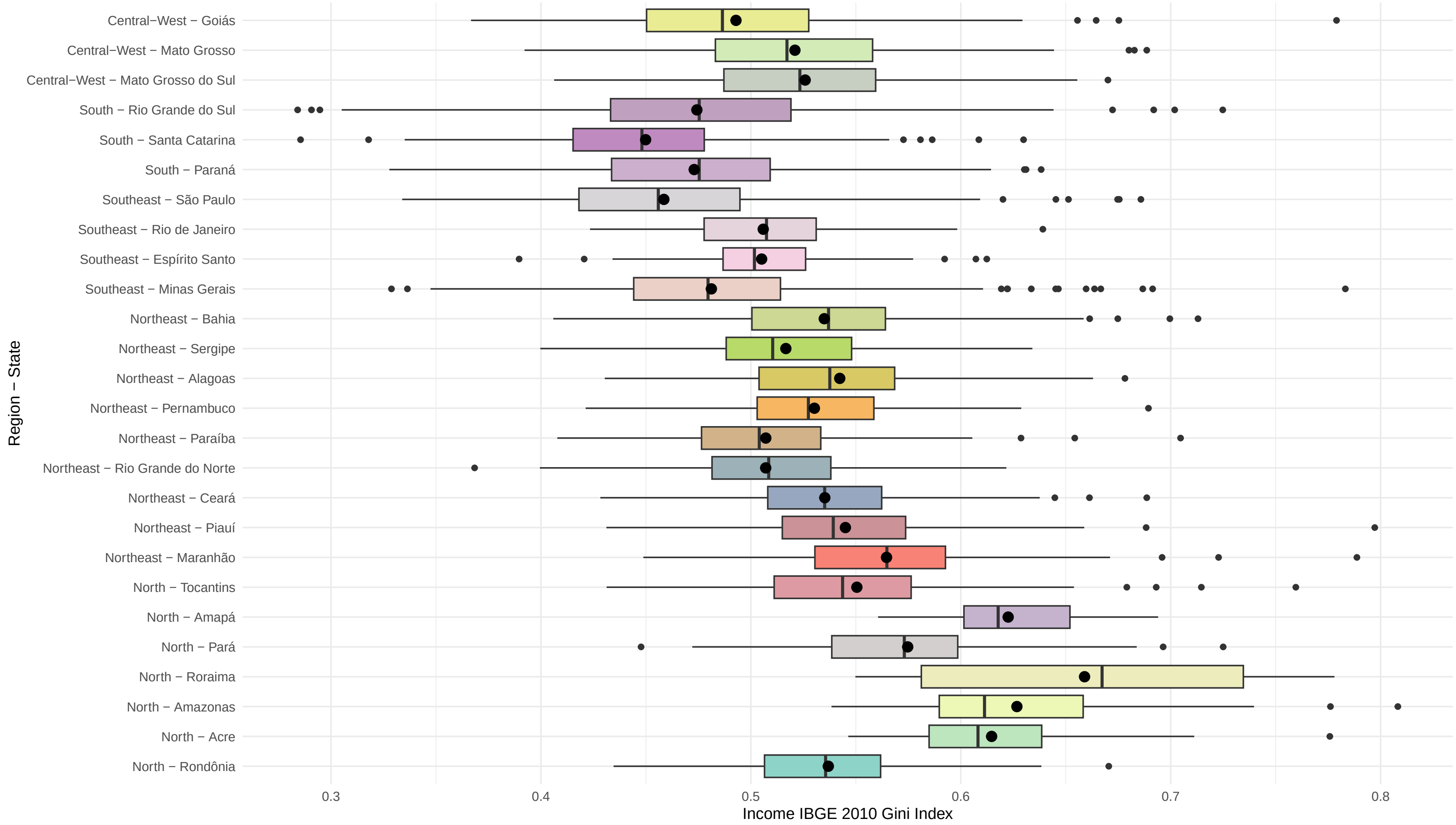


Figure 22 – Boxplots of municipal IBGE 2010 income Gini coefficients highlighting variability among states.



APPENDIX H – SQL Query for Computing the Gini Index

Let ‘T_PAG_INDIV’ be a table with the monthly average of electronic payments, ‘C’, made by each individual, uniquely identified in each row of the table, residing in the municipalities identified in ‘MUNICIPIO_PAG’. The SQL query for calculating the Gini index for each municipality is then given by:

```sql
WITH
T AS (
  SELECT
    MUNICIPIO_PAG,
    C,
    ROW_NUMBER() OVER(PARTITION BY MUNICIPIO_PAG ORDER BY C) AS i
  FROM T_PAG_INDIV
  )

SELECT
  MUNICIPIO_PAG,
  2.0 * SUM(C * i) / (COUNT(C) * SUM(C)) - 1.0 - (1.0 / COUNT(C))
      AS Gini
FROM T
GROUP BY MUNICIPIO_PAG;
```

APPENDIX I – IBGE activity group codes not considered for consumption data

Table 20 – List of IBGE activity group codes identifying institutions where payments made via Pix are not considered as consumption.

Code	**Activity**
050	Coal Mining
060	Oil and Natural Gas Extraction
071	Iron Ore Mining
072	Extraction of Non-Ferrous Metal Ores
081	Extraction of Stone, Sand and Clay
089	Extraction of Other Non-Metallic Minerals
091	Support Activities for Oil and Natural Gas Extraction
099	Support Activities for Mining, Except Oil and Natural Gas
411	Real Estate Development
421	Construction of Roads, Railways, Urban Works and Special Works
422	Infrastructure Works for Electricity, Telecommunications, Water, Sewer and Pipeline Transport
641	Central Bank
646	Activities of Holding Companies
647	Investment Funds
649	Unspecified Financial Services Activities
653	Reinsurance
841	Administration of the State and Economic and Social Policy
842	Collective Services Provided by Public Administration
843	Compulsory Social Security
990	International Bodies and Other Extraterritorial Institutions

Annex

ANNEX A – Anticipated municipal legislation and planning instruments

Table 21 – Anticipated Legislation and Planning Instruments.

Legislation/Instrument
City Master Plan
Legislation on Special Social Interest Area/Zone
Legislation on Special Interest Area/Zone
Urban Perimeter Law
Land Subdivision Legislation
Zoning or Land Use and Occupation Legislation
Created Soil Legislation or Onerous Granting of the Right to Build
Improvement Contribution Legislation
Consortium Urban Operation Legislation
Neighborhood Impact Study Legislation
Building Code
Environmental Zoning or Ecological-Economic Zoning Legislation
Administrative Servitude Legislation
Heritage Preservation Legislation
Conservation Unit Legislation
Special Use Concession for Housing Legislation
Special Urban Property Usucapion Legislation
Surface Right Legislation
Land Regularization Legislation
Possession Legitimization Legislation
Preliminary Environmental Impact Study Legislation
Code of Postures

Source: IBGE. Base year: 2021. <https://www.ibge.gov.br/estatisticas/sociais/saude/10586-pesquisa-de-informacoes-basicas-municipais.html>. Accessed on 2023-12-18.