Pontifícia Universidade Católica
do Rio de Janeiro

Leonardo Caio de Ladalardo Martins

From Micro to Macro: Essays in Textual Analysis

Dissertação de Mestrado

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

Advisor: Prof. Marcelo Cunha Medeiros

Rio de Janeiro
March 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Leonardo Caio de Ladalardo Martins

From Micro to Macro: Essays in Textual Analysis

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

Prof. Marcelo Cunha Medeiros
Advisor
Departamento de Economia – PUC-Rio

Prof. Eduardo Zilberman
Departamento de Economia – PUC-Rio

Prof. Marcelo Fernandes
Departamento de Economia - FGV/EESP

Rio de Janeiro, March the 14th, 2022

Leonardo Caio de Ladalardo Martins

B.A. in Economics, Universidade de São Paulo (USP), 2019

Bibliographic data

de Ladalardo Martins, Leonardo Caio

From Micro to Macro: Essays in Textual Analysis / Leonardo Caio de Ladalardo Martins; advisor: Marcelo Cunha Medeiros. – 2022.

101 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2022.

Inclui bibliografia

1. Economia – Teses. 2. Econometria – Teses.
3. Dados Textuais. 4. Google Trends. 5. Covid-19. 6. Mobilidade. 7. Causalidade. 8. Efeitos Fixos. 9. Previsão. 10. Nowcasting. 11. Modelos de Shrinkage.
I. Medeiros, Marcelo Cunha. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 620.11

To my father (in memorian).

Acknowledgments

Firstly, I would like to thank my advisor Marcelo Medeiros for not being just an advisor, but a very good friend. Thank you for all your support, meetings, projects, lectures and discussions. I hope this friendship could last very long!

I also would like to thank the examinnation comitee and all professors and staff members from PUC-Rio who helped us through the very tough moments of the Covid-19 pandemics. It was tough, but we are alive!

I could not have finished this program without all support for my dearest friends. In particular I would like to thank Alexandre, Thales, Gabriel, Marcelo, Barbara, Maria, Marina, Manuela and Renata. Thank your for being together in this journey and close whenever I needed!

Finally, I would also like to thank my beloved family for all the support and unconditional love. For Antonio, Wilma, Marina, Manon, André, Daniel and my nephews Sofia, Bernardo and Victor. Unfortunately, I cannot read this dedicatory to my father in person, but he knows that I think about him everyday. Thank you for being such an amazing dad!

This study was partially financed by the Coordenação de Aperfeiçoamento Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Abstract

de Ladalardo Martins, Leonardo Caio; Medeiros, Marcelo Cunha (Advisor). **From Micro to Macro: Essays in Textual Analysis**. Rio de Janeiro, 2022. 101p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

This study exploits non-conventional data sources such as newspaper textual data and internet searches from Google Trends in two empirical problems: (i) analysing the impacts of mobility on cases and deaths due to Covid-19; (ii) nowcasting GDP in high-frequency. The first paper resorts to unstructured data to control for non-observable behavioural effects and finds that an increase in residential mobility significantly reduces Covid-19 cases and deaths over a 4-week horizon. The second paper uses unstructured data sources to nowcast GDP on a weekly basis, showing that textual data and Google Trends can significantly enhance the quality of nowcasts (measured by MSE, MAE and other metrics) compared to Focus's market expectations as a benchmark. In both cases, unstructured data was revealed to be a valuable source of information not encoded in structured indicators.

Keywords

Text-Data; Google Trends; Covid-19; Mobility; Causality; Fixed Effects; Forecasting; Nowcasting; Shrinkage Models; .

Resumo

de Ladalardo Martins, Leonardo Caio; Medeiros, Marcelo Cunha. **De Micro à Macro: Ensaios em Análise Textual**. Rio de Janeiro, 2022. 101p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Este estudo explora fontes de dados não convencionais como dados textuais de jornais e pesquisas de internet do Google Trends em dois problemas empíricos: (i) analisar o impacto da mobilidade sobre o número de casos e mortes por Covid-19; (ii) *nowcasting* do PIB em alta-frequência. O primeiro artigo usa fontes de dados não estruturados como controle para fatores comportamentais não observados e encontra que um aumento na mobilidade residencial diminui significativamente o número de casos e mortes num horizonte de quatro semanas. O segundo artigo usa fontes de dados não estruturadas para fazer um *nowcasting* semanal do PIB, mostrando que dados textuais e Google Trends pode aumentar a qualidade das projeções (medido pelo EQM, EAM e outras métricas) comparado com as expectativas de mercado do Focus como base. Em ambos casos, dados não estruturados reveleram-se fontes ricas de informação não codificadas em indicadores estruturados convencionais.

Palavras-chave

Dados Textuais; Google Trends; Covid-19; Mobilidade; Causalidade; Efeitos Fixos; Previsão; Nowcasting; Modelos de Shrinkage; .

Table of contents

List of figures

List of tables

1
The Impacts of Mobility on Infectious Diseases Dynamics Using Soft and Hard Data: The Case of Covid-19

1.1
Introduction

The Covid-19 pandemic has created a new dynamic in terms of social behavior. Its impacts on society have been widely studied over many fields and various scientific reports. Many studies were born based on different analysis of the Covid-19 effects: from psychology (e.g. Kontoangelos et al. (2020)) to economic impact studies (e.g. Deb et al. (2020)). This paper focuses on the impacts of restrictions on mobility in the Covid-19 number of cases and deaths. We resort to panel data from 2020 and 2021 at the municipality level from Brazil to measure the short-term impacts of reduction in mobility on the dynamics of Covid-19.

Throughout the first year of the pandemic (2020), many countries adopted circulation restrictions intending to reduce the spread of the disease on the population (see Goldstein et al. (2021), and Scherbina (2021)). However, there is considerable heterogeneity in terms of the restriction degrees across the countries: while New Zealand imposed a high-level centralized lockdown strategy (see Stannard et al. (2020)), Brazil only imposed decentralized mobility restrictions.

In this scenario, evaluating the causal effects of mobility on infection proliferation is a challenge, as we cannot divide countries and regions according to its lockdown policies and evaluate the effects of circulation restrictions as in a randomized experiment. There is an additional complication of isolating pure effects of mobility impacts, as individual behavior may pollute the effects. For example, behavioral (non-observable) variables such as usage of masks, social distancing, and the adoption of better hygiene measures may affect mobility and infection levels, generating an omitted bias issue. Agglomerations, a non-observable variable, may also affect mobility and Covid-19 spread. On the other hand, when the infection rates are high, people tend to comply more with restrictive measures, generating a simultaneity bias.

This paper has the objective of empirically analyzing the effects of restric-

tions to circulation on the infection spread without resorting to epidemiological models or theoretical formulation of individual behavior. While recurring to a weekly panel of municipalities, we construct a fixed-effects model based on a simple identification scheme, illustrated by a Direct Acyclic Graph (DAG). The DAG motivates the inclusion of textual data controls to control for non-observable effects and the usage of lagged mobility and control variables to address the simultaneity bias. Our results suggest that even in the absence of controls, an increase in residential mobility (i.e., a decrease in overall mobility) can reduce the number of Covid-19 cases and deaths. Adding controls for behavioral effects only affects the magnitude of the coefficients and not their qualitative properties as its sign and statistical significance.

Some papers focused on analyzing the effect of Covid-19 spread over mobility (i.e., the converse of our causal identification), which is the case of Engle et al. (2020). Also, other papers analyzed the effects of non-pharmaceutical interventions in terms of Covid-19 spread as Kong & Prinz (2020) and also some country-specific analyzes which focus on evaluating the effects on mobility after Covid-19 as Batty et al. (2021) for London, Janiak et al. (2021) for Chile and Benítez et al. (2020) for Latin-America Countries. However, as our causal hypothesis rules out the possible simultaneity effect, we focus mainly on the effects of mobility over the infection dynamics.

The literature on mobility impact over Covid-19 infection has focused mainly on two different aspects: prediction and causality. The first group focuses on predicting the impact of lockdown policies in evaluating Covid-19 spread. Within this class, there is a distinction in the methodology used on different papers: (a) adoption of synthetic controls; (b) adoption of alternative data sources and non-linear models. Our model resorts to alternative data sources without pursuing a synthetic control identification. The second group focused on panel data estimations to identify the causal relationship between mobility and Covid-19.

Regarding the usage of synthetic controls and artificial counterfactual (ArCo) approaches, we highlight Carneiro et al. (2020). The authors adopted an ArCo approach to verify the impacts of the short-run evolution of the number of cases (and deaths) at the state level in the United States. The prediction suggests that the number of cases would have been twofold in the absence of the restriction measures. On the same line, Bayat et al. (2020) recurs to a synthetic control methodology to analyze the effects of lockdown measures and the potential impact of those policies on the development of herd immunity. The authors suggest that reducing cases and deaths would have been much smaller if lockdown policies had been imposed earlier and

reopening postponed. Also, Born et al. (2021) resorting to a counterfactual approach finds that a lockdown policy during Covid-19 first wave in Sweden would have reduced cases by about 78% and deaths in 38%.

On the other side, we briefly describe papers that used alternative data sources to predict the impacts of lockdown policies. This is the case of Gerlee et al. (2021), Schwabe et al. (2021) and Vespe et al. (2021), recurring to alternative mobility data in order to assess its impacts on the disease infection levels. In particular, Barboza et al. (2021) aim to infer the effects of changes on mobility on the dynamics of the transmission of the Covid-19 in Costa Rica while using Google Mobility and Google Trends data to evaluate the effects of sanitary measures. The author finds evidence that reducing mobility can decrease the infection spread in Costa Rica while recurring to alternative data sources.

An important remark regards mobility variables used in these papers: independently of its source, mobility measures arrive from individual data aggregated at some level. Usually, mobility arrives from unstructured data sources, such as mobility data from telecommunication providers or centralized companies such as Google. Arnal et al. (2020) explores the differences between those data sources. In addition to using Google Mobility data as the baseline measure of mobility, this paper also adopts textual data sources to generate causal identification. As behavioral effects are non-observable, we suggest the usage of internet searches from Google Trends and newspaper articles from a national news platform (Globo - G1) as controls in our identification scheme. Up to this point, such usage of textual elements has not been adopted in other papers, and we argue that this inclusion displays a fundamental role in capturing the omitted variable bias to generate identification.

In terms of causality analysis, which constitutes a larger share of the empirical work and is based on different methodologies, the main objective is to assess the causal effects of mobility restrictions in terms of the evolution of the Covid-19 cases and deaths. Based on the availability of Covid-19 infection data, the models are predominantly analyzed in a panel of weekly cases and deaths (to reduce noise effects on daily published data) between or within countries. The usage of weekly variables observed at municipalities is the main focus of this paper, although some control variables are only available at the state level.

Methodologically, the usage of panel data motivated a debate regarding the validity of fixed-effects estimation (see Gauthier (2021)). For example, Liu et al. (2021) suggests the usage of a dynamic panel data model to generate forecasts to capture the inertial component that may affect the current

infection situation. The authors opt to model the growth rate of infections, assuming that this variable can be represented by fluctuations around a downward sloping deterministic trend (with a break).

As for the effects of reducing mobility over Covid-19 spread, Nouvellet et al. (2021) finds that over 52 countries, reducing mobility were able to decrease infection in 73% of the countries analyzed. Also, the ongoing literature of causal analysis reveals a consistent negative effect of mobility impacts over Covid-19 dynamics, reducing cases and deaths through different horizons. Finally, Huang (2020) makes use of the growth rate modeling based on counterfactual analysis to find that social distancing intervention is effective in reducing the weekly growth rate by 9.8% and deaths by 7.0% at the state-level in the United States.

In terms of Brazilian data, evidence supports the negative effect of mobility over Covid-19 infections. Chagas et al. (2021) while resorting to epidemiological shows that through simulations and hypothetical scenarios, the infection spread would be lower if the government imposed more mobility restrictions. More similarly to our methodology, Resende & Maciel (2021) explored a panel-data regression for São Paulo municipalities using labor market dynamics, medical infrastructure, and government transfers as controls. The authors found that increasing 1% on social distancing reduces infections by 4.14% in a week and diminishes deaths by 2.8% after two weeks.

The approach that we have adopted is inserted in the causal identification category with some elements of the prediction group, as we aim to identify how mobility (even in the absence of a strict lockdown) affected the infection evolution by recurring to alternative data sources to control for non-observable effects on a causal relation framework. We focus on modeling the level difference of Covid-19 cases and deaths (i.e., increase or decrease of cases and deaths on the reference week compared to the prior week) as our panel is inflated with equal counts (mainly on small counties), which generates a zero change from one week to another.

For the empirical modeling, we created a weekly based panel data for Brazilian municipalities to evaluate the effects of restrictions in mobility in terms of the pandemic evolution. In addition, our paper contributes to the ongoing literature of causal identification of mobility effects based on panel-data evaluation by adding textual data (Google Trends and News-indexes) to generate proxies for non-observable behavioral variables that affect the Covid-19 spread.

We considered a sample that comprehends 91 weeks and 5565 municipalities (i.e., 99.91% of all counties in Brazil) covering the period of April

2020 - December 2021 (significantly wider than the studies that focused on the effects of mobility). Estimation results suggest that increasing residential mobility (reducing overall mobility) significantly diminishes the number of cases and deaths over a four-week horizon. We also conduct separate sub-sample analyzes for 2020 and 2021, and the effects of reducing mobility are similar to the complete sample analysis, but the effects are lower for the first year of the pandemics. In addition, the results are robust to variations in mobility variables added to the model, geographical aggregation of cases, and different vaccination campaign variables.

The remainder of this paper is structured in four additional sections. The second section describes the structured and unstructured data sources used as regressors. The third section describes the identification strategy, the Direct Acyclic Graphs (DAG) illustration of the causal hypothesis, the fixed-effects model, and all variables used as controls for non-observable behavioral effects. The fourth section describes the estimation results for three different strategies: (i) a naive estimation considering only mobility as regressor without any controls; (ii) a complete sample estimation (2020-2021) with all suggested control variables; (iii) an estimation for each sub-sample of 2020 and 2021 threaten separately. We also include robustness estimations that corroborate the results obtained in the complete sample analysis. The last section concludes this paper.

1.2 Data

There are mainly two types of data that have been used to estimate our model. The first set denotes usual quantitative data, such as the number of cases and deaths due to Covid-19, vaccination campaign data, and even consolidated mobility data from Google. The second set denotes textual data from Google Trends or local newspaper articles that may or not be pre-processed in an early stage. Therefore, whenever the unstructured data is not pre-organized, we conduce an effort to generate a structured object to use as controls in the regression scheme.

The first set of hard indicators is the number of cases and deaths by Covid-19, which constitutes our target variables. Those have been extracted from SRAG data[1] available at the OpenDataSUS website.[2] The main difference

[1]SRAG is the acronym for “Vigilância de Síndrome Respiratória Aguda Grave", which consolidates all data related severe acute respiratory syndrome in Brazil, including Covid-19, for 2020 and 2021. Note that SRAG data only consider patients that effectively enter hospitals due to a respiratory syndrome and therefore does not represent mild cases.

[2]OpenDataSUS is an initiative of the Ministry of Health of Brazil.

between the construction of Covid-19 cases and deaths series regards the filtering date: for the number of cases, we have set the first symptom date as reference for aggregation, whereas for the number of deaths, we set the obit date as reference. In both cases, we construct a series at the municipality level based on the residence and notification area of the Covid-19 cases and deaths. Such distinction produces different aggregations as they constitute different hypotheses in terms of disease contamination process.[3] In Figure 1.1 we compare the effects of different types of date aggregation for the number of cases and deaths, e.g., first symptom (or obit) date versus notification date.

The next set of indicators comprehends mobility measures that constitute the object of interest in our estimations. These variables have been broken down into six mobility categories (workplace, residential, parks, transit, grocery, and retail). In terms of mobility measurement, Google Mobility Data considers Residence mobility as the time spent in-locus at home, whereas the other five categories are measured in terms of the number of visitors or individuals in each specific category. Those six mobility variables are consolidated, and each weekly window is compared to the 5-week baseline period of January 3, 2020, to February 6, 2020, in terms of the percentage change. We extract mobility data from the Covid-19 Community Mobility Reports from Google.

As for controls, we start by collecting vaccination microdata regarding timing and immunization type (first, second or third dose) also from the OpenDataSUS website. We also collect vaccination microdata from inside the SRAG data set used to collect cases and deaths. The two data sets differ as the first is broader and considers the national vaccination campaign, whereas the second only consider individuals inside the SRAG accounting (i.e., those infected by Covid-19). Finally, both types of vaccination data are consolidated into a panel of municipality observations at a weekly frequency.

The other two sets of variables, internet searches and news regarding Covid-19, correspond to our set of soft controls and are only observed at the state level. We used Google Trends data extracted using the Google Trends API for internet searches, which revealed the number of internet searches of a given topic for a certain period using the Google search mechanism. All series have been aggregated accordingly and normalized between zero and unity. To use news regarding Covid-19, we created a News-Index based on a dictionary

[3]The hypothesis of the contamination process are different depending on the aggregation. By filtering by residence area, we impose that patients are counted based on where they live; filtering by notification area imposes that counts are made where cases or deaths have been reported. However, if an individual is traveling, his residence does not coincide with its notification area. Also, regions have adopted different degrees of restrictions to mobility and also different vaccination campaigns.

method (similar to Baker et al. (2016)) from news collected from G1.[4] that possess an in-depth coverage of Covid-19 in Brazil[5]

These two sets of Google Trends and News controls (GT-series and N-index from now on) rely on a subjective categorization of search terms and keywords selection that needs to be specified to generate data series for our estimates. We adopted four categories denoted by (i) general Covid-19 evolution; (ii) fake-news regarding Covid-19; (iii) vaccination campaign evolution; (iv) prevention-related measures. Over each category, the terms selected for both GT-series and N-index are the same and display internal coherence in describing the same overall topic (e.g., usage of masks, washing hands, hand sanitizer belongs to the prevention group). The formulation of the indexes and the categories/keywords are presented in detail in Appendix A.

In Figure 2.1 we plot the evolution of the number of news regarding Covid-19 in Brazil, suggesting a particular common trend of a higher number of articles at the beginning of the pandemic and a subsequent increase in the first half of 2021, marking the second wave of cases as observed in Figure 1.1. We also present a comparison between the GT-series and N-index for each of the four categories described in Appendix A presented in Figure 1.4. It is important to notice that both control series present similar behavior but not identical patterns. For instance, Figure 1.4 presents the evolution of the number of news for the vaccination group, which is not matched by the evolution of internet searches in Google for the same terms on the period considered, i.e., the series displays positive correlation, but they are not perfectly correlated.

The overall data constitutes 91-weekly based observations from 5565[6] municipalities in Brazil starting on April 4, 2020, up to December 26, 2021. We opt to start the panel data in April 2020 and not in March 2020 to obtain a smoother series, as most small counties only displayed Covid-19 cases from April 2020 onwards.[7]

Tables 1.1 and 1.2 display descriptive statistics (from the whole period of 91 weeks and full 5565 municipalities) such as counts, mean, quantiles, and standard deviation of the variables used in the estimation. Table 1.3 consists of an additional descriptive table for vaccination variables solely for the 2021

[4]G1 is a local news platform that belongs to Grupo Globo.

[5]For April 2020 to December 2021, articles mentioning Covid-19 corresponds for more than 180,000 observations. The advantage of considering G1 articles is that they are divided into subregions, i.e., news data is locally stamped at the state level.

[6]Only five municipalities are not considered on the SRAG sample. The overall number of municipalities constitutes 99.91% of the total number of municipalities in Brazil, which is negligible for such a sample size.

[7]Between the end of February and March 2020 only big counties as São Paulo displayed cases. As our analysis focus on capturing the general effect of mobility over Covid-19 infection in Brazil, we opt to drop the first month of the pandemic to avoid local results.

year. As the national vaccination campaign only started in January 2021, all prior weeks were assigned with zero counting in vaccines administered. Such assignment is important while estimating the average effect of interest for the full 2020-2021 period. We also include in Tables 2.6 and 1.5 the correlation matrix for all variables used on the estimation for the same reference period of 2020-2021.

1.3 Identification Strategy

We motivate our identification strategy recurring to a Direct Acyclic Graph (DAG) approach, following Elwert (2013). In Appendix B, we provide an introduction to the concept of DAGs. In our specification, we want to estimate the impact of mobility on cases and deaths due to Covid-19. However, many confounding factors may affect mobility (or even the infection situation) that generates an omitted variable bias problem. To overcome such an issue, we specify carefully some of those factors following our hypothesis regarding the causal relationship between the variables.

We start by conjecturing that some preventative measures such as using masks, washing hands, and hand sanitizer may affect the virus infection, i.e., through individual behavior. However, such variable (denoted B, from now on) is non-observable. Therefore, to correctly evaluate the impact of mobility on the infection, we should include some control variables to capture this effect and clean the confounding factor. We use internet searches and news both regarding Covid-19 prevention behavior (denoted b from now on) in order to capture this omitted effect resulting in GT_b and N_b variables, respectively, each represented by coefficients $\boldsymbol{\gamma_b}$ and $\boldsymbol{\eta_b}$, where b denotes behavioral variables. We also consider that internet searches and news regarding the general Covid-19 situation (denoted g from now on) may reflect the widespread infection situation and can be affected by the evolution of the vaccination campaign, represented by GT_g and N_g variables, respectively. Therefore, we use such general situation variables as a proxy for lagged effects of Covid-19 infection situation in order to capture inertial effects. In Figure 1.5 we plot the DAG representing the causal relationship between the variables that we adopt in order to identify our model.

We then structure the channels that are represented on the DAG of Figure 1.5 in the following manner: we consider that vaccination may affect the number of cases and deaths, mobility, and individual behavior (measured through GT-series and N-index proxies to Covid-19 related keywords and searches for those behavioral channels). However, we also consider that these

GT-series and N-index should affect only mobility and may not affect the spread of the disease through direct channels. Therefore, mobility is mainly determined by vaccination and individual behavior (controlled by GT-series and N-index). Nonetheless, we consider that Covid-19 spread is determined by mobility, individual behavior, and vaccination.

However, if we isolate the first year of the pandemic (2020), the vaccination variable turns out to be innocuous (as the vaccination only started in Brazil by January 2021). Therefore, all the channels that relate vaccination with mobility measures, internet searches, and news regarding Covid-19 disappear from the DAG proposed in the Appendix. The result is the DAG presented in Figure 1.6, which is a simplified version of the first graph. Note that the identification of such a model is mainly a reduced form of the overall sample model but is still subject to omitted bias effects.

After motivating the causal relation that we aim to identify, the econometric specification adopted is straightforward. Given a panel data, where the indexes $j \in \{1, \cdots, J\}$ represent municipalities, $t \in \{t_0, \cdots, T\}$ is a weekly time-index and $m \in \{1, 2, 3, 4\}$ is the time-structure that we impose over the dependent variables of the model, denote:

$$\Delta Y_{j,t+m} = \beta_0 + \boldsymbol{X'_{j,t}}\boldsymbol{\beta} + \boldsymbol{Z_{j,t}} + \alpha_j + \delta_t + u_{j,t} \tag{1-1}$$

and $\boldsymbol{Z_{j,t}}$ denotes controls used in the overall fixed-effects estimate:

$$\boldsymbol{Z_{j,t}} = \boldsymbol{GT'_{j,t}}\boldsymbol{\gamma} + \boldsymbol{N'_{j,t}}\boldsymbol{\eta} + \boldsymbol{V'_{j,t}}\boldsymbol{\nu}$$

where $\Delta Y_{j,t+m} = Y_{j,t+m} - Y_{j,t+m-1}$ is the in-level difference between number of cases (or deaths) from week $t+1$ with respect to week t in municipality j. Also, consider $\boldsymbol{X'}_{j,t}$ as a vector of mobility measures for county j measured in percentage.[8] In terms of controls, $\boldsymbol{Z_{j,t}}$ represents a linear combination of GT-series ($\boldsymbol{GT'_{j,t}}$ a vector of internet searches from Google Trends composed by four categories, normalized between 0 and 1), N-index ($\boldsymbol{N'_{j,t}}$ a vector of News counts composed by four categories, normalized between 0 and 1) and vaccination campaign ($\boldsymbol{V'_{j,t}}$ is either a scalar or a vector of counts of doses applied on individuals depending on the vaccination data used). The coefficients $\beta_0, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\nu}, \alpha_j$ and δ_t are parameters to be estimated. Note that $\boldsymbol{\gamma} = (\boldsymbol{\gamma_g}, \boldsymbol{\gamma_b})$ and $\boldsymbol{\eta} = (\boldsymbol{\eta_g}, \boldsymbol{\eta_b})$ represent GT-series and N-index channels through general Covid-19 situation (g) and through proxies for non-observable behavioral effects (b), respectively. Also, α_j is an individual fixed-effect, whereas δ_t is a time fixed-effect. The constant β_0 imposes that the panel fixed

[8] To generate internal compatibility, all percentage variables have been transformed into level factors, i.e. 100% is presented as 1.

effects sum zero across all municipalities.[9]

In terms of the time-structure indicator $m \in \{1, 2, 3, 4\}$, we make two remarks. Firstly, this indicator has the objective of capturing delayed effects of regressors over Covid-19 cases and deaths. For example, vaccination may affect cases and deaths only after the immunization hits (about two weeks after the second dose). Secondly, by imposing non-overlapping subsequent weeks estimations (e.g., $m = 1$ denotes the change in the number of cases in the first week, $m = 2$ denotes the change in the number of cases in the second week and so on), we can construct an upper-bound measurement to estimate the cumulative impact of the regressors over the Covid-19 infection. This approach can deliver valuable insights at longer horizons while breaking down each weekly effect.

Focusing on the estimation methodology, we adopted a fixed-effect model imposing a two-way transformation, i.e., removing both time and individual effects. In this case, we used a within transformation with respect to each municipality, and we included time-dummies to capture heterogeneous time-effects throughout time. To validate the adoption of a fixed-effect model instead of a random-effect model (i.e., by considering that the unobservable factor is not correlated with the regressors), we used a Hausman test for selecting the models. All estimates used correction for residual correlation and robust variance estimators. Finally, we used Stata software to estimate all models used in this paper.

1.4 Results

The results of the estimations are divided into four different subsections. The first subsection constitutes the "naive estimation" as it focuses on direct effects of mobility over Covid-19 cases (deaths) through all samples (2020-2021) without resorting to the usage of any controls. The second subsection analyzes the "overall estimation" focusing on estimating the effects of mobility over Covid-19 cases and deaths following the specification of Equation (1-1) adopting controls for general and behavioral channels. The third subsection restricts the model to two separate sub-sample periods (2020 and 2021) to separate the estimation coefficients. Finally, the last subsection consists

[9] All estimations include an intercept due to the normalization adopted by Stata. Instead of setting the intercept equal to zero, the program adopts the following normalization:

$$\sum_{j=1}^{N} \sum_{t=t_0}^{T} (\beta_0 + \alpha_j) = 0$$

of robustness checks adopting different geographical aggregations, different vaccination variables, and the usage of all mobility measures from Google Mobility Reports. However, restrictions to mobility displayed a relevant impact on the evolution of the Covid-19 infection in all cases.

Prior to estimating Equation (1-1), we need to make some considerations regarding the number of lags $m \in \{1, 2, 3, 4\}$ used on our specification. In terms of Covid-19 cases, it is known that the average number of days taken up to the first symptom is about five days (see Cintra & Fontinele (2020)). Therefore, inside the 5-days window, the individual may spread the virus without knowing his infection situation. In terms of deaths, we computed the median number of days between the first symptom and obit, represented in Figure 1.2. The results suggest that, in the SRAG sample, the median is about seventeen days (third week onwards), while the minimum is about eight days (second week) and the maximum is twenty-two days (fourth week). As we consider weekly windows, the relevant horizon that we may consider for cases is from one to four weeks after the infection and two up to four weeks for deaths. However, we opt to display all four-week lags for cases and deaths, as we observe that relevant estimates for deaths only tend to appear after the third week.

Regarding mobility measures, Table 1.1 reveals that only workplace mobility is present in almost all municipalities that are available on Google's Mobility website, while residential mobility is also present in a high share of overall municipalities. The other mobility measures are not observed throughout the sample, which impacts the panel size significantly and generates a higher probability of missing values in the estimates. We, therefore, choose to run the regressions only on the first two residential and workplace mobility measures through the estimations. We also present a model with all six mobility measures, but all the results that we find are robust to this inclusion.

1.4.1
Naive Estimation: Mobility only

Prior to estimating the complete model, we estimated a simple model as described in Equation (1-1) but without any control variables (i.e. excluding $\boldsymbol{Z_{j,t}}$ from the equation). The idea of such estimation is to motivate the adoption of a model with all control variables, as we suggest in the DAG of Figure 1.5 to address the omitted variable bias correctly.

The naive estimation results are displayed in Table 1.6, based on fixed-effects estimations for Covid-19 cases (columns 1 to 4) and deaths (columns 5 to 8). In each column, we highlight the number of lags m growing from $m = 1$ up to $m = 4$ weeks for cases and deaths. The first important thing to notice

is the negative sign associated with the coefficient of residential mobility and the positive sign associated with the coefficient of workplace mobility. Those results suggest that an increase in residential mobility (i.e., a decrease in overall mobility) decreases the number of cases and deaths for the whole 2020-2021 sample. As we analyze in the next subsection, magnitudes of the coefficients are under-estimated compared to the full control model. Our hypothesis to explain such an effect is that mobility embeds different elements that may attenuate the effects over the infection spread in the absence of control variables.

There are introductory remarks to this impact analysis of mobility over Covid-19 infections, even with such a simple model. The first element that we highlight is that the associated R^2 measure for each estimate is low since we opt to model the variation in the number of cases from one week to another despite modeling the level cases.[10] Another important fact is that even with such a simple model, we reject the hypothesis that all coefficients estimated are statically zero. Another important fact is the Hausman test[11] used to select fixed or random-effects models suggests that for $m = 4$ for cases and $m = 1$ for deaths, the fixed-effect model is consistent but not efficient and rejects the null for all other time structures. Therefore, the Hausman test provides additional statistical evidence for a fixed-effects estimate.

1.4.2 Overall Estimation: Full Model

In this subsection, we estimate the main specification of this paper, in which we adopt soft data variables as controls in our regression. The overall estimation results are displayed in Table 1.7, based on fixed-effects estimations for Covid-19 cases (columns 1 to 4) and deaths (columns 5 to 8). In each column, we highlight the number of lags m growing from $m = 1$ up to $m = 4$ weeks for cases and deaths. The coefficient of interest is the residential and workplace mobility, as they are associated with the highest number of observations for municipalities. The first important result is that increasing residential mobility (i.e., decreasing overall mobility) reduces the sample's number of cases and deaths.

The impact over Covid-19 cases is higher over the first reference week, where a 100% increase in workplace mobility diminishes at most 60.36 cases and slowly decreases, reaching a 26.94 reduction in four weeks. Regarding

[10]Modelling the level of cases generates a higher fit but wrongly incorporates inertial behavior intrinsic to the pandemic evolution.

[11]The Hausman test is used to test through models in panel data estimates. In the null hypothesis, random effects estimate is more efficient than fixed effects. The alternative hypothesis is that random effects estimates are not consistent.

Covid-19 deaths, we also observe that increasing residential mobility reduces the number of deaths. However, this effect is only significant after two weeks. In this case, a 100% reduction in residential mobility reduces 19.89 deaths in one week and 26.07 deaths in four weeks.

The results are similar for workplace mobility: reducing workplace mobility negatively affects the number of cases and deaths but with a lower impact than increasing residential mobility. The result suggests that the combined effect of reducing overall mobility (i.e., increasing residential mobility and decreasing workplace mobility) has substantial effects while reducing the number of cases and deaths over the four-week horizon. These effects diminish over time for cases and increase the coefficients for deaths, suggesting that the time structure is fundamental to map the heterogeneous effects depending on the time horizon.

Considering that this effect can be combined week-by-week (inducing an over-estimate of the overall four-week effect), we can determine that the (maximum) upper bound effect of a 100% increase in residential mobility (ceteris paribus) results in a cumulative reduction of 177.42 cases (over the four weeks) and a cumulative reduction of 73.38 deaths (over three weeks). In terms of a 100% decrease in workplace mobility, the combined effects for cases are a reduction of 28.63 cases and 14.13 deaths. On average[12], considering a conservative scenario where the average residential mobility increased only 7% and workplace mobility was 0%, the weighted effect over four weeks is a 12.41 diminish in cases and a 5.13 reduction in the number of deaths. Considering that the average number of cases was 4.67 and the average number of deaths on the period was 1.8, the reduction in mobility impact over Covid-19 infections is significant.

The vaccination, GT-series, and N-index controls display an essential role while capturing the effects suggested on the DAGs compared to the naive regression of Table 1.6. The R-squared is significantly higher (mainly the Between R^2 which describes the adjustment over the time-demeaned the model), the F-test rejects the null hypothesis that all coefficients are zero for all estimates. More importantly, all vaccination coefficients are negative, which indicates that an increase in the number of individuals vaccinated diminishes the number of cases and deaths for the following weeks. Finally, the Hausman tests reject the null hypothesis through all regressions, and the fixed-effects estimate is both consistent and efficient.

[12]Following the descriptive statistics of Table 1.1.

1.4.3 Sub-sample Estimations: 2020 and 2021

This subsection has the objective of estimating the impact of mobility on Covid-19 cases (deaths) in two separate sub-samples: 2020 and 2021. The first sub-sample marks the first year of the Covid-19 pandemic, whereas the second sub-sample marks the beginning of the vaccination campaign in Brazil. Therefore, limiting the sample solely to 2020 and 2021 can redefine the causal relation that we aim to identify, as the vaccination variable plays a different role in each sub-period. Furthermore, as described in Section 1.3, the analysis of the first year of the pandemic has the natural advantage of isolating the effect of the vaccination campaign from the complete model. Thus, we can remove the vaccination node from the DAG and solely focus on mobility-related variables, as presented in the simplified DAG of Figure 1.6.

The estimation results for the first sub-sample (2020 only) are presented in Table 1.8. The results suggest three common elements: (i) the negative sign associated with the residential mobility coefficient and the positive sign associated with workplace mobility is still present, and all coefficients are statistically significant at the 1% level; (ii) the magnitude of the effects is much lower than the ones from the overall estimation displayed in Table 1.7, accounting for almost half of the size of the complete sample estimate; (iii) the coefficients associated with soft data variables aimed to control for the vaccination through the behavioral channel mostly do not display statistical significance.

For the second sub-sample analysis (2021 only) estimation, the DAG which motivates the causal identification scheme is the same as the complete sample model, present in Figure 1.5. The two main differences are the sample size and the exclusion of the all-zero vaccination variable through the first year of the pandemic.[13] The estimation results for the 2021 sample are presented in Table 1.9. The results are similar to the complete sample model: (i) the negative sign associated with residential mobility coefficient and the positive sign associated with workplace mobility are still present and significant; (ii) the dimension of the effects is compatible with the ones displayed in Table 1.7; (iii) the immunization provided by the vaccination campaign displayed an important role while controlling the number of cases and deaths in the second year of the pandemic.

The two sub-sample estimations reveal that the overall impact of residential and workplace mobility has an essential role in controlling the infection

[13]To estimate an average effect for vaccination variable through the complete sample, we threaten the missing observations as zeros, and the effects are downsized by construction.

and deaths due to Covid-19 over the last years. Another important insight from the estimation results is that after controlling by vaccination, the observed effects of residential mobility have grown in size, and reductions in workplace mobility revealed a less prominent role while maintaining the infections at a stable level. Finally, the behavioral controls such as the evolution of searches and news regarding vaccination used to capture the effects of non-observable factors displayed statistical significance on the later estimation, suggesting the existence of the causal channel present on DAG of Figure 1.5.

1.4.4 Robustness

After estimating the effects of mobility on the infection dynamics, we provide some additional robustness to validate the results obtained on the overall and sub-sample estimations in this subsection. Next, we address the geographical aggregation of cases and deaths in terms of notification area despite the residence areas adopted over the last subsections. Next, we supply a full estimate of all six mobility measures to argue that the inclusion of the other categories does not affect the qualitative properties (the sign of the coefficients) of the estimated model but only the quantitative aspects (size of the coefficients). Lastly, we expand the vaccination data and recur to national vaccination campaign data from OpenDataSUS to capture spillover effects of the immunization.

The first robustness estimation aggregates Covid-19 cases and deaths by notification area instead of residence area. Such change is vital to address individuals not at their residence counties when diagnosed with Covid-19 inside the SRAG data accounting (e.g., traveling individuals). Results are shown in Table 1.10 and are closely related in terms of qualitative and quantitative aspects of the estimations present in Table 1.7.

The second robustness estimation adds all six mobility measures as regressors in the model suggested in Equation (1-1). The estimation output is displayed in Table 1.11 and the qualitative aspects described on the baseline regression follow. The main differences while adding the mobility measures consist in: (i) the cumulative impact of a 7% increase in residential mobility and a 0% change in workplace mobility over the four reference weeks is a reduction in 25.87 cases and a reduction in 10.91 deaths, i.e., more significant than the results of Table 1.7 suggests; (ii) the sample is significantly smaller due to the lack of observations in small counties. These results suggest that in larger municipalities, the effects of a reduction in mobility can be even more

significant than the average effects of Table 1.7.[14]

Finally, we added the national vaccination campaign for Covid-19 as an alternative for the SRAG vaccination data that only consider individuals inside hospitals accounting. The results are displayed in Table 1.12 and consider the number of vaccines administered at each county and each week in terms of first, second or third dose. The sign and magnitude of the mobility coefficients are similar to the ones present in Table 1.7. It is also interesting to note that mainly first and third doses revealed negative coefficients for the estimation result while second dose data displayed much lower coefficients and mixed signs. However, the combined effects of the vaccination coefficient are negative, as expected.

1.5 Conclusion

In this paper, we aimed to develop an empirical framework to address the questions related to the causal effects of mobility restrictions in terms of impacts over Covid-19 cases and deaths. First, by recurring to a DAG to illustrate the causal hypothesis, we conjectured causal channels that required the design of “soft data" proxies to capture non-observable effects of individuals. Then, we estimated a simple naive model and compared it to the adoption of soft data controls in terms of qualitative results (the sign of the coefficients) and qualitative properties (size of the coefficients).

The methodology adopted to estimate the effects is in line with the panel-data estimates by Liu et al. (2021) and Huang (2020), but instead of modeling the change in the growth rate, we opt to model the change in the level number of cases. Our results suggest that a 1% reduction in mobility (increase in residential mobility) reduces the number of cases and deaths in a one to a four-week horizon for both the complete sample (2020-2021) and over each sub-sample. The average ceteris paribus effect of an increase in residential mobility over the four-week horizon reduces 12.41 cases and 5.13 deaths. This effect is significant compared to the average of 4.67 cases and 1.8 deaths over the whole sample.

Our analysis has some limitations based on potential noise in data and non-linearity on the effect of vaccination over mobility: vaccines can induce higher or lower mobility depending on the overall vaccination campaign or even on the Covid-19 infection level. We, therefore, avoid interpreting vaccination-related coefficients as they can be misleading. The same happens

[14]Only large municipalities have data for all six-mobility measures. Therefore, when estimating a regression using all variables, smaller counties are ruled off the sample, and therefore the result is biased towards average effects over large municipalities.

while analyzing other mobility coefficients due to the intrinsic lower number of observations and possible measurement error in those variables.

Finally, the paper suggests that restricting mobility can reduce the number of cases and deaths with particular robustness throughout the sample and methodological evaluation. The objective of developing a causal framework based on a DAG illustration to identify the model is sustained by the estimation outputs, and the mobility effects over the number of cases and deaths are solely determined by the causal relation conjectured through the identification strategy.

Tables

Table 1.1: Descriptive Statistics for Model's Variables (Part I)

	cases	deaths	srag_vac*	1st_dose*	2nd_dose*	3rd_dose*	residential	workplace	transit	parks
count	393066	323880	393066	398304	398304	398304	84040	175421	52658	70924
mean	4.67	1.80	0.54	382.61	316.01	38.19	0.07	0.00	-0.23	-0.27
std	41.66	14.69	6.52	4557.98	4241.88	1021.44	0.05	0.16	0.36	0.32
min	0.00	0.00	0.00	0.00	0.00	0.00	-0.71	-0.77	-1.00	-1.00
25%	0.00	0.00	0.00	0.00	0.00	0.00	0.04	-0.08	-0.47	-0.50
50%	1.00	0.00	0.00	1.00	0.00	0.00	0.07	0.00	-0.27	-0.30
75%	2.00	1.00	0.00	157.00	87.00	0.00	0.10	0.09	-0.06	-0.08
max	5676	1635	853	899970	691182	194269	0.33	0.83	3.88	3.90

[a] The descriptive statistics table reveals the statistics in relation to municipality level variables (5565 counties) and state variables (27 states) for the 91-week period from April 4, 2020, up to December 26, 2021.
[*] The descriptive statistics for vaccination consider the whole 2020-2021 sample, i.e. prior to the beginning of the vaccination campaign, zero individuals have been vaccinated. This framework is used to understand the whole sample analysis (i.e. 2020-2021 estimation).

Table 1.2: Descriptive Statistics for Model's Variables (Part II)

	grocery	retail	n_covid	n_prevention	n_fakenews	n_vaccine	gt_covid	gt_prevention	gt_fakenews	gt_vaccine
count	70595	74629	2457	2457	2457	2457	2457	2457	2457	2457
mean	0.20	-0.18	0.19	0.04	0.01	0.08	0.39	0.23	0.21	0.28
std	0.26	0.25	0.13	0.11	0.07	0.13	0.22	0.16	0.21	0.25
min	-0.91	-0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.04	-0.34	0.09	0.00	0.00	0.00	0.21	0.11	0.06	0.10
50%	0.19	-0.17	0.18	0.00	0.00	0.00	0.38	0.18	0.16	0.17
75%	0.36	-0.02	0.25	0.00	0.00	0.14	0.54	0.30	0.30	0.40
max	1.67	1.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

[a] The descriptive statistics table reveals the statistics in relation to municipality level variables (5565 counties) and state variables (27 states) for the 91-week period from April 4, 2020, up to December 26, 2021.

Table 1.3: Descriptive Statistics for Vaccination Variables (2021 only)

Descriptive Statistics

	srag_vac	1st_dose	2nd_dose	3rd_dose
count	224985	208685	208685	208685
mean	0.94	730.20	603.14	72.89
std	8.59	6276.77	5845.52	1410.26
min	0	0	0	0
25%	0	48	13	0
50%	0	143	77	0
75%	0	405	290	0
max	853	899970	691182	194269

Correlation Table

	srag_vac	1st_dose	2nd_dose	3rd_dose
srag_vac	1.00	-	-	-
1st_dose	0.72	1.00	-	-
2nd_dose	0.57	0.59	1.00	-
3rd_dose	0.15	0.06	0.45	1.00

[a] The descriptive statistics table reveals both statistics and correlation between vaccination variables (SRAG data versus national vaccination campaign data) from the 5565 counties and for the 51-week period from January 3, 2021, up to December 26, 2021.

Table 1.4: Correlation Matrix for Model's Variables (Part I)

	cases	deaths	srag_vaccine	1st_dose	2nd_dose	3rd_dose	residential	workplace	transit	parks
cases	1.00	-	-	-	-	-	-	-	-	-
deaths	**0.89**	1.00	-	-	-	-	-	-	-	-
srag_vaccine	0.14	0.18	1.00	-	-	-	-	-	-	-
1st_dose	-0.10	-0.12	**-0.71**	1.00	-	-	-	-	-	-
2nd_dose	-0.01	-0.02	-0.31	0.38	1.00	-	-	-	-	-
3rd_dose	-0.02	-0.03	-0.30	0.37	0.36	1.00	-	-	-	-
residential	-0.04	-0.05	**-0.53**	**0.64**	0.47	**0.53**	1.00	-	-	-
workplace	-0.10	-0.12	**-0.68**	**0.74**	**0.50**	**0.56**	**0.74**	1.00	-	-
transit	**0.53**	**0.55**	0.03	0.01	0.06	0.04	0.06	0.00	1.00	-
parks	0.41	0.42	0.02	0.02	0.06	0.05	0.07	0.02	**0.72**	1.00
grocery	0.24	0.25	-0.02	0.05	0.08	0.07	0.09	0.05	**0.57**	**0.59**
retail	0.03	0.04	-0.03	0.04	0.07	0.05	0.06	0.05	0.15	0.06
n_covid	0.03	0.04	0.32	-0.28	-0.19	-0.15	-0.26	-0.33	0.00	0.00
n_prevention	0.03	0.04	0.34	-0.30	-0.18	-0.11	-0.26	-0.31	-0.01	-0.01
n_fakenews	0.01	0.01	0.03	-0.09	0.00	0.01	-0.02	-0.03	0.00	0.00
n_vaccine	0.02	0.02	-0.20	0.16	0.10	0.04	0.24	0.14	0.06	0.06
gt_covid	0.05	0.06	0.29	-0.35	-0.22	-0.26	-0.26	-0.39	0.04	0.03
gt_prevention	0.04	0.04	0.45	-0.45	-0.29	-0.24	-0.43	-0.47	-0.02	-0.02
gt_fakenews	0.04	0.06	0.35	-0.39	-0.26	-0.26	-0.35	-0.42	0.00	-0.01
gt_vaccine	0.02	0.03	-0.15	0.14	0.10	0.02	0.25	0.12	0.08	0.08

[a] This table reveals the correlation between model's variables. Coefficients higher than 0.5 in modulus are exhibited in bold. Note that mobility measures display internal consistence: by construction, residence mobility is negatively correlated to all non-residential mobility series.

Table 1.5: Correlation Matrix for Model's Variables (Part II)

	grocery	retail	n_covid	n_prevention	n_fakenews	n_vaccine	gt_covid	gt_prevention	gt_fakenews	gt_vaccine
cases	-	-	-	-	-	-	-	-	-	-
deaths	-	-	-	-	-	-	-	-	-	-
srag_vaccine	-	-	-	-	-	-	-	-	-	-
1st_dose	-	-	-	-	-	-	-	-	-	-
2nd_dose	-	-	-	-	-	-	-	-	-	-
3rd_dose	-	-	-	-	-	-	-	-	-	-
residential	-	-	-	-	-	-	-	-	-	-
workplace	-	-	-	-	-	-	-	-	-	-
transit	-	-	-	-	-	-	-	-	-	-
parks	-	-	-	-	-	-	-	-	-	-
grocery	1.00	-	-	-	-	-	-	-	-	-
retail	0.45	1.00	-	-	-	-	-	-	-	-
n_covid	-0.02	-0.03	1.00	-	-	-	-	-	-	-
n_prevention	-0.02	-0.01	0.21	1.00	-	-	-	-	-	-
n_fakenews	0.00	0.00	0.04	0.02	1.00	-	-	-	-	-
n_vaccine	0.03	0.00	0.34	-0.08	-0.02	1.00	-	-	-	-
gt_covid	-0.03	-0.05	0.42	0.17	0.02	0.33	1.00	-	-	-
gt_prevention	-0.04	-0.03	0.39	0.28	0.02	-0.04	**0.56**	1.00	-	-
gt_fakenews	-0.04	-0.03	0.35	0.22	0.06	0.02	**0.70**	0.46	1.00	-
gt_vaccine	0.02	-0.03	0.14	-0.13	-0.01	**0.61**	**0.59**	-0.06	0.10	1.00

[a] This table reveals the correlation between model's variables. Coefficients higher than 0.5 in modulus are exhibited in bold. Note that controls for same category display positive correlation as expected (e.g. prevention series from Google Trends and prevention index from news).

Table 1.6: Results for Mobility Regressors: Naive Model

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-76.76***	-74.25***	-60.63***	-47.45***	-4.622*	-20.88***	-31.43***	-32.39***
	(11.28)	(11.30)	(9.561)	(9.010)	(2.678)	(4.561)	(5.654)	(5.482)
workplace	8.096***	6.964***	5.155***	3.095***	1.305***	2.885***	3.719***	3.765***
	(1.580)	(1.563)	(1.155)	(1.059)	(0.437)	(0.678)	(0.724)	(0.721)
_cons	21.85***	23.17***	22.11***	18.17***	7.325**	8.581***	12.65***	17.31***
	(3.647)	(4.197)	(4.022)	(3.656)	(3.080)	(2.196)	(2.922)	(4.019)
Within R^2	0.033	0.033	0.031	0.029	0.023	0.026	0.031	0.032
Between R^2	0.012	0.011	0.019	0.012	0.037	0.013	0.002	0.016
Overall R^2	0.024	0.025	0.026	0.026	0.022	0.020	0.021	0.021
Observations	82460	81651	81647	81649	77022	76615	76601	76604
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.546	1.000	0.000	0.000	0.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the complete sample (2020-2021) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by residence area) and residential and workplace mobility as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

Table 1.7: Estimation for Complete Sample (2020-2021)

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-60.36***	-52.79***	-37.33***	-26.94***	-6.074	-19.89***	-27.42***	-26.07***
	(10.76)	(9.894)	(8.585)	(8.050)	(3.795)	(5.137)	(5.501)	(4.946)
workplace	10.15***	8.081***	6.295***	4.111***	2.155***	3.640***	4.261***	4.082***
	(1.957)	(1.998)	(1.490)	(1.114)	(0.565)	(0.798)	(0.810)	(0.779)
srag_vac	-0.294***	-0.302***	-0.304***	-0.263***	-0.0272*	-0.0461***	-0.0697***	-0.0821***
	(0.0849)	(0.0889)	(0.0899)	(0.0739)	(0.0156)	(0.0176)	(0.0208)	(0.0240)
n_covid	0.345	-2.408**	-2.366**	-2.297**	1.906***	0.477	0.129	-0.881*
	(1.072)	(1.057)	(1.165)	(1.131)	(0.730)	(0.524)	(0.367)	(0.499)
n_prevention	-0.0452	-0.948	-2.042**	-1.611**	-0.143	-0.403	-0.786	-0.897*
	(0.918)	(0.918)	(0.987)	(0.649)	(0.490)	(0.310)	(0.482)	(0.511)
n_fakenews	-1.535	-0.304	2.728	3.409	-0.695	-0.468	-0.683	0.320
	(1.540)	(2.274)	(3.304)	(3.875)	(0.561)	(0.566)	(0.990)	(1.611)
n_vaccines	1.272	1.723	2.624**	2.112**	-1.672***	-0.477	-0.152	1.126**
	(1.122)	(1.522)	(1.085)	(0.901)	(0.619)	(0.528)	(0.400)	(0.477)
gt_covid	-1.351	-6.964***	-8.851***	-8.228***	4.297***	2.246***	-0.0712	-1.371*
	(1.615)	(1.964)	(1.996)	(1.708)	(1.181)	(0.820)	(0.639)	(0.713)
gt_prevention	3.300**	3.032***	-1.179	-2.427	0.528	1.345*	1.081**	0.723*
	(1.345)	(0.838)	(1.093)	(1.478)	(0.597)	(0.770)	(0.469)	(0.378)
gt_fakenews	-2.551**	-4.544***	-2.772***	-1.495	0.936**	0.364	-0.516	-1.775***
	(1.295)	(1.152)	(0.926)	(0.977)	(0.373)	(0.652)	(0.660)	(0.453)
gt_vaccines	0.779	3.667***	3.683***	3.110***	-1.865***	-0.991**	0.446	0.530*
	(0.956)	(0.979)	(0.886)	(0.656)	(0.527)	(0.461)	(0.326)	(0.295)
trend	-0.221***	-0.259***	-0.282***	-0.227***	-0.0605*	-0.0804***	-0.148***	-0.195***
	(0.0460)	(0.0534)	(0.0542)	(0.0470)	(0.0319)	(0.0234)	(0.0351)	(0.0486)
_cons	17.28***	21.23***	22.78***	19.65***	4.518*	6.176***	10.75***	16.04***
	(3.773)	(4.427)	(4.548)	(4.044)	(2.506)	(1.834)	(2.958)	(4.185)
Within R^2	0.066	0.070	0.070	0.058	0.028	0.033	0.045	0.054
Between R^2	0.313	0.395	0.455	0.366	0.080	0.118	0.109	0.213
Overall R^2	0.050	0.055	0.060	0.056	0.026	0.025	0.032	0.041
Observations	82460	81651	81647	81649	76286	76214	76064	75976
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the complete sample (2020-2021) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by residence area) and residential and workplace mobility together with SRAG vaccination, Google Trends and News as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

Table 1.8: Estimation for Sub-Sample I: 2020 only

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-35.80***	-45.98***	-39.27***	-34.59**	-2.236	-14.08***	-22.52***	-21.83***
	(9.034)	(10.29)	(12.05)	(14.79)	(4.034)	(4.837)	(5.946)	(6.734)
workplace	24.09***	21.65***	22.20***	16.12***	7.947***	9.043***	10.16***	10.42***
	(5.256)	(5.730)	(5.534)	(5.257)	(2.234)	(2.149)	(2.373)	(2.843)
n_covid	1.155	-1.491	-1.998	-3.107*	2.842**	1.761*	0.776	-0.341
	(1.494)	(1.393)	(1.771)	(1.651)	(1.348)	(0.962)	(0.713)	(0.698)
n_prevention	-0.124	-1.134	-2.966***	-1.082	-0.345	-0.837*	-1.396**	-1.473***
	(1.055)	(0.825)	(0.997)	(0.708)	(0.559)	(0.508)	(0.620)	(0.569)
n_fakenews	3.372	7.339	16.34	19.17	0.122	1.306	3.216	6.339
	(4.417)	(8.398)	(13.27)	(15.46)	(2.403)	(2.272)	(4.928)	(8.984)
n_vaccines	-2.656	-3.094	-0.204	2.606	-6.234*	-2.251	-2.868	-3.068
	(6.103)	(5.763)	(3.426)	(3.876)	(3.306)	(2.514)	(2.448)	(2.621)
gt_covid	4.541	-2.207	-8.939***	-11.79***	6.180***	4.075**	2.937	0.619
	(3.195)	(2.518)	(3.109)	(3.049)	(1.688)	(1.812)	(1.801)	(1.348)
gt_prevention	0.964	3.235**	3.769**	1.328	-0.642	-0.404	-0.618	0.314
	(1.264)	(1.315)	(1.626)	(1.561)	(0.678)	(0.841)	(0.711)	(0.685)
gt_fakenews	-4.420*	-6.325***	-4.271*	-2.234	-0.0476	-0.855	-2.523*	-2.654**
	(2.344)	(2.429)	(2.536)	(1.825)	(0.506)	(0.805)	(1.372)	(1.170)
gt_vaccines	-0.777	1.135	3.511**	1.882	-0.440	-0.594	0.0678	-0.653
	(2.559)	(2.826)	(1.743)	(1.781)	(1.231)	(1.245)	(1.029)	(1.941)
trend	-0.435***	-0.486***	-0.596***	-0.557***	-0.239***	-0.237***	-0.325***	-0.431***
	(0.103)	(0.118)	(0.137)	(0.121)	(0.0843)	(0.0680)	(0.0860)	(0.120)
_cons	16.55***	22.74***	25.95***	24.43***	5.007**	6.906***	11.77***	16.58***
	(3.763)	(4.838)	(5.347)	(4.943)	(2.397)	(2.063)	(3.187)	(4.448)
Within R^2	0.023	0.027	0.032	0.030	0.023	0.021	0.027	0.029
Between R^2	0.034	0.021	0.004	0.005	0.019	0.027	0.022	0.001
Overall R^2	0.012	0.015	0.019	0.023	0.016	0.012	0.015	0.018
Observations	31687	31955	31955	31955	28564	29232	29232	29232
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.000	0.023	0.000	0.000	0.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the first sub-sample (2020) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by residence area) and residential and workplace mobility together with SRAG vaccination, Google Trends and News as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

Table 1.9: Estimation for Sub-Sample II: 2021 only

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-82.75***	-57.90***	-33.94***	-19.26**	-7.119	-26.48***	-36.10***	-31.48***
	(14.32)	(10.04)	(10.57)	(9.634)	(6.270)	(7.510)	(6.739)	(4.552)
workplace	7.820***	6.360**	2.859	0.731	1.144**	2.390***	2.615***	2.819***
	(2.285)	(2.670)	(1.750)	(1.743)	(0.459)	(0.701)	(0.832)	(1.021)
srag_vac	-0.389**	-0.402**	-0.423***	-0.413***	-0.00262	-0.0433	-0.0862*	-0.114**
	(0.169)	(0.162)	(0.155)	(0.135)	(0.0338)	(0.0408)	(0.0473)	(0.0471)
n_covid	0.817	-2.403**	-2.000	-0.729	1.383**	-0.213	0.0671	-0.876*
	(1.013)	(1.087)	(1.400)	(0.949)	(0.687)	(0.504)	(0.322)	(0.468)
n_prevention	0.352	0.0995	0.517	-1.331	0.117	0.317	0.315	0.202
	(1.434)	(1.663)	(1.462)	(1.045)	(0.889)	(0.772)	(0.526)	(0.559)
n_fakenews	-3.543**	-3.159**	-2.091**	-2.412*	-0.714*	-0.842*	-1.667***	-1.109*
	(1.451)	(1.365)	(1.052)	(1.296)	(0.389)	(0.438)	(0.597)	(0.623)
n_vaccines	0.601	1.359	2.298**	1.260	-1.485**	-0.413	-0.449	1.005**
	(0.874)	(1.403)	(0.969)	(0.846)	(0.625)	(0.520)	(0.315)	(0.442)
gt_covid	-3.606	-9.656***	-8.333***	-4.808**	3.517**	1.705**	-1.341	-2.157*
	(2.564)	(3.042)	(2.994)	(1.996)	(1.588)	(0.770)	(1.165)	(1.155)
gt_prevention	3.941*	2.936**	-3.694*	-4.144**	1.227	2.316**	1.989***	0.971*
	(2.156)	(1.399)	(2.122)	(2.097)	(0.751)	(0.938)	(0.592)	(0.539)
gt_fakenews	-2.296	-3.908***	-1.554*	-0.430	1.603**	1.078	0.554	-1.534***
	(1.499)	(1.323)	(0.934)	(1.584)	(0.660)	(0.948)	(0.718)	(0.420)
gt_vaccines	1.448	4.439***	3.120***	1.674*	-2.026***	-1.085**	0.733*	0.655
	(1.394)	(1.131)	(0.956)	(0.959)	(0.730)	(0.482)	(0.434)	(0.415)
trend	-0.229***	-0.194***	-0.126***	-0.0420**	0.0201*	-0.0361**	-0.0995***	-0.0882***
	(0.0440)	(0.0392)	(0.0305)	(0.0177)	(0.0118)	(0.0151)	(0.0230)	(0.0201)
_cons	18.05***	15.69***	9.464***	3.473**	-1.722**	2.750***	6.726***	6.733***
	(3.284)	(2.833)	(2.711)	(1.670)	(0.703)	(0.908)	(1.587)	(1.309)
Within R^2	0.093	0.097	0.099	0.091	0.030	0.038	0.060	0.085
Between R^2	0.610	0.573	0.614	0.385	0.039	0.275	0.314	0.284
Overall R^2	0.064	0.072	0.075	0.066	0.029	0.031	0.042	0.057
Observations	50773	49696	49692	49694	47722	46982	46832	46744
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the second sub-sample (2021) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by residence area) and residential and workplace mobility together with SRAG vaccination, Google Trends and News as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

Table 1.10: Estimation for Complete Sample (2020-2021): Notification Area

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-63.77***	-56.00***	-40.74***	-30.44***	-3.315	-21.37***	-30.17***	-30.16***
	(12.40)	(11.72)	(10.85)	(10.46)	(4.528)	(6.042)	(6.427)	(5.997)
workplace	12.90***	10.53***	8.920***	5.501***	3.145***	4.762***	5.645***	5.278***
	(2.812)	(2.951)	(2.315)	(1.613)	(0.788)	(1.066)	(1.094)	(1.080)
srag_vac	-0.317***	-0.327***	-0.327***	-0.283***	-0.0354**	-0.0527***	-0.0757***	-0.0876***
	(0.0874)	(0.0913)	(0.0924)	(0.0762)	(0.0138)	(0.0157)	(0.0200)	(0.0230)
n_covid	0.794	-3.247**	-2.849*	-3.090**	2.491***	0.784	0.150	-1.207*
	(1.498)	(1.468)	(1.588)	(1.527)	(0.931)	(0.721)	(0.541)	(0.687)
n_prevention	-0.796	-0.946	-2.258**	-1.911**	-0.231	-0.680*	-0.826	-1.116*
	(1.237)	(1.182)	(1.150)	(0.778)	(0.551)	(0.395)	(0.571)	(0.627)
n_fakenews	-1.560	-0.285	3.294	4.277	-0.944	-0.340	-0.794	0.463
	(1.984)	(2.981)	(4.270)	(4.974)	(0.721)	(0.688)	(1.355)	(2.134)
n_vaccines	1.535	2.051	3.119**	2.886**	-2.103**	-0.817	-0.264	1.650**
	(1.379)	(2.095)	(1.432)	(1.291)	(0.815)	(0.707)	(0.517)	(0.647)
gt_covid	0.254	-6.882***	-9.598***	-9.514***	5.644***	3.238***	0.504	-1.026
	(1.946)	(2.122)	(2.235)	(2.042)	(1.419)	(1.037)	(0.781)	(0.849)
gt_prevention	2.953	2.939***	-1.566	-2.872	0.438	1.353	0.883	0.560
	(1.841)	(1.044)	(1.389)	(1.895)	(0.721)	(0.951)	(0.585)	(0.485)
gt_fakenews	-3.607**	-5.342***	-3.078**	-1.386	0.899*	0.278	-0.760	-2.146***
	(1.600)	(1.398)	(1.214)	(1.162)	(0.469)	(0.769)	(0.790)	(0.546)
gt_vaccines	-0.0733	3.340***	3.751***	2.799***	-2.207***	-1.272**	0.250	0.293
	(1.169)	(1.172)	(1.046)	(0.830)	(0.620)	(0.581)	(0.403)	(0.389)
trend	-0.261***	-0.269***	-0.306***	-0.235***	-0.0723*	-0.0811***	-0.172***	-0.241***
	(0.0596)	(0.0642)	(0.0723)	(0.0659)	(0.0402)	(0.0293)	(0.0460)	(0.0658)
_cons	19.38***	24.24***	26.71***	23.65***	4.458	6.991***	12.62***	19.05***
	(4.550)	(5.366)	(5.613)	(4.964)	(2.825)	(2.239)	(3.605)	(5.209)
Within R^2	0.071	0.075	0.075	0.063	0.033	0.038	0.051	0.061
Between R^2	0.354	0.467	0.438	0.366	0.001	0.128	0.115	0.161
Overall R^2	0.053	0.059	0.063	0.059	0.029	0.028	0.036	0.045
Observations	67729	67069	66969	66939	62149	62072	61868	61767
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the second sub-sample (2021) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by notification area) and residential and workplace mobility together with SRAG vaccination, Google Trends and News as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

Table 1.11: Estimation for Complete Sample (2020-2021): All Mobility Variables

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-129.1***	-109.5***	-78.20***	-52.85***	-11.66	-41.65***	-58.29***	-55.95***
	(24.78)	(21.88)	(17.30)	(16.48)	(10.20)	(13.00)	(13.04)	(10.77)
workplace	17.41***	17.00***	14.16***	11.28***	4.645***	6.777***	7.765***	7.979***
	(4.554)	(4.525)	(3.330)	(2.580)	(1.452)	(2.024)	(1.925)	(1.893)
transit	0.436	0.142	0.285	0.305	0.230	0.268*	0.229	0.221
	(0.423)	(0.417)	(0.414)	(0.375)	(0.157)	(0.162)	(0.203)	(0.201)
retail	8.809***	3.901*	-2.135	-5.268*	1.621	2.588***	2.375**	1.222
	(2.275)	(2.146)	(2.464)	(2.701)	(1.064)	(0.950)	(1.000)	(1.023)
grocery	-5.223***	-6.109***	-3.979***	-2.126**	0.552	-0.633	-1.860***	-2.120***
	(1.589)	(1.290)	(1.009)	(0.971)	(0.516)	(0.569)	(0.645)	(0.593)
parks	-1.354***	-0.911	0.273	0.543	-0.595**	-0.525**	-0.381*	-0.203
	(0.517)	(0.593)	(0.647)	(0.732)	(0.296)	(0.237)	(0.197)	(0.229)
srag_vac	-0.288***	-0.300***	-0.303***	-0.261***	-0.0259	-0.0444**	-0.0678***	-0.0802***
	(0.0874)	(0.0920)	(0.0932)	(0.0766)	(0.0161)	(0.0182)	(0.0214)	(0.0246)
n_covid	1.978	-3.360	-4.153	-4.845*	4.478**	1.707	0.841	-1.194
	(2.634)	(2.434)	(2.708)	(2.687)	(1.808)	(1.290)	(0.911)	(1.185)
n_prevention	-0.326	-1.378	-3.919*	-3.001**	-0.246	-0.622	-1.641	-1.848*
	(2.021)	(2.011)	(2.097)	(1.370)	(1.035)	(0.634)	(0.999)	(1.076)
n_fakenews	-2.001	0.259	5.374	6.425	-1.278	-0.701	-1.074	1.107
	(2.804)	(4.479)	(6.547)	(7.750)	(0.939)	(0.965)	(1.711)	(2.862)
n_vaccines	0.614	1.068	4.846*	4.855**	-3.990**	-1.467	-1.297	1.255
	(2.927)	(3.861)	(2.729)	(2.265)	(1.613)	(1.396)	(1.032)	(1.172)
gt_covid	-2.867	-15.29***	-19.05***	-19.44***	9.829***	4.709**	-0.452	-3.052*
	(4.101)	(4.706)	(4.663)	(3.943)	(2.823)	(2.061)	(1.559)	(1.700)
gt_prevention	7.580**	6.733***	-0.905	-3.914	0.873	2.645	2.512**	1.690*
	(3.470)	(2.227)	(2.540)	(3.470)	(1.494)	(1.979)	(1.216)	(0.947)
gt_fakenews	-4.920	-8.942***	-5.900***	-1.848	1.196	0.751	-1.205	-3.605***
	(3.075)	(2.736)	(2.165)	(2.236)	(0.873)	(1.630)	(1.625)	(1.051)
gt_vaccines	0.504	6.860***	6.387***	5.842***	-3.654***	-1.975	0.647	0.679
	(2.489)	(2.448)	(2.184)	(1.640)	(1.356)	(1.233)	(0.813)	(0.750)
trend	-0.450***	-0.479***	-0.476***	-0.377***	-0.0861**	-0.151***	-0.249***	-0.305***
	(0.0990)	(0.108)	(0.103)	(0.0870)	(0.0409)	(0.0435)	(0.0619)	(0.0755)
_cons	38.83***	43.40***	42.05***	34.61***	5.105*	11.89***	20.04***	26.74***
	(8.266)	(9.330)	(9.008)	(7.824)	(3.016)	(3.466)	(5.296)	(6.624)
Within R^2	0.090	0.095	0.094	0.083	0.047	0.052	0.066	0.077
Between R^2	0.239	0.333	0.472	0.501	0.007	0.028	0.069	0.107
Overall R^2	0.067	0.076	0.083	0.081	0.042	0.038	0.046	0.058
Observations	32847	32584	32574	32577	30936	30992	30939	30918
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.271	0.998	0.000	0.000	0.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the second sub-sample (2021) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by residence area) and all six mobility measures (residential, workplace, transit, retail, grocery and parks) together with SRAG vaccination, Google Trends and News as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

Table 1.12: Estimation for Complete Sample (2020-2021): National Vaccination Data

	(1) m=1	(2) m=2	(3) m=3	(4) m=4	(5) m=1	(6) m=2	(7) m=3	(8) m=4
residential	-68.28***	-62.24***	-46.86***	-35.45***	-4.760	-19.98***	-28.63***	-28.33***
	(10.24)	(9.790)	(7.782)	(7.663)	(3.285)	(4.730)	(5.284)	(4.826)
workplace	9.693***	7.198***	5.382***	3.240***	2.622***	3.881***	4.343***	3.957***
	(1.728)	(1.549)	(1.149)	(0.921)	(0.618)	(0.813)	(0.782)	(0.707)
1st_dose	-0.000304***	-0.000273***	-0.000226***	-0.000172***	-0.0000776***	-0.0000856***	-0.0000953***	-0.0000799***
	(0.0000438)	(0.0000343)	(0.0000511)	(0.0000636)	(0.00000627)	(0.00000935)	(0.0000124)	(0.0000104)
2nd_dose	0.0000556**	0.0000810***	-0.000000613	-0.0000306	-0.0000171**	0.0000100*	0.0000197***	0.0000170**
	(0.0000240)	(0.0000265)	(0.0000442)	(0.0000525)	(0.00000718)	(0.00000581)	(0.00000692)	(0.00000788)
3rd_dose	-0.000357***	-0.000350***	-0.000129	0.0000222	-0.000168***	-0.000160***	-0.000156***	-0.000112***
	(0.0000470)	(0.0000756)	(0.000122)	(0.000131)	(0.0000461)	(0.0000264)	(0.0000306)	(0.0000167)
n_covid	0.400	-2.336**	-2.294*	-2.234*	1.874***	0.471	0.133	-0.855*
	(1.020)	(1.091)	(1.204)	(1.182)	(0.719)	(0.513)	(0.363)	(0.508)
n_prevention	-0.00250	-0.863	-1.986*	-1.568**	-0.179	-0.413	-0.788	-0.894*
	(0.973)	(0.979)	(1.039)	(0.677)	(0.479)	(0.319)	(0.493)	(0.529)
n_fakenews	-1.562	-0.338	2.719	3.415	-0.709	-0.471	-0.697	0.305
	(1.531)	(2.253)	(3.272)	(3.847)	(0.554)	(0.551)	(0.979)	(1.594)
n_vaccines	1.483	1.897	2.798***	2.256**	-1.578***	-0.405	-0.0767	1.188**
	(1.075)	(1.450)	(1.066)	(0.905)	(0.592)	(0.514)	(0.392)	(0.466)
gt_covid	-2.684	-8.386***	-10.35***	-9.562***	4.219***	2.082**	-0.363	-1.738**
	(1.666)	(1.983)	(1.948)	(1.642)	(1.213)	(0.854)	(0.644)	(0.713)
gt_prevention	3.973***	3.752***	-0.457	-1.796	0.581	1.438*	1.227**	0.884**
	(1.467)	(0.929)	(1.039)	(1.386)	(0.599)	(0.781)	(0.490)	(0.406)
gt_fakenews	-2.379*	-4.362***	-2.521***	-1.254	0.959***	0.378	-0.485	-1.738***
	(1.253)	(1.118)	(0.877)	(0.927)	(0.371)	(0.647)	(0.650)	(0.435)
gt_vaccines	1.423	4.340***	4.264***	3.586***	-1.815***	-0.865*	0.628*	0.719**
	(1.157)	(1.213)	(1.060)	(0.786)	(0.551)	(0.468)	(0.357)	(0.350)
trend	-0.228***	-0.270***	-0.294***	-0.240***	-0.0499*	-0.0780***	-0.142***	-0.187***
	(0.0442)	(0.0526)	(0.0542)	(0.0475)	(0.0303)	(0.0227)	(0.0340)	(0.0483)
_cons	18.95***	23.04***	24.62***	21.27***	4.082	6.229***	11.11***	16.66***
	(3.529)	(4.373)	(4.559)	(4.099)	(2.514)	(1.757)	(2.833)	(4.061)
Within R^2	0.049	0.047	0.044	0.038	0.039	0.037	0.043	0.042
Between R^2	0.217	0.202	0.284	0.222	0.169	0.175	0.089	0.122
Overall R^2	0.037	0.039	0.040	0.039	0.031	0.027	0.030	0.032
Observations	82460	81651	81647	81649	77022	76615	76601	76604
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Hausman	0.000	0.000	0.000	0.180	0.000	0.000	0.000	0.000

Standard errors in parentheses

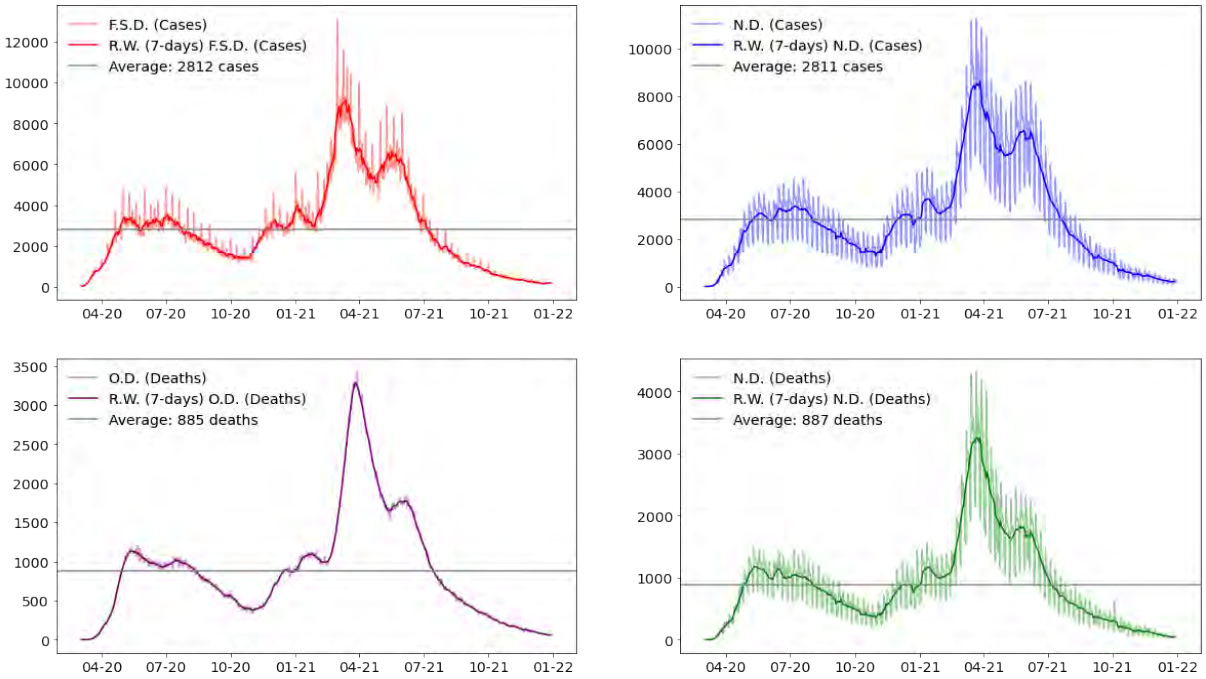
* $p < .1$, ** $p < .05$, *** $p < .01$

[a] Results for Fixed-Effects Estimation for the second sub-sample (2021) using Covid-19 cases (columns (1) to (4)) and deaths (columns (5) to (8)) as dependent variable (aggregated by notification area) and residential and workplace mobility together with national vaccination data, Google Trends and News as regressors. The time-structure $m \in \{1, 2, 3, 4\}$ refers to the week-by-week evolution of the dependent variable.

[b] p-value is the associated probability of the overall statistic significance of coefficients from the F-Test. Hausman is the p-value associated with Hausman Test for fixed versus random effects estimation.

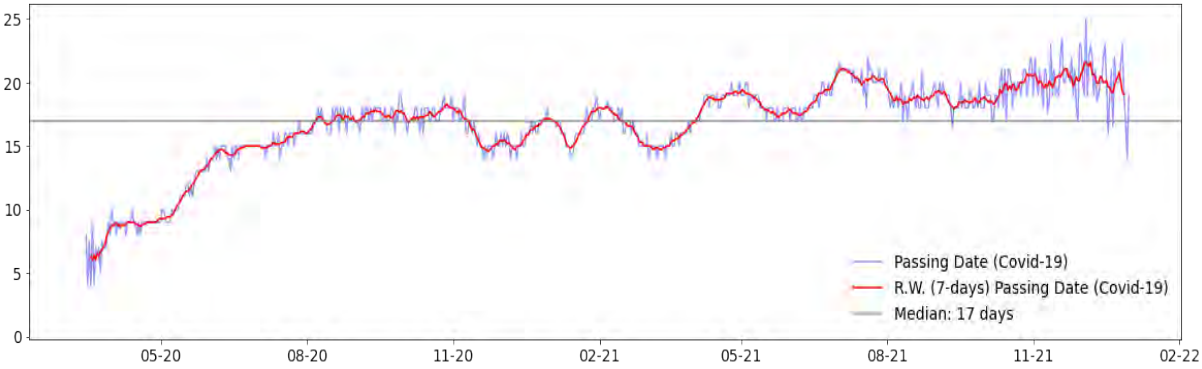
Figures

Figure 1.1: First Symptom and Obit Date Versus Notification Date



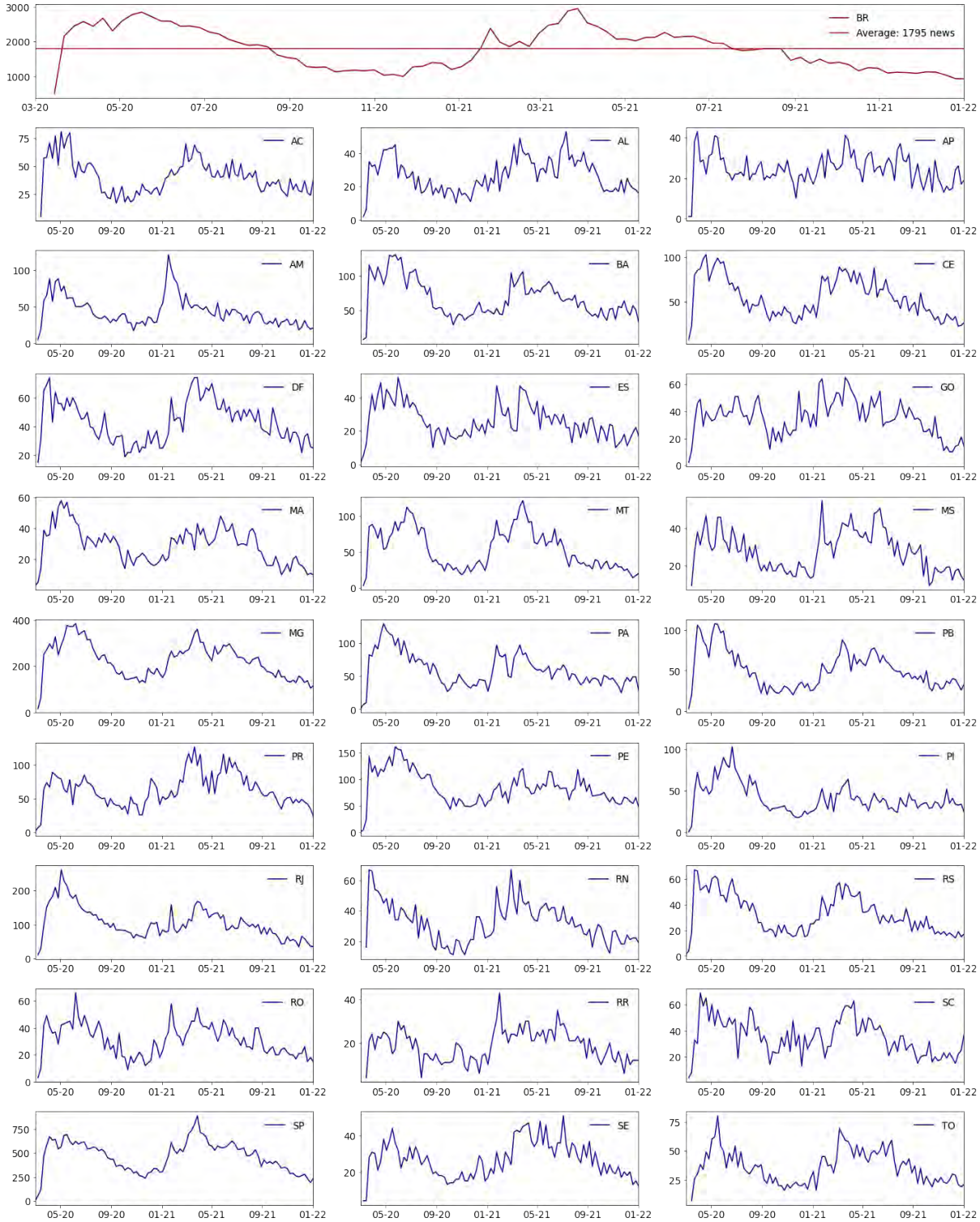
Comparison between First Symptom Date (FSD) for cases or Obit Date (OD) for deaths and the Notification Date (ND) daily from March 2020 to December 2021. Note how the notification date is much more volatile than the FSD series for cases or the OD series for deaths. The average number of cases and deaths also changes depending on the date aggregation.

Figure 1.2: Median Number of Days from First Symptom to Obit



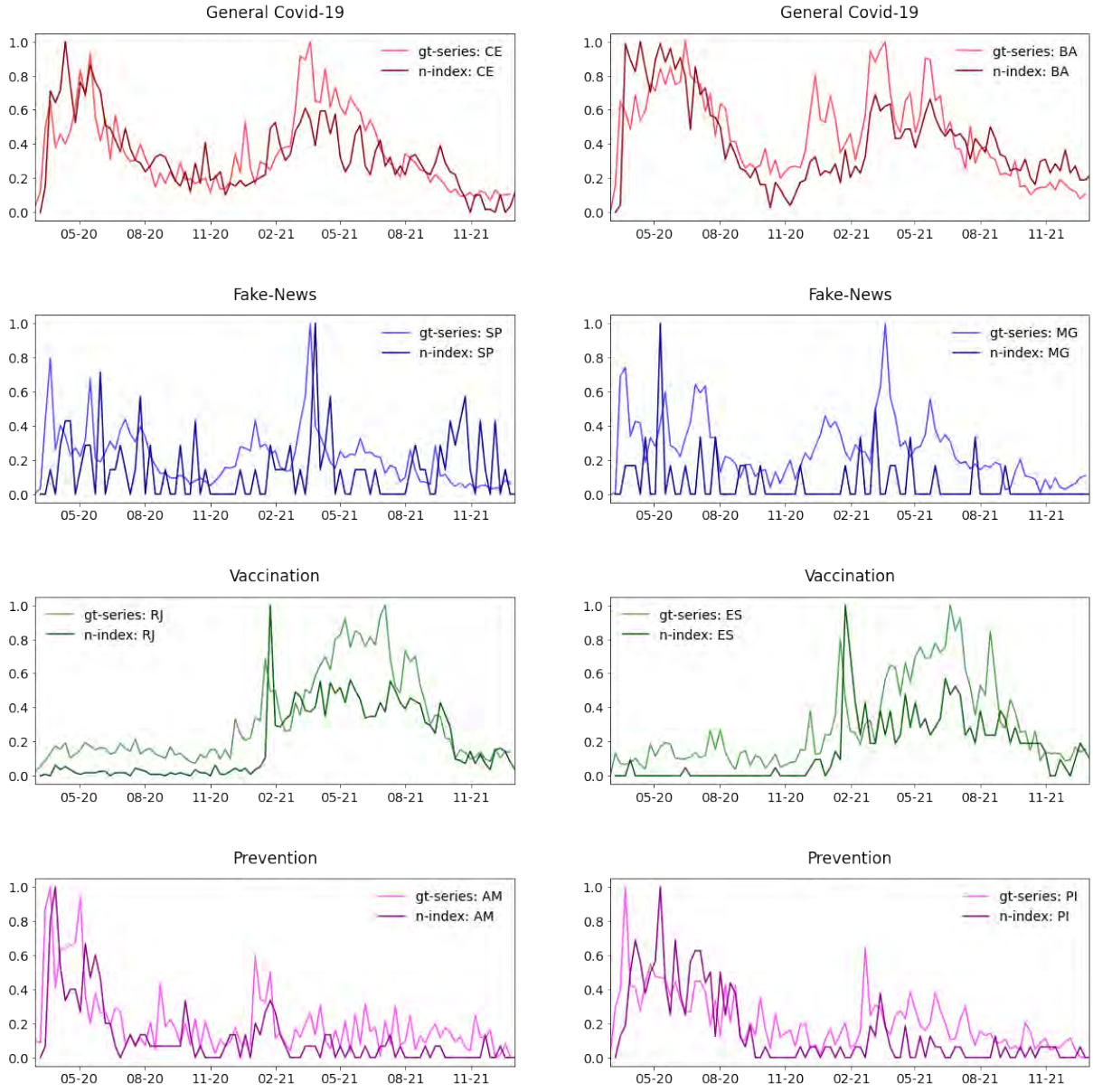
The plot reveals the median number of days taken from first symptom up to obit from patients that are considered on the SRAG data over the period of March 2020 up to December 2021. We also represent a 7-days rolling window (R.W.) as a smoothing and the overall median (17 days).

Figure 1.3: News Regarding Covid-19



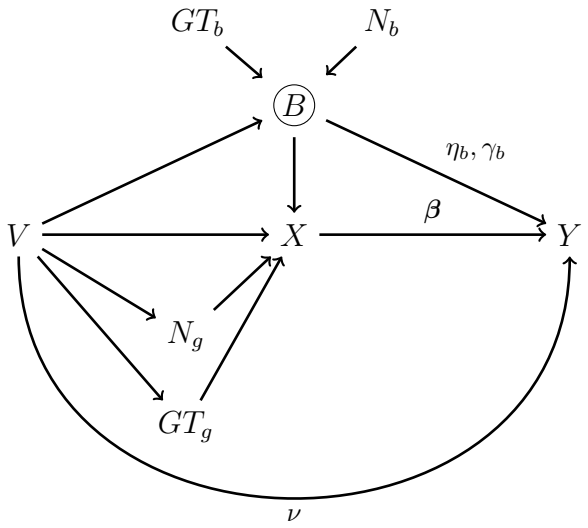
The plot reveals the evolution of the number of news from the G1 newspaper at country and state level for the period of March 2020 up to December 2021. The average weekly number of news regarding Covid-19 for Brazil was 1795 on the period.

Figure 1.4: Google Trends Series and News-Index



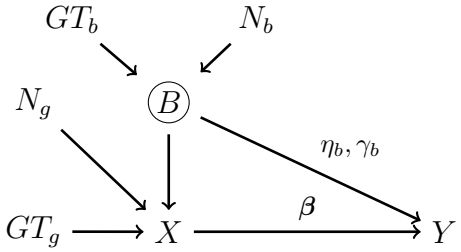
The plot reveals the evolution of the Google Trends Series (GT-series) and the News-Index (N-index) over the four categories used as controls for selected states for the period of March 2020 up to December 2021.

Figure 1.5: DAG for 2021 Identification Representing Causal Chain Within Model's Variables



The complete version of the DAG suggests how mobility (X) interacts with vaccination (V), Google Trends (GT_b), News (N_b), proxies for the behavioral effects (B). Each arrow suggests a connection between two variables. Following Elwert (2013), only missing arrows makes assumptions regarding causal relations. The circle around the behavioral variable (B) denotes a non-observable variable. Also, general Covid-19 situation variables as GT_g and N_g serve as proxies for lagged effects of Y and may be related to vaccination (V) and mobility (X).

Figure 1.6: DAG for 2020: Shutting Down Vaccination Channel



The shrunk version of the DAG (only for 2020) shuts down the vaccination channel that affects mobility, behavior, and cases (deaths). Such transformation would imply a cleaner identification of mobility effect, but with local validation only, not extending to the subsequent second wave of Covid-19 spread.

2
Nowcasting GDP with Unstructured Data

2.1
Introduction

Assessing the real-time situation of the economy is a topic of particular concern for applied economists. In particular, central banks and financial institutions dedicate considerable effort to predict the evolution of macroeconomic variables such as Gross Domestic Product (GDP) and Inflation. While central banks focus on amplifying the information set to make monetary policy decisions, financial institutions anticipate market behavior to make profitable transactions.

In order to map the real-time evolution of macroeconomic variables, the nowcasting literature naturally surges to "predict the present, the very near future and the very recent past", as initially suggested by Evans (2005), and Giannone et al. (2008). Indeed, many of the statistical foundations for nowcasting previously developed focused mainly on "hard" data indicators, consisting of usual macroeconomic indicators as GDP components, activity measures (e.g., industrial production, retail sales), inflation, and other indirect quantities.

As the idea of nowcasting is to generate a series of predictions, the problem is naturally high-dimensional for two different reasons: (i) the forecasting problem is seeded with a vast number of predictors in order to capture economic behavior;[1] (ii) the nowcasting problem naturally embeds a mixed frequency problem: while the target variable is usually low-frequency, researchers are interested in high-frequency measures of such series.[2]

In order to address the two problems simultaneously, Giannone et al. (2008) adopted the usage of (Dynamic) Factor Models and Kalman smoother to capture components that affect the current (non-observable) state of the target variable. Such an approach can handle the growing availability of economic

[1]For additional reference in big data and alternative sources for nowcasting, refer to Bok et al. (2018) survey, which analyzes different predictors and econometric solutions to work in a data-rich environment.

[2]For additional reference in mixed-frequencies models, refer to Foroni & Marcellino (2013) survey, which provides an in-deep review of usual approaches to deal with different frequencies.

indicators and access their marginal impacts over the "term-structure" of nowcasting.[3] This is also the case of later nowcasting models developed in Bańbura et al. (2010), Bańbura & Rünstler (2011) and Jansen et al. (2016).

On the other hand, the usage of Mixed-Data Sampling models also gained space in the nowcasting literature due to its flexibility while matching series frequencies. Those models allow for the inclusion of variables of different frequencies while transforming the higher frequency variable into a lower frequency series of variables.[4] However, two main strands differ on the parametrization of the model. While the "unrestricted" (U-MIDAS) front allows each variable to have its coefficient, the "restricted" (MIDAS) imposes a polynomial structure over the coefficients to limit the number of parameters to be estimated. The difference between these two approaches is analyzed in Foroni & Marcellino (2014), and the gains of each strand depend mainly on the econometric problem.

While Dynamic Factor Models have dominated the core literature of nowcasting, the usage of U-MIDAS combined with machine learning models that treat the inherent high-dimensional problem with dimensionality reduction solutions (from shrinkage models to Neural Networks) is gaining space due to its higher flexibility and lower computational effort.[5] This is the case of Borup et al. (2021), which uses a U-MIDAS model together with machine learning techniques as Least Absolute Shrinkage and Selection Operator (LASSO), Random Forests (RF), and Neural Networks (NN), in order to deal with the mixed-frequency problem (daily predictors, weekly target) and high-dimension setup based on alternative data usage (Google Trends) to nowcast unemployment.

After briefly describing the mixed frequencies problem, we now turn our attention to the data problem. Most of the current literature's theoretical and empirical works focused on "structured" data, usually categorized as quantitative or highly organized data formats. In this category, we highlight the continuous effort to create common data sources that can supply informative variables into big-chunk models. This includes the efforts made by McCracken & Ng (2016, 2020) to consolidate a series of real-time indicators ("Big Data")

[3]As the informational flow is non-decreasing, one can generate a term-structure of projections indexed by the current state of the economy based on the flow of information (e.g., economic indicators) that becomes available during the sampling-period of the target variable.

[4]This transformation is done by skip-sampling the variable. For example, suppose that the target variable is sampled at a quarterly frequency. Therefore, a monthly regressor is transformed into three quarterly regressors, where each series denotes a month of the quarter in analysis.

[5]As an example, in Ellingsen et al. (2020), the authors needs to limit the number of textual elements in order to estimate a DFM model.

for the United States economy.

However, new efforts to consider alternative data sources have been made over the last few years. For example, one of the first papers in order to consider "soft" data variables such as market expectations, financial variables, and surveys came from Rünstler et al. (2007), which explored the additional information provided by such alternative data sources in order to forecast GDP in the Euro area. More recently, a new form of "unstructured" data, categorized as qualitative data, displayed in non-usual pre-processed formats, has been used for nowcasting purposes.[6] Internet searches and textual data are two examples of unstructured data recently explored in nowcasting applications.[7]

In terms of internet searches, the usage of Google Trends data in Choi & Varian (2009, 2012) to forecast economic activity (firstly unemployment benefits, then automotive industry evolution, respectively), kick-started a new branch of unstructured data analysis that revealed to be important predictors in nowcasting problems (see Carrière-Swallow & Labbé (2013), Smith (2016), and more recently Woloszko (2020) and Borup et al. (2021)). The main issue regarding Google Trends data is that the process of selecting terms is not innocuous: one needs to manually select terms and actively provide its selection to the model to generate predictions.[8]

Also, textual data, mainly from newspaper articles, has been an important source of alternative information in econometric problems. The precursor paper of Baker et al. (2016) using counts of news in order to develop an uncertainty index motivated a deeper analysis of newspaper data in economics. Also, Manela & Moreira (2017) shows that news embeds valuable information in order to predict implied volatility from VIX. Nonetheless, a large literature regarding central bank communication (e.g., Hansen & McMahon (2016) and Hansen et al. (2018)) also resorts to textual data to extract its underlying information and evaluate its impacts through the economy. In terms of nowcasting research, newspaper articles displayed important relevance in order to predict macroeconomic variables and sparked a new branch in the nowcasting literature (see Thorsrud (2016), Ke et al. (2019), Bybee et al. (2020) and more recently Saiz et al. (2021)).

[6]For additional discussion and reference, visit: `https://www.ibm.com/cloud/blog/structured-vs-unstructured-data`.

[7]Note that unstructured data is composed basically by any format of data that cannot be processed by recurring to conventional tools. Other examples of unstructured data are audio, images, video, and maps.

[8]In stressful environments such as the Covid-19 crisis, new information has been added to the model's information set, and without a cautious addition of such terms, one may lose important information or even commit look-ahead bias over the estimations. Adding an important term ad-hoc generates an implicit hypothesis that this term was correctly inserted into the model at the correct period to capture the predictive gains over the period.

Compared to internet searches, newspaper data provides a more data-driven approach due to the common textual-to-data approaches. However, while internet searches data from Google Trends is already provided in a structured format, news articles need to be pre-processed and structured in a useful format. As a result, the natural language processing (NLP) literature was born to provide a series of transformations from raw data into structured formats, which allow the usage of text in econometric problems.[9] Text elements go through a whole process of cleaning up to an organization that is sensible to the hypothesis made in terms of text structure. As an example, while Baker et al. (2016) simply count the number of news, Manela & Moreira (2017) focus on terms-counting, Bybee et al. (2020) focus on media-attention series and Kelly et al. (2019) focus on the inclusion (or exclusion) rather than on a count of a given term. Each approach yields a different length in terms of dimensionality reduction and different results in terms of forecasting capabilities.

With the objective of nowcasting the current state of the economy (i.e., the Brazilian GDP), this paper recurs to a combination of three different elements: (i) addition of alternative data sources (newspaper and Google Trends) with conventional data (as macroeconomic indicators, market expectations, temperature, the balance of trade data and energy data); (ii) usage of a mixed-data sampling framework (U-MIDAS); (iii) adoption of machine learning models for high-dimensional settings (shrinkage models as LASSO and AdaLASSO). Indeed, we show that such a combination of data sources is a powerful tool as it generates higher predictive power when compared to models that only recur to usual macroeconomic indicators.

This paper aims to propose the usage of alternative data sources together with conventional data to check if the former has information not encoded in the latter. We define alternative data as internet searches collected via Google Trends through a set of pre-defined terms. We also collect and use textual data from the three prominent publicly available newspapers in Brazil (Folha de São Paulo, Estadão, and Valor Econômico). Those articles have been cleaned and transformed into structured data sets, where each comprehends counts of terms or phrases (N-grams) and a general media-attention series.[10] Each of these three data sources (Google Trends, N-grams, and media attention) is included one at a time and all together to evaluate the additional information provided by their inclusion.

[9]For additional reference in textual data and application in economic problems, refer to Gentzkow et al. (2019) survey on textual analysis.

[10]Following Bybee et al. (2020), based on Latent Dirichlet Allocation (LDA) modelling. For additional reference for LDA, we refer to Blei et al. (2003).

The modeling approach is similar to Borup et al. (2021), as we adopt a U-MIDAS setup in order to make compatible weekly, monthly, and quarterly frequencies without imposing restrictions over the coefficients of each variable. To address the high-dimensionality of the problem, we also adopt linear shrinkage models that can induce sparsity (LASSO, as suggested by Tibshirani (1996)) and variable selection (AdaLASSO, as suggested by Zou (2006)). The U-MIDAS approach allows the construction of a term structure of nowcasts, while linear shrinkage models allow simple marginal impact evaluation of each predictor, allowing for the evaluation of the impacts of each different data source.

We set a benchmark model to evaluate the nowcasting model the market expectations collected from "Relatório Focus", from the Central Bank of Brazil. Preliminary results based on simple linear specifications suggest that pure shrinkage models cannot beat the benchmark, on average. However, we show that generating ensembles (combining market expectations with shrinkage models that use alternative data sources) is a powerful approach. With such models, we can over-perform the benchmark and systematically enhance forecasting quality, measured by a significant decrease in the mean squared (MSE) and mean absolute errors (MAE). Furthermore, all nowcasts have been compared and tested with the benchmark using a Diebold Mariano's 2002 test, ensuring the statistical validation of the results.

Nonetheless, we also show that textual data is more selected by the shrinkage models whenever usual indicators are not available. In the first weeks after the last GDP number has been released, alternative data sources generate a significant broader gain in reducing the MSE and MAE. We argue that the success of such predictors comes from the fact that only a few usual macroeconomic indicators are available, and Google Trends and news articles can provide additional valuable information for nowcasting that is not encoded on such indicators. However, as time evolves, in the last nowcast prediction prior to the actual release of the GDP, alternative data sources seem not to provide additional information, and the nowcasts cannot over-perform the benchmark.

We also discuss further details of the nowcasting model, evaluating the variables selected over the shrinkage specifications. First, we present one example pointing to the higher importance of alternative data sources when usual indicators are not available. Next, we explain the usage of both linear and quantile regression to attenuate the nowcasting term-structure evolution, arguing that modeling the median instead of the average provides a less volatile nowcast path. Finally, we argue that an ensemble model that considers all four

types of shrinkage specification (LASSO and AdaLASSO; median or average regression) provides a best-performer for this data set, as it can reduce the noise of each nowcast.

This paper is structured into five sections. Section 2.2 starts by describing the textual data collected online from each newspaper, all pre-processing and cleaning, and the structure of the textual data set. Then, we describe internet searches data and the "enchainment" process adopted to mimic a weekly series based on the longer Google Trends monthly observations. Next, we describe all hard data sources, including usual macroeconomic indicators and alternative structured sources such as market expectations, temperature, the balance of trade, and energy data. Section 2.3 presents the methodology used to develop the U-MIDAS nowcasting model, as also the specifications and the estimation framework along with the information flow of the model. Section 2.4 displays the out-of-sample results, the particularities of the term structure of nowcasts, and also further details and discussions. Section 2.5 concludes this paper.

2.2 Data

This section describes all data series used in the nowcasting problem. We cover the period of January 1, 2009, up to December 31, 2021, i.e., a whole thirteen-year period, or 52-quarters. The data sources may differ in terms of their frequency, as described in this section. Note, however, that all series have been collected using first-vintages, and we only consider real-time data, i.e., we do not include data that was not available when the forecast was made.

A real-time nowcasting problem induces the necessity to reconcile different date types. In particular, the nowcasting with unstructured data induces three different dates: (i) reference date, describing the period that the series is referring to (e.g., Brazilian GDP accounting for 2020Q1); (ii) release date, describing when the data source becomes firstly available (e.g., GDP regarding 2020Q1 was only released at May 29, 2020); (iii) forecast date, describing when the nowcast prediction was generated (e.g., the first nowcast was calculated on March 13, 2020, while the last nowcast prediction was made on May 28, 2020). Note that the reference date and release date coincide for financial and textual series, as there is no lag between the release and the period covered by the data. In addition, these three different date elements generate a multi-indexed object describing the state of the economy at each point in time.

2.2.1 Textual Data

At this point, we describe the news-based textual data used on the nowcasting model. We summarize the process of finding what matters inside the unstructured data, up to generating highly organized structured data, which can be included as a usual time-series in the specifications. We focused solely on publicly digital news as it can be easily accessed does not require advanced techniques for text recognition. As guidance for textual analysis, we followed mainly Eisenstein (2018), and Kelly et al. (2019) as back-references.

Initially, we need to select the digital sources from which we collect the articles. The main feature that needs to be considered is the relevance and dominance of the newspaper in terms of news diffusion. As a proxy for relevance and diffusion, we focused on each newspaper's number of digital subscriptions. Following the yearly report of the Instituto Verificador de Comunicação (IVC) and Poder 360, the six main newspapers in terms of digital subscriptions are: (i) Folha de São Paulo; (ii) O Globo; (iii) Estadão; (iv) Valor Econômico; (v) Zero Hora; and (vi) Super Notícia. Following the same study, only Estadão, Folha, and Valor represent about 54% of the total number of digital subscriptions for the first quarter of 2021.[11] Therefore, we opt to collect publicly available news from Estadão, Folha, and Valor Econômico.[12]

We developed a framework to use the textual elements of each article of the newspapers, consisting of a bottom-up approach. The first step was to collect and generate a time-indexed object with all publicly available articles from each of these three leading websites. The second step consists of tokenizing, cleaning, and pre-processing text elements. The third and fourth steps consisted of structuring the text data in different manners. The first strategy used Principal Component Analysis over simple N-grams counting. The second strategy adopted an unsupervised learning model (based on a Latent Dirichlet Allocation model of Blei et al. (2003)) to generate media-attention series (similar to Bybee et al. (2020)). In both cases, we used PCA or LDA groups to generate a treatable structured object that can be used as regressors, avoiding introducing too much information prone to overfitting or even noise in the estimation. All articles have been treated as arriving from different sources to control specific vocabulary and only aggregated at a final step in the structured-series construction.

[11]IVC is a non-profit national entity responsible for audit media multiplatform. For additional information, refer to the IVC website: `https://ivcbrasil.org.br/#/home`.

[12]The development of the algorithm and the usage of all news collected are in agreement with the terms and conditions of use and with the robots.txt file of each of the newspaper's website.

2.2.1.1 News Collection

The first step involves collecting articles' vintages based on the website search mechanism and a simple web-scraping algorithm. The algorithm filtered news from January 1, 2009, to December 31, 2021, covering politics, economy, and international affairs sections.[13] In addition, all article's dates considered when the article was first published, despite later revisions.

Table 2.1 presents a year-by-year summary of news collected for each newspaper and in general. Also, Figure 2.1 presents the graphical evolution of the total counts of news and (cleaned) words, as we describe over the next section. The evolution of words and news is closely related, a typical pattern in newspaper data, e.g., Bybee et al. (2020). Finally, Figure 2.1 highlights an important point: Valor Econômico only has an archive of news from 2012 onwards. However, we opt to start the model's estimation from 2009 onwards, as the overall news (the relevant measure) is well represented by Estadão (and Folha for 2010 onwards) articles. Moreover, as we deal with quarterly data, this inclusion of 2009 provides additional relevant data points, covering the financial crisis of 2008, essential to estimate and validate the in-sample model.

In Table 2.2 we present the descriptive statistics for the number of articles weekly, which is the nowcasting computation frequency. We highlight that Folha has a lower average number of articles, while Estadão and Valor have a similar average number of articles. Finally, as we aggregate news from all sources (forming a single "national" newspaper), the general evolution is much less volatile and a better representation of the overall media, as presented in Figure 2.1.

2.2.1.2 Tokenization and Cleaning

Collecting all textual elements from each article over the three prominent newspapers constitutes an essential part of the data structuring problem. However, as Gentzkow et al. (2019) describes, working with raw text elements generates an inherent high-dimensional problem. Therefore, to organize the data in a treatable format, we need to proceed with some transformations and simplifications over the raw text data to generate a treatable lower dimensional object.

[13]In particular, sections adopted in each newspaper were: (i) Folha de São Paulo: "Política, Mercado e Mundo"; (ii) Estadão: "Economia, Política e Internacional"; (iii) Valor Econômico: we collected all public news (there is no section filter on the website search mechanism).

At this level, we are going to introduce some mathematical notation. Denote the Corpus $\mathbb{C}_i$ as the set of articles from a particular newspaper $i \in \{\text{Folha}, \text{Estadão}, \text{Valor}\}$. As we deal with a time-series of articles, note that each element of the Corpus is a time-indexed object, i.e. $c_{i,t,j} \in \mathbb{C}_i$, where j denotes a single article published at date t. While $\mathbb{C}_i \subset \mathbb{C}$ are sets, the element $c_{i,t,j}$ is a string, i.e. a non-numerical element. For illustration, consider the following example:

$$c_{Estadão,31/12/2020,1} = \text{‘GRAFTON, Wisconsin - Quinhentas doses da vacina contra a covid-19 [...]’}$$

The first step, namely tokenization, consists in taking the whole text (i.e., the element $c_{i,t,j}$ in a string format) and dividing it into small pieces (tokens) of mainly words and punctuation. To apply such a transformation, we detect empty spaces between elements on the text data and use them as start-and-finish markers. After tokenizing all articles, we take our element $c_{i,t,j}$ and apply a transformation to create a set of elements composed by words and punctuation, which we are calling from now on as “text element". It is important to know that we treat capitalized words differently from lower case words. Only in a final step we normalize all words to lower case.[14] After performing tokenization, we should have:

$$c_{Estadão,31/12/2020,1} = \{ \text{‘GRAFTON’, ‘,’ , ‘Wisconsin’, ‘-’ , ‘Quinhentas’, ‘doses’, , ‘da’, ‘vacina’, ‘contra’, ‘a’, ‘covid-19’, ...}\}$$

The second step (which is a hypothesis, additionally) consists in extracting stopwords[15], rare words, digits, ordinal numbers, punctuation, and other non-alphanumerical elements. The implicit hypothesis is that such elements do not provide valuable information on the article.[16] Therefore, we remove all

[14]We have adopted such a rule to avoid ruling out proper names from our database. For example, “São” and “são” may refer to different words, as the first may be a city name, “São Paulo”, while the second is a simple connective.

[15]Stopwords have been removed based on a dictionary method and with a Python module named “Natural Language Toolkit (NLTK)”. For additional reference, refer to Bird et al. (2009).

[16]The idea is that connectors only provide easier oral interpretation (marking pauses, adding explanations or examples), while digits cannot be interpreted without context. Rare words may be too specific to have a clear, direct interpretation. Finally, other non-alphanumerical elements may refer to emoticons, which may not be common newspapers articles.

those elements from the articles from the Corpus. At the above example, we are left with:

$$c_{Estadão,31/12/2020,1} = \{ \text{‘GRAFTON’, ‘Wisconsin’, ‘doses’,‘vacina’, ‘contra’, ‘covid-19’, ...}\}$$

A final step proceeded with lemmatization. This process takes words and reduces them to their canonical form. Lemmatization is an important step that needs to be addressed carefully to provide a significant dimension-reduction on the text/feature space to be used as nowcasting variables. [17] After performing the lemmatization, we normalize all words to lower case. Under our primary example, we should observe something like:

$$c_{Estadão,31/12/2020,1} = \{ \text{‘grafton’, ‘wisconsin’,‘dosar’, ‘vacinar’, ‘covid-19’, ...}\}$$

In Figure 2.2 we provide a reference for the cleaning and pre-processing for the complete news used as an example through this subsection. Although many transformations have been adopted to clean the textual elements, some steps may rely on a subjective hypothesis regarding the type of content being analyzed. This is a fundamental problem in textual analysis, as different text pre-processing may lead to different data sets.

2.2.1.3 Words Counting - N-grams and PCA

Tokenization and text pre-processing provide a cleaner text representation but are still in a non-structured format. A widespread form to treat text as data is to proceed with a simple count of occurrences of a given term for a specific period. One of such representations is the bag-of-words methodology. In this representation, the order of words is ignored altogether, and a (cleaned) phrase of N-length is classified as an N-gram.

Most applications of textual data in economics recurred to simple N-grams counting, as precursor papers of Hansen & McMahon (2016), Manela & Moreira (2017), and Thorsrud (2016). Such an approach provides a simple yet direct form of representing text as data, as larger “N” generates more complex textual structures of “-grams”. However, as Gentzkow et al. (2019)

[17] To proceed with lemmatization, we used the Python package Spacy, which is an advanced open-source Natural Language Processing software. There is a full implementation of lemmatization in Portuguese made by native Portuguese speakers under development. For more information, refer to Honnibal et al. (2020).

highlights, higher-order N-grams induce both higher computational cost and also heavily zero-inflated series of counts, as terms and sentences become much more specific.

A simple mathematical formulation can be described as: index each unique term over the documents $\mathbb{D}_i$ in the corpus $\mathbb{C}_i$ by, $v_i \in \{1, \cdots, V_i\}$ where V_i is the total number of unique terms. Then, for each document $d_i \in \{1, \cdots, D_i\}$, compute the count x_{d_i,v_i} of occurrences of the term v_i in the document d_i. Then, the $D_i \times V_i$ matrix $\boldsymbol{X}_{D_i,V_i}$ of all such counts is called the document-term matrix for newspaper $i \in \{\text{Folha}, \text{Estadão}, \text{Valor}\}$. Note that we consider documents d_i as the aggregate set of news for each day. Therefore, d_i works as a single representative news for each day.

In terms of the nowcasting application, we opt to compute 1-gram up to 3-grams daily for each newspaper accounted separately. After that, we exclude terms that occur less than 10% in the overall sample size. As the distribution of counts is highly right-skewed (see Ellingsen et al. (2020)), such a threshold is fundamental to exclude words that occur only in a few days and can be interpreted as rare words (or particular vocabulary). After generating a cleaned version of the document-term matrix for each newspaper, we then aggregate the counts as a single representative newspaper, represented by $\boldsymbol{X}_{D,V}$. The document-term matrix consists of daily-indexed terms with about 16,000 observations in the nowcasting problem.

However, some critical remarks follow from the preceding aggregation: (i) we opt first to clean all articles, count and then aggregate as a single newspaper to avoid penalizing common words that may appear in one newspaper and not in another (e.g., business-related terms); (ii) whenever a new term is introduced on the vocabulary (e.g., Covid-19), its counts prior to the inclusion is set to zero. The latter remark is addressed by the aggregation of counts, as a term is considered new only when the three other newspapers do not account for it. Therefore, such an approach generates a document-term matrix with more observations and less zero-inflated counts despite particular terms or new vocabulary.

Figure 2.3 display the distribution of unique terms, for each newspaper and in aggregate, over the total sample size (in days). We highlight that the distribution of unique terms is highly right-skewed: few terms dominate the vocabulary of articles, while many terms only appear a few times. Such empirical regularity has already been studied over the Natural Language literature in terms of Zipf's Law behavior (see Eisenstein (2018)). Also, in Figure 2.4 we display the evolution of selected N-grams to motivate the usage of such series in the nowcasting model. We also highlight that series behavior

spikes are accompanied by periods when the word received more attention in a given topic, e.g., the word "pandemia" had practically zero counts up to the Covid-19 pandemics in 2020.

As the total number of terms exceeds 16,000, this significant representation may not be feasible in terms of the nowcasting estimation described in Section 2.3. Therefore, to generate a lower-dimensional setting, we apply a Principal Components Analysis (PCA) over the counts of terms in the newspaper corpus. Such a model can reduce the total number of counts to only a fraction of the original data, which also helps to explain a significant fraction of the total variance over the data (e.g., see Drikvandi & Lawal (2020)).

2.2.1.4 Media Attention - LDA Groups

Word counting can be helpful to describe text structure in an organized format, but it may lose internal coherence between phrases that can enhance the informational component of text data. A simple form to deal with such a problem is to consider higher-order N-grams. However, higher-order N-grams introduce nonlinear computational costs and exponentially larger datasets. Another form to deal with the growing complexity of higher-order N-grams is to consider clustering algorithms to group words with similar content and meaning.

Clustering algorithms have been widely studied in the Natural Language literature (see Eisenstein (2018) and Gentzkow et al. (2019) for example), as it may require particular properties that occur in text data, such as mixed-membership and nonindependent distributions. As an example, an optimal cluster algorithm may not treat words as belonging to one exclusive group, as words can describe different topics depending on its context.[18] A very popular approach, widely used in economics papers as in Bybee et al. (2020), Hansen & McMahon (2016) and Thorsrud (2016), is to use Latent Dirichlet Allocation (LDA) model, developed by Blei et al. (2003) to generate non-exclusive mixed-membership groups of words.

The idea of the LDA model is simple: assign each (unique) word $v \in \{1, \cdots, V\}$ in each day t to a topic β_k. Note that the k-topics may not be exclusive, so a word v can simultaneously belong to more than one topic.[19] This model is based on a pure Bayesian computation:

[18] As an example, the word "crisis" may belong to an economic group describing "economic crisis" or a sanitary group, describing the "Covid-19 sanitary crisis".

[19] Note that the LDA model considers word counts for the aggregate document-term matrix instead of treating each newspaper as separate. The idea is to generate a representative media-attention series.

$$P(\boldsymbol{\beta}|\boldsymbol{x_t}) = \frac{P(\boldsymbol{x_t}, \boldsymbol{\beta})}{P(\boldsymbol{x_t})} = \frac{P(\boldsymbol{x_t}|\boldsymbol{\beta})P(\boldsymbol{\beta})}{P(\boldsymbol{x_t})}$$

where $\boldsymbol{x_t} = (x_{1,t}, \cdots, x_{V,t})$ is a vector of counts of words $v \in \{1, \cdots, V\}$. The idea is to compute the posterior distribution $P(\boldsymbol{\beta}|\boldsymbol{x_t})$ based on the likelihood function $P(\boldsymbol{x_t}|\boldsymbol{\beta})$, a prior distribution over the topics $P(\boldsymbol{\beta})$ and a normalizing constant $P(\boldsymbol{x_t})$.

The LDA resorts to the fact that the Dirichlet Distribution is a popular choice in Bayesian models due to its conjugate prior property.[20] Given the aggregated document-term matrix $\boldsymbol{X}_{D,V}$ with D documents and V unique terms, denote α as the parameter of a Dir(α) prior regarding document-topic distribution, and β the parameter of a Dir(β) prior regarding topic-word distribution. Also, define $x_{d,v}$ as a specific word and $z_{d,v}$ as the topic of the v-th word in a document d.

Intuitively, as Bybee et al. (2020) and Thorsrud (2016) suggests, the LDA algorithm produces two outputs: (i) a distribution of words for each topic, i.e., $\phi_k \sim \text{Dir}(\beta)$ for $k \in \{1, \cdots, K\}$; (ii) a distribution of topics for each document, i.e., $\theta_d \sim \text{Dir}(\alpha)$ for $d \in \{1, \cdots, D\}$. The LDA generative process used to infer the topics in the corpus consists in a three steps process: (i) Draw ϕ_k independently for $k = 1, \cdots, K$ from Dir(β); (ii) Draw θ_d independently for $d = 1, \cdots, D$ from Dir(α); (iii) For each word $x_{d,v}$ in document d, draw topic assignment $z_{d,v}$ from θ_d and word assignment to topic from $\phi_{z_{d,v}}$. We fix scalar values for α and β.

Note that the two inputs given to the model are the vector of word counts $x_{d,v}$ and the number of topics k. A popular approach while selecting the number of topics is to fix k and inspect the clusters generated manually. In this article, we use a similar approach as the one provided by Bybee et al. (2020) which consists of a data-driven method for selecting the optimal number of topics k: generate a grid of topics (in our case, starting from 10 growing up to 200) and calculate the likelihood score. Then, we select the number of topics that maximize the score. Instead of dividing the sample into sub-samples and proceeding with cross-validation or Bayesian factor analyses, we consider the number of topics relatively stable over time and already known. After calculating the number of topics, the number of media attention series are held fixed (i.e., the probability of each word of each article belonging to a specific topic), and only tokens counts affect the evolution of such series.

The results regarding the optimal number of topics are displayed in

[20] When the posterior distribution $P(\boldsymbol{\beta}|\boldsymbol{x_t})$ is in the same probability distribution family as the prior probability distribution $P(\boldsymbol{\beta})$, the prior and posterior are called conjugate distributions.

Figure 2.5. By both data-driven procedures and manual inspection, the optimal number of topics is around $k = 60$. More topics induce similar groups, while fewer topics lose relevant news subjects. After calculating the LDA probabilities of a particular article that belong to a given topic, one can generate media-attention series, as suggested by Bybee et al. (2020). The attention given to a specific topic k is defined as the scalar product between the counts of occurrences of word v given by $x_{d,v}$ and the probabilities of occurrence of this word inside a given topic $\phi_{z_{d,v}}$. Formally, we have:

$$ATT(d) = \sum_t ATT(d,k) = \sum_d \sum_v x_{d,v}\phi_{z_{d,v}}$$

therefore, for each topic, we can construct an aggregated media-attention time-series that embed textual information in a lower-dimensional scheme. In Figure 2.6 we plot the evolution of the media attention series for selected topics. Note that the media-attention series is normalized between 0 and 1, where a unity denotes the full media attention over the optimal number of topics.

2.2.2 Google Trends Data

Another important source of unstructured data used on the nowcasting problem consists in internet searches. We opt to collect data from Google Trends, as Google is the most used search engine currently.[21] Also, a vast literature has used Google Trends data as an unstructured data source in economic problems since Choi & Varian (2009).[22]

The usage of this data source demands an explicit set of keywords and terms that agents search on the internet. In Table 2.3 we display over six categories (economy, investment, energy, companies, risk, and politics), all 87 search terms that have been collected every week. Note that all terms have been selected following these criteria: (i) the terms should not be a low-interest term; (ii) the terms should be either observable throughout the sample or very important to describe a pattern (e.g., Covid-19); (iii) all terms have selected based on Google Trends recommendation algorithm to avoid non-representative searches. We highlight, however, that some terms may have ambiguous meanings, such as "ir", which can denote income tax (abbreviation for "imposto de renda") or the intransitive verb "go". We studied the series behavior to validate the search relation for the target search term in such cases.

[21] StatCounter and Statista reveal a more than 90% market share as of January 2022.

[22] See, among others, Choi & Varian (2012) for macroeconomic forecasting exercises, Smith (2016) for MIDAS approach in unemployment forecasting, and more recently Borup et al. (2021) for inflation nowcasting.

As Google only publishes weekly series for Google Trends (simply GT from now on) for periods lower than 5-years, we adopted an "enchainment" strategy to replicate the monthly series available for horizons longer than 5-years. The idea is a two-step weighting methodology. Figure 2.7 summarizes the idea of the algorithm used to replicate the GT-series at a weekly frequency based on the double-weighting methodology that we propose.

2.2.2.1 Google Trends Double Weighting

The first step consists in generating $T-1$ series for Google Trends with equal length, i.e., 4-years, in our case. Also, each GT_j series with $j \in \{1, \cdots, T-1\}$ have an overlap with respect to the previous period GT_{j-1} series, with $j \in \{2, \cdots, T-1\}$ i.e. a 2-years overlap, in our case. Therefore, we are left with $T-1$ series covering the weekly period of T years, with $T-2$ overlapping periods, as Figure 2.7 suggests. These overlapping periods are used in order to join subsequent series, calculating a factor of adjustment, i.e., given two vectors of weekly observations $\boldsymbol{GT_j}$ and $\boldsymbol{GT_{j-1}}$, for each week t inside the overlapping period j calculate:

$$w_{t,j} = \frac{GT_{t,j}}{GT_{t,j-1}}$$

where $\boldsymbol{w_j}$ denotes a vector of weights of weekly observations t inside a year j, i.e. $\boldsymbol{w_j} = (w_{t_0,j}, \cdots, w_{t_j,j})$. After calculating the series of weights $\boldsymbol{w_j}$, the adjusted Google Trends weekly-series ($\boldsymbol{GTW}$ from now on) will be simply given by:

$$\boldsymbol{GTW} = (\boldsymbol{GT_1 w_1}, \boldsymbol{GT_2 w_2}, \cdots, \boldsymbol{GT_{T-2} w_{T-2}}, \boldsymbol{GT_{T-1} w_{T-1}})$$

note that $\boldsymbol{w_1} = \boldsymbol{w_{T-1}} = (1, \cdots, 1)$ as there are no overlapping periods.

The second weighting procedure consists in using the observed Google Trends monthly-series ($\boldsymbol{GTM}^{\text{observerd}}$ from now on) as a final adjustment to the weekly series. The idea is to force the weekly series to match with the monthly series in terms of its monthly average. Using the same idea as we described in the first step, given a month m with w_m weekly observations, calculate:

$$\omega_m = \frac{1/w \sum_w GTW_{w,m}}{GTM_m^{\text{observerd}}}$$

where $m \in \{m_0, \cdots, m_T\}$ represents the months inside the $j \in \{0, \cdots, T\}$ periods. After calculating each monthly-weight ω_m, apply it over the weekly-

observations and define the final-adjusted series as $\boldsymbol{GTF}$:

$$\boldsymbol{GTF} = (GTW_{w_{0,m_0}}\omega_{m_0}, GTW_{w_{1,m_0}}\omega_{m_0}, \cdots, GTW_{w_{0,m_j}}\omega_{m_j}, \cdots, GTW_{w_{W,M}}\omega_M)$$

We exemplify this strategy after defining the enchainment method that we use to replicate the monthly series every week. For example, figure 2.8 reveals the result after applying this method over a "Inflation" search term. Note that after joining the series and applying the first-weighting, the monthly average of such series does not coincide with the observed "Inflation" monthly series that Google provides. However, the second weighting procedure forces the monthly average of the weekly series to be centered over the monthly observed series. Thus, we can replicate the weekly behavior of the series for longer horizons and use the variables as regressors on the nowcasting problem.

2.2.3 Hard Data

After describing the two sets of unstructured data, we now focus on the usual macroeconomic indicators on the nowcasting problem. First, we divide the usually structured indicators as they have been collected from different sources. In all cases, we collected first-vintages of all time-series to avoid "look-ahead biases" on the out-of-sample validation of the forecasts.

The first primary source was Bloomberg ("BBG") historical data for Brazilian and international macroeconomic and financial indicators. The BBG data comprehends 99 series divided into financial, commodities, activity, credit and loans, inflation rates and cost indexes, unemployment and employment, confidence, fiscal policy and tax revenues, inflation rates and monetary policy, and external accounts time-series. We present a list of all series in detail in Table 2.4.

The second source was micro-level data from the balance of trade collected from the Ministry of Economy website.[23] We extracted only imports and exports data from the highest aggregation considering ten categories.[24]

The third data source considered alternative structured indicators. Firstly we collected hydrologic stored energy, demand for energy, and the

[23]The Balance of Trade data has been downloaded from the "CUCI" spreadsheet from the "Balança Comercial - Dados consolidados" section.

[24]Categories: 1) food products and live animals; 2) Beverage and Tobacco; 3) Raw materials, inedible, other than fuel; 4) Mineral fuels, lubricants, and related materials; 5) Animal and vegetable oils, fats, and waxes; 6) Chemical and related products; 7) Manufactured articles, classified primarily by material; 8) Transport machinery and equipment; 9) Miscellaneous manufactured articles and 10) Goods and transactions not specified elsewhere.

marginal cost of operation data from the ONS website[25] which consolidates energy data at the regional level for Brazil. Next, we also collected temperature data at the regional level from the "Instituto Nacional de Meteorologia" website.

Finally, we collect market expectations from "Boletim Focus" from the Central Bank of Brazil for quarterly Gross Domestic Product (GDP) on a Year-over-Year basis. Such variable is fundamental on the nowcasting model as it will work as the nowcasting benchmark. Expectation data is provided daily for a full-week period covering Monday-to-Friday and released over the following Monday. As the central bank of Brazil ranks institutions according to their forecast accuracy ("Top-5 Institutions"), market agents aim to provide good forecasts for the short, medium, and long term for GDP and other macroeconomic variables. Therefore, we collect market expectation data for the respective nowcasting release and longer horizons (about six quarter-ahead).

A critical remark is that the frequency of structured indicators may vary across the series and their release lags. For example, financial, commodities, and temperature data are observed daily, but they are released with different lags: the first two are daily-disposable, whereas the last is only disposable after a month. The balance of trade data may be released at a weekly frequency (preview of the headline), while its components are only released after a month. Fiscal policy and activity indicators such as industrial production, retail sales, and services are monthly indicators that may be released at least one month after the reference date. Finally, GDP data is observed only quarterly and released almost a quarter later. Therefore, the nowcasting model should deal with mixed-frequency time-series observations to extract as much information as possible.

2.3 Nowcasting Model

Dealing with economic nowcasting in a data-rich environment is challenging, as new issues may arise due to mixed-frequencies indicators. Furthermore, while usual forecasting models consider regressors and target variables sampled at the same frequency, nowcasting induces the necessity to deal with multiple frequencies simultaneously. Therefore, we opt to start describing all sources of data indicators prior to the model specification in order to motivate the mixed-sampling problem and the respective modeling solution adopted.

[25]ONS is the acronym of "Operador Nacional do Sistema Elétrico", i.e., the national operator of the electric system.

Before describing the model, we explore some differences in data frequencies: textual news data (N-grams and media-attention) have been captured at daily-frequency, Google Trends and Financial indicators are sampled at weekly-frequency. In contrast, usual macroeconomic indicators may be disposable solely at monthly or quarterly-frequency. Therefore, we choose a weekly frequency as the highest-frequency variable to simplify the model and reduce noise. Therefore, all news counts and media attention have been resampled and aggregated at a weekly frequency.[26]

As the target variable of the nowcasting model is the quarterly Year-over-Year (YoY) Gross Domestic Product (GDP), and we deal with inter-quarterly data, we adopt a Mixed Data Sampling model, a common approach in nowcasting models (see Bai et al. (2013)). Such models allow the usage of high-frequency variables (and their lags) in order to explain low-frequency variables. Furthermore, we opt for the U-MIDAS approach, as it reveals good properties while forecasting quarterly variables, as described by Foroni et al. (2015).

Following, we define a nowcasting model with similar characteristics to Borup et al. (2021) suggestion, but with proposing three additional modifications: (i) Brazilian GDP data is only released about one-quarter of its official reference date, which allow the inclusion of backcasting variables; (ii) we add an error structure to Borup et al. (2021) formulation; (iii) instead of working with two different forecasting frequencies, we generalize the model in order to forecast GDP with both structured, unstructured and textual data in many frequencies (weekly, monthly and quarterly).

The nowcasting modelling idea consists in defining a U-MIDAS model, as in Foroni et al. (2015), with the following functional form:

$$y_t = f^{(j)}\left(y_{t-1}, \cdots, y_{t-j}, \boldsymbol{X_t}^{(j)}, u_{t-1}^{focus(j)}, \cdots, u_{t-m}^{focus(j)}; \boldsymbol{\beta}^{(j)}\right) \tag{2-1}$$

where y_t is the quarter-t target variable; $(y_{t-1}, \cdots, y_{t-j})$ are lags of the dependent variable; $(u_{t-1}^{focus(j)}, \cdots, u_{t-m}^{focus(j)})$ incorporates an order m market expectation forecasting error (namely simply as "focus" from now on), which is defined as the difference between the market expectations forecast and the actual release value of the GDP, i.e. $u_t^{focus(j)} = y_t^{focus(j)} - y_t$ to the structure of the model; $\boldsymbol{X_t^{(j)}}$ is a $K_x \times 1$ vector of predictors, that can be decomposed in:

$$\boldsymbol{X_t}^{(j)} = \begin{bmatrix} \boldsymbol{x_{w,t}^{(j)}}' & \boldsymbol{x_{m,t}^{(j)}}' \end{bmatrix}' \tag{2-2}$$

where $\boldsymbol{x_{w,t}^{(j)}} = \begin{bmatrix} \boldsymbol{x'_{w,t-j/W_t}} & \boldsymbol{x'_{w,t-(j+1)/W_t}} & \cdots & \boldsymbol{x'_{w,t}} & \cdots & \boldsymbol{x'_{w,t+(k)/W_{t+1}}} \end{bmatrix}'$

[26]For N-grams, the counts of terms have been summed, and the media-attention series was calculated averaging daily media-attention.

denotes a vector of $W_t + W_{t+1}$ weekly observed variables on quarter t.[27] As we also include backcasting possibility, the vector of variables $x'_{w,t+(k)/W_{t+1}}$ allows the usage of weekly-variables that refer to quarter $t+1$ on the nowcasting exercise for quarter t.[28]

Also, $\boldsymbol{x}^{(j)}_{\boldsymbol{m,t}} = \left[\boldsymbol{x'_{m,t-i/3}} \quad \boldsymbol{x'_{m,t-(i+1)/3}} \quad \boldsymbol{x'_{m,t-(i+2)/3}} \quad \cdots \quad \boldsymbol{x'_{m,t+(k)/3}}\right]'$ denotes a vector of monthly observed variables.[29] Therefore, the vector of regressors $\boldsymbol{X^{(j)}_t}$ is a $K_x \times 1$ has the dimension of $K_x = (W_t + W_{t+1})K^w_x + (M_t + M_{t+1})K^m_x$ variables, where K^w_x denotes the total number of weekly variables, W_t denotes the total number of weeks on quarter-t (including backcasting possibility W_{t+1}), K^m_x the total number of monthly variables and M_t the total number of monthly variables considered (also allowing backcasting variables inclusion M_{t+1}).

Also, in terms of Equation (2-1) parameters $\boldsymbol{\beta}$, we have:

$$\boldsymbol{\beta}^{(j)} = \left[\alpha^{(j)} \quad \boldsymbol{\phi}^{(j)\prime} \quad \boldsymbol{\beta}^{(j)\prime}_{\boldsymbol{w}} \quad \boldsymbol{\beta}^{(j)\prime}_{\boldsymbol{m}} \quad \boldsymbol{\xi}^{(j)\prime}\right]' \tag{2-3}$$

where α^j denotes a typical constant term, a vector of coefficients for lagged dependent variables $\boldsymbol{\phi}^{(j)\prime}$, a vector of coefficients for weekly independent variables $\boldsymbol{\beta}^{(j)\prime}_{\boldsymbol{w}}$, a vector of coefficients for monthly independent variables $\boldsymbol{\beta}^{(j)\prime}_{\boldsymbol{m}}$ and a vector of coefficients for lagged market expectation error terms $\boldsymbol{\xi}^{(j)\prime}$.

Two important points regarding the $f^{(j)}$ functional of Equation (2-1) arise. Firstly, the information flow is such that when a given variable is released, it is included in the model of Equation (2-1) and stays in the information set of the model. Therefore, for each week $w_t \in \{1, \cdots, W_t\}$ of quarter t, we generate a new forecast, generating a term-structure of predictions that are indexed by the time in which the forecast was generated, i.e., the forecasting date. Note that the highest frequency variable determines the forecasting frequency, i.e., even in the absence of (new) monthly data, forecasts are generated weekly.

A second point regards the functional form $f^{(j)}$ used in Equation (2-1).

[27] As an example, in a usual quarter $W_t = 12$. However, as some quarters may have more or less weeks, we set W_t as an arbitrary indicator of the number of weeks of quarter t. But as we also allow the inclusion of backcasting variables, we should also consider the number of weeks of the following quarter W_{t+1} prior to the GDP data release.

[28] For notation simplicity, suppose that a quarter is composed by twelve weeks. Therefore, for a typical quarter t, the vector $\boldsymbol{x_{w,t-j/12}}$ denotes the first week of the quarter, $\boldsymbol{x_{w,t-(j+1)/12}}$ denotes the second week of the quarter and so on. However, as we may include backcasting possibility, we can use variables that refer to quarter $t+1$ to nowcast the GDP of period t. In this case, $x'_{w,t+1/W_{t+1}}$ denotes the first week of the following quarter, $x'_{w,t+2/W_{t+1}}$ the second week and so on.

[29] For a typical quarter t, the vector $\boldsymbol{x'_{m,t-i/3}}$ represents the first month of the quarter, $\boldsymbol{x'_{m,t-(i+1)/3}}$ the second month and $\boldsymbol{x'_{m,t-(i+2)/3}}$ the third month of the quarter. As in the weekly-variables exercise, we allow backcasting variables. Therefore, $\boldsymbol{x'_{m,t+1/3}}$ denotes a monthly variable that refers to the first month of the subsequent quarter $t+1$. Thus, a typical quarter-t may consider $M_t + M_{t+1}$ monthly nowcast and backcast variables.

We adopt only linear specifications in order to generate comparable results and a simple marginal analysis: in a linear specific, verifying the marginal impact of a given release over the dependent variable is straightforward. Also, as we lie in a high-dimensional setting, the shrinkage methods discussed in the next section embed good theoretical properties that allow the correct model identification.

2.3.1 Specifications

There are mainly four specifications used on the nowcasting exercise and an additional market expectations benchmark: a simple AR(1) process, a linear expectation model, a shrinkage specification, and a final ensemble of the shrinkage model with the benchmark forecast. The market expectations benchmark model consists of weekly market expectations for GDP provided by the Central Bank of Brazil on "Relatório de Mercado - Focus". Such indicator is the median forecast collected week by week from all leading financial institutions in Brazil.[30] The focus market expectations are usually set as a benchmark to be beaten by Brazilian financial institutions, as it is almost a real-time measure of the economy (e.g., Garcia et al. (2017)). Therefore, from now on, we will denote the focus benchmark forecast as $\widehat{y_t^{\text{focus}(j)}}$, where j denotes the week of reference in which the nowcast has been made.

The first specification consists in a simple first-order autoregressive model, which is a common adoption as benchmark estimation on the literature (e.g. Ellingsen et al. (2020), Borup et al. (2021) and also in Brazil for Garcia et al. (2017) and Medeiros et al. (2020)).

$$y_t = \mu + \phi_1 y_{t-1} + u_t \quad \Rightarrow \quad \widehat{y_t^{\text{AR}}} \tag{2-4}$$

where μ is a constant term, ϕ_1 is the coefficient term associated with autoregressive term and $\widehat{y_t^{\text{AR}}}$ is the AR forecasting.

The second specification is a linear model over market expectations (LM). The idea is to use market expectations as the sole regressor on a univariate linear model as a form of rationality test (see the precursor paper of Lovell (1986)). In this framework, there are rationality deviations if we can generate systematic better forecasts using a simple linear expectation model.

$$y_t = a + b\widehat{y_t^{\text{focus}(j)}} + v_t \quad \Rightarrow \quad \widehat{y_t^{\text{LM}(j)}} \tag{2-5}$$

where a is a constant, b is the slope term for the LM model, $y_t^{\text{focus}(j)}$ is the j-th week market expectations for each quarter t, and $\widehat{y_t^{\text{LM}}}$ is the resulting LM

[30] The survey covers about 140 institutions.

forecasting. On such a framework, the simple linear regression has the objective of verifying if there exists a linear combination of market expectations that may perform better than itself in terms of forecasting accuracy.[31]

The third set of models consider a more general and complete linear functional form for the $f^{(j)}$ function. Consider a multiple regression scheme composed by a vector $\boldsymbol{X_t}$ with dimension $K_x \times 1$ (recall that $\boldsymbol{X_t^{(j)}} = \left[\boldsymbol{x_{w,t}^{(j)\prime}} \quad \boldsymbol{x_{m,t}^{(j)\prime}}\right]'$ and $K_x = (W_t + W_{t+1})K_x^w + (M_t + M_{t+1})K_x^m$) with the addition of j-dimensional vector of lagged dependent variables $\boldsymbol{y_{t-j}}$ and m-dimensional vector of lagged market expectation error $\boldsymbol{u_{t-m}^{focus(j)}}$ term:

$$y_t = \alpha^{(j)} + \boldsymbol{\phi_h^{(j)\prime}}\boldsymbol{y_{t-j}} + \boldsymbol{\beta_w^{(j)\prime}}\boldsymbol{x_{w,t}^{(j)}} + \boldsymbol{\beta_m^{(j)\prime}}\boldsymbol{x_{m,t}^{(j)}} + \boldsymbol{\xi_t^{(j)\prime}}\boldsymbol{u_{t-m}^{focus(j)}} + e_t \quad \Rightarrow \quad \widehat{y_t^{\mathrm{LS}(j)}} \tag{2-6}$$

however, as typically the vector $\boldsymbol{X_t^{(j)}}$ has dimension $K_x \gg T$, i.e., the functional form suggested in Equation (2-6) cannot be estimated by Ordinary Least Squares (OLS) regression, as it yields a high-dimensional problem.

There are many forms to proceed with dimensionality reduction.[32] However, we opt to adopt two linear shrinkage models to estimate Equation (2-6) based on penalized regression: Least Absolute Shrinkage and Selection Operator (LASSO), developed by Tibshirani (1996), and Adaptative-LASSO, introduced by Zou (2006).

The idea of using LASSO and AdaLASSO (LA and AL from now on) are directly associated with its basic properties: both are linear models (allowing direct marginal impact calculation by taking the derivative with respect to the variable), which recur to a penalty function that "shrinks" to zero parameters associated with redundant predictors. Moreover, as Zhao & Yu (2006) shows, LASSO can generate model selection consistency, i.e., the probability of LASSO selecting the true model (the one which assigns non-zero coefficients only to relevant predictors) approaches to one at an exponential rate.[33]

However, the LASSO model, when in a high-dimension setup with $K_x \gg T$ tends to select arbitrarily one predictor from a group of highly correlated predictors (see Meinshausen & Buhlmann (2004) critique). To address such

[31]Note that if there exists a and b such that yields better forecasting than the expectation $y_t^{\mathrm{focus}(j)}$ itself, then there are systematic gains over market expectations. Therefore, a lack of "Rational Expectations" results in systematic gains. Thus, if "Rational Expectations" hold, we would expect that there is no linear combination that performs better in terms of forecasting accuracy than itself, i.e., an application of what Lovell (1986) suggests.

[32]e.g., Principal Component Analysis (PCA), shrinkage (linear) models, machine learning (nonlinear) models as Random Forests (RF), and Neural Networks (NN)

[33]Necessary conditions are that the number of relevant variables p and irrelevant variables s are large and assume only a finite second moment of the noise.

a problem, we also estimate an AdaLASSO model, which consists of a two-step method: (i) on the first-stage estimation of LASSO and a rescaling of the regressors; (ii) in a second-stage we estimate a LASSO model based on the scaled regressors. As showed by Zou (2006), the AdaLASSO presents the oracle properties (i.e., selecting the correct model by setting irrelevant coefficients to zero) over much milder conditions than the LASSO model.

Mathematically, the problem of the penalized regression model consists in selecting a hyperparamenter λ such that given the vector of variables, minimize the loss-function $L_j(.)$ with respect to the vector $\boldsymbol{\beta}^{(j)}$ of coefficients (including a constant term) and also minimizing the penalty function $\rho_{\lambda,k}(.)$ with respect to the vector of coefficients solely (given by $\boldsymbol{\beta_x}^{(j)} = [\boldsymbol{\phi}^{(j)\prime} \quad \boldsymbol{\beta_w}^{(j)\prime} \quad \boldsymbol{\beta_m}^{(j)\prime} \quad \boldsymbol{\xi}^{(j)\prime}]'$, with dimension $(j + K_x + m) \times 1$:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\boldsymbol{\beta}\in\boldsymbol{B}} \left[\sum_{t=1}^{T} L_j(e_t) + \lambda \sum_{k=1}^{j+K_x+m} \rho_{\lambda,k}(\beta_x; data, w_k) \right]$$

where e_t is the error term of Equation (2-6), and the loss-function $L_j(.)$ can denote either a quadratic loss-function $L_j(e_t) = e_t^2$ or a quantile check-function $L_j(e_t) = \tau e_t \mathbb{I}(e_t > 0) + (\tau - 1)e_t \mathbb{I}(e_t \leq 0)$. This two loss-functions allows modelling the average (due to the quadratic loss) and median (due to the check-function with $\tau = 0.5$), respectively. These two approaches (mean vs median) yields different properties when generating the nowcasting term-structure, as described in Section 2.4.

The penalty function ρ_k can denote either a simple l1-norm (in the case of LASSO), where $\rho_k(\beta_k; data, w_k) = |\beta_k|$ or a pre-weighted l1-norm (in the case of AdaLASSO), where $\rho_k(\beta_k; data, w_k) = w_k|\beta_k|$. In terms of weights, w_k, the LASSO model set $W_k = 1$, while the AdaLASSO sets $w_k = |\tilde{\beta_k}|^{-1}$ where $\tilde{\beta_k}$ is an initial LASSO estimator.[34] Therefore, the AdaLASSO model is simply a generalization of the LASSO model, allowing the usage of different weights for variables, in order to produce the Oracle property as described by Zou (2006).

In order to select the optimal λ, we recurred to the usage of the Bayesian Information Criteria (BIC) to select the $\lambda_{min} \in \text{argmin}_{\lambda\in\Lambda} BIC(\lambda)$.[35] Due to the arbitrary selection of regressors issue of LASSO, we also recur to a Principal Component Analysis (PCA) scheme over textual regressors. As the

[34]This is the reasoning why the penalty function $\rho_k(\beta_k; data, w_k)$ depends on the data. For additional information, check Zou (2006).

[35]The BIC is simply given by $BIC = \log\left[\hat{\sigma}^2(\lambda)\right] + df(\lambda)\frac{\log(T)}{T}$, where df($\lambda$) denotes the degrees of freedom of LASSO (or AdaLASSO) associated with λ, which are simply the number of variables in the active set (variables associated with non-zero coefficients) and $\hat{\sigma}^2(\lambda) = \frac{1}{T-df(\lambda)}\sum_{t=1}^{T} e_t^2$. For additional referece, check Zou et al. (2007).

N-grams matrix of term-counts per week has more than 16,000 unique terms, to avoid term-overfitting, we proceed with PCA over the N-grams matrix selecting components that explain 80% of the total (N-gram) variance. Such an approach can reduce the problem to around 100 components. The result is a series of components that summarize information of the N-gram matrix, similar to the media-attention topics generated using LDA.

Finally, we generated ensemble models based on simple linear combinations of the forecasts generated based on LASSO and AdaLASSO linear (mean) and quantile (median) regression and focus market expectations. A critical remark is that market expectations are already included inside the weekly hard-data regressors. Therefore, an ensemble considers both, directly and indirectly, market expectations:

$$\widehat{y_t^{\mathrm{EM}(j)}} = \frac{\widehat{y_t^{focus(j)}} + \widehat{y_t^{LS(j)}}}{2} \tag{2-7}$$

the motivation to ensemble models comes from an empirical observation that the correlation between the residuals of the LASSO and AdaLASSO with market expectations from Focus is small. We describe such observation in detail over Section 2.4.

2.3.2 Estimation Framework and Information Flow

The estimation framework is based on a recursive train-and-test procedure. The observations from 2009Q1 to 2016Q4 are used solely to estimate the model in-sample (32 data points). As the number of periods is relatively small, we opt to use an expanding-window procedure to expand the number of observations to test the accuracy of the model.[36] The method consists of a recursive evaluation of the model accuracy, e.g., for the 2017Q1 GDP, all observations from 2009Q1 to 2016Q4 are used to estimate the model (in-sample) and evaluate the forecast (out-of-sample) for this reference GDP;[37] for 2017Q2, all observations from 2009Q1 to 2017Q1 are used to estimate the model, and we evaluate the results for 2017Q2. This procedure is iterated until 2021Q3 last GDP forecast.

Adopting an expanding window rather than a rolling-window procedure has advantages and disadvantages. The main advantages are that setting

[36]The algorithm used to develop the model has the option of selecting expanding or rolling window. However, preliminary tests suggested that rolling-window accuracy is worse than expanding-window training. This effect can be attributed to the relatively small number of data points to train the model.

[37]Recall that for a quarter t, we made typically W_t different forecasts, where W_t represents the number of weeks in quarter t. We make at least $W_t = 12$ different forecasts for a typical quarter.

the initial period of the out-of-sample evaluation does not affect the sample size of the subsequent periods, and later periods embed a higher number of observations to test the model. However, the main disadvantage is that expanding window typically propagates the inertial behavior of the series throughout the sample, e.g., a variable that helped to explain the financial crisis of 2008 may not explain the Covid-19 crisis of 2020, although it may be selected as a predictor in both different periods.

Following Borup et al. (2021) we opt to carefully describe the information flow of forecasts as it helps to explain the model. Recall that the reference date of the GDP (e.g., 2017Q1, i.e., January 2017 to March 2017) does not coincide with the release date of the GDP (usually, the GDP is only released with a lag of three months, around one-quarter after its reference). Therefore, some monthly and weekly variables from the next quarter may be disposable and can backcast the GDP. Furthermore, our nowcasting design allows future variables to forecast past variables, as they are available at the moment and maybe part of the information set.

Table 2.5 presents one example for 2017Q1 and 2017Q2 relative to the information flow. The nowcasting is only computed after the last release, so the forecast date does not coincide with the first week of the reference date. Therefore, when the forecast has been made in the first week, some usual macroeconomic indicators may be available for the first or second months. Textual data and Google Trends for the whole reference date are available for the first weeks, and then data covering the next quarter can also be used to forecast the current quarter GDP. However, only in the last weeks may all macroeconomic indicators (such as industrial production) be available to predict the current GDP target.

As a result, Table 2.5 shows a pattern that happens throughout the sample: unstructured and financial data becomes available sooner than usual macroeconomic indicators, which allows the inclusion of helpful information on the nowcasting model prior to the availability of structured indicators. Therefore, on the first two or three weeks relative to the current quarter, unstructured data may be available to cover all information relative to the reference period that GDP data is covering. Such observation is vital to understand the results displayed in the next section.

2.4 Results

We define six different data-based models estimated using the methodology and data described earlier. The idea is to isolate the effects of using

different sources of data and their impacts in terms of forecasting capabilities. The first model (HARD) includes only hard-data variables displayed in Table 2.4. The second model (GT) estimates the first model with the addition of Google Trends search data terms from Table 2.3. The third model (PCA) estimates the first model by adding principal components of the N-grams counts. The fourth model (PCA/ATT) estimates the third model with all sixty media-attention series. The fifth model (PCA/GT) estimates the third model with Google Trends search terms. Finally, the sixth model (ALL) estimates the nowcasting using all available data sources.

Also, as described earlier, after estimating the specifications of Section 2.3.1 we collected the residuals generated for each week of forecasting and compared them with the residuals of the market expectations forecast. The correlation analysis is displayed in Table 2.6, all using market expectation (Focus) as the benchmark. Note that the correlation of the shrinkage models with Focus is relatively small (less than 50%) and positive. The positive sign is expected since the model includes Focus as a predictor, but the low correlation signals that there may be a combination between both forecasts that may overperform each one individually, which is described in the next section.

2.4.1 Out-of-Sample Results

The out-of-sample results, covering all the term-structure of nowcasts made through the weeks over each quarter from 2017Q1 up to 2021Q3, are displayed in Table 2.7. All metrics have been normalized with respect to the Focus benchmark, as it is commonly used as a reference in Brazil (as in Garcia et al. (2017)). In addition, all estimations adopted an expanding window framework.

The first important remark is that shrinkage models cannot over-perform the Focus benchmark in forecasting accuracy, measured by the Mean Squared Error (MSE) and Mean Absolute Error (MAE).[38] However, as stated previously, due to the low correlation between the residuals of the shrinkage models and the Focus benchmark, ensemble models displayed good prediction accuracy throughout the weeks, measured by a lower than benchmark MSE and MAE. In order to evaluate if the predictions are different from the Focus benchmark,

[38]The Mean Squared Error (MSE) is simply defined as: $\text{MSE} = \frac{1}{n}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$, while the Mean Absolute Error (MAE) is defined as: $\text{MAE} = \frac{1}{n}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|$. As the MSE squares the prediction errors, it may over-penalize periods with higher uncertainty and, therefore, it will generate high MSE. Conversely, the MAE, while taking the absolute value of the prediction errors, induces linearly-weighted penalties to abnormal errors. Therefore, the two measures are often used to evaluate the model, focusing on different qualitative characteristics of the nowcasting model.

we also proceeded with Diebold-Mariano[39] test in terms of MSE and MAE to evaluate the difference between the model's forecast and the benchmark.

The second important aspect regards the best-performer specifications. In particular, ensemble models using only hard data, despite the LASSO ensemble, do not provide better results than the benchmark, as the ratio of the MSE and MAE with respect to Focus is close to one (Table 2.7). The surprisingly good results of the ensemble using the LASSO model and hard data are not followed by other specifications. Such particular behavior does not happen over all other data sources, raising a possible flag for non-robust results or local validity only.

The third and most important aspect regards the inclusion of unstructured data sources and their crucial role in enhancing the overall nowcasting quality. All columns from Table 2.7 that considered any combination of hard data with alternative data displayed significantly better results than solely considering hard data (or even the own benchmark). In particular, the combination of the principal components of N-grams and Google Trends data with hard data provided a robust decrease (through all metrics) of the forecasting error.

Nonetheless, Table 2.8 presents the Diebold-Mariano test, suggesting that all predictions based on the ensemble model, besides having better accuracy, are statistically different from the benchmark. Therefore, through all weeks of reference, the models that use unstructured data sources can deliver nowcasting capabilities that systematically deliver statistically better forecasts than the own benchmark when combined with market expectations.

We emphasize that the overwhelming performance of the weekly specifications using ensemble between the benchmark and shrinkage models allowed systematic (and not isolated cases of) better prediction accuracy. As we describe over the next section, such a result can be explained due to the increase in the informational component of Focus (mainly at the beginning of the nowcasting horizon) that tends to converge to the actual data release.

2.4.2 Zooming In The Term-Structure of Nowcasts

Figure 2.9 shows the evolution of the best-performers ensemble models for nowcasting GDP using hard data with the addition of PCA over newspaper terms counts and Google Trends series (PCA/GT columns of Table 2.7). The period in reference covers all weeks of each quarter starting at the release of

[39]Diebold-Mariano test consists in evaluating if the distribution of the residuals from a benchmark model and the model in interest have similar (gaussian) distribution. For more details, see Diebold & Mariano (2002).

the last GDP number up to one day prior to the reference quarter GDP data release. Such plots illustrate two crucial points that help explain the systematic over-performance of the ensemble models: (i) using the benchmark on the ensemble model anchors the term structure of predictions to a consensus that, by its characteristics, embeds all the information set disposable at the moment; (ii) unstructured data introduces valuable information at the beginning of the term structure of predictions, which can generate initial forecasting accuracy over the first weeks.

To illustrate both points, consider zooming Figure 2.9 to the first week of prediction of all quarters. The result is the plot present in Figure 2.10. Note that the shrinkage models are presented in green, while ensemble models are presented in blue. The shrinkage models, by themselves, are often much more volatile than the forecasts from the benchmark. However, combining the benchmark with shrinkage specification reduces the series volatility and anchors forecasts around the market consensus. Such a combination still embeds valuable information that induces better predictions.

In Table 2.9 we present the MSE and MAE solely for the forecasts made in the first week of the term structure (the ones present in Figure 2.10). Note that the ensemble model based solely on hard data cannot overperform the benchmark. However, including alternative data sources as news data (either from media-attention series or PCA of the term counts) or Google Trends can reduce the MSE by at least 30% and the MAE by 10%.

Another essential point regards the last prediction of each quarter, i.e., one day prior to the GDP release. Figure 2.9 shows that the benchmark tends to smoothly converge to the actual number throughout the weeks, while the nowcasting model presents a more volatile path. When zooming in to the last prediction, Figure 2.11 shows that the primary deviations of the benchmark and the actual value tend to disappear, while shrinkage may suggest different predictions. Table 2.10 shows that, in terms of MSE, all ensemble prediction displays lack-of-performance compared to the benchmark.

A possible explanation to the phenomenon present in Figures 2.10 and 2.11 regards the information set disposable at each period. At the beginning of the nowcast term structure, only a few usual macroeconomic indicators are disposable, and the agents only have access to a limited quantity of information regarding the GDP. In such a case, alternative data sources introduce new information that is valuable in order to forecast GDP. However, as time passes, agents incorporate all disposable information (and even other non-quantitative sources) that may help explain the GDP. Therefore, the market expectation converges to the actual release.

2.4.3 Model Interpretation and Further Details

A final important element regards the selection of the predictors. As LASSO and (mainly) AdaLASSO shrink irrelevant variable coefficients towards zero, one can evaluate the frequency in which a particular variable has a non-zero coefficient and generate a frequency plot throughout the out-of-sample analysis.

In order to illustrate such a procedure, we continue to use the best-performer hard data with the addition of PCA/GT as the variable reference. Using the ensemble model as a reference, the benchmark is selected with 50% frequency by construction. Therefore, we use the shrinkage model solely to generate a variable selection to understand the variable selection over the out-of-sample analysis.

Consider a single LASSO model using a check function, i.e., a median quantile regression. The results are displayed in Figure 2.12. We start by highlighting the fact that we suggested in the last section: only a limited number of hard data indicators are available (mainly the ones regarding the first month of the reference period). Therefore, only the first Industrial Production number and later market expectation variables are selected. The other variables come mainly from Google Trends ("_gt" suffix variables) or are components from the N-grams matrix ("comp_" prefix variables). However, as time evolves, the last prediction of the nowcast prior to the data release selects all variables covering the current quarter as Industrial Production from all months from the current quarter and other hard data indicators. Nonetheless, Google Trends variables, principal components from N-grams, and error terms based on lagged market expectations appear for last week's prediction of the nowcast.

Finally, we explore an observation based on the option to model the mean and median of the GDP forecasts, based on a quadratic loss function and a check-function for quantile regression as stated in Section 2.3. Note that Figure 2.9 reveals that quantile regression can generate a term-structure path much less volatile than the quadratic loss-function general model. Such a pattern is observed through all the out-of-sample results. This observation emerges naturally while modeling the median of the GDP, as this quantity is much more robust to outliers and less volatile than the average measure.

Note also that the forecasting accuracy of quantile regression is the best-performer for some types of data mixtures. Furthermore, we opt to use quantile (median) and linear (mean) regression to generate a more robust approach. After that, we generated a final ensemble model that can embed the lower

volatile patterns from the quantile regression with good forecasting accuracy from the mean modeling, reducing the estimation noise elsewhere.

Figure 2.13 presents the evolution of this general ensemble model for all data types discussed earlier. We highlight that the general approach clarifies better prediction accuracy due to alternative data adoption. Also, we highlight that the term structure is much less volatile and presents the convergence towards markets expectation as stated previously.

2.5 Conclusion

This paper develops a nowcasting model to track the Brazilian GDP on a weekly basis. We collect alternative data sources like newspaper articles and Google Trends search data and conventional macroeconomic indicators and other structured series such as market expectations, temperature, the balance of trade data, and energy to generate a data-rich environment. The data sources have been combined with an Unrestricted Data-Sampling (U-MIDAS) framework in such a form that each variable was skip-sampled at a quarterly frequency. We estimate each variable's coefficient for each data frequency using either a LASSO or an AdaLASSO shrinkage model that can induce sparsity and also has the model selection property for the latter. We estimate the models for average and median (linear and quantile regression), and we generated a term structure of nowcasts for each quarter.

The results suggested that alternative data sources on a simple shrinkage model cannot beat the benchmark given by market expectations for GDP. However, we highlight that combining the shrinkage specification with the market expectations can generate systematic gains while decreasing the MSE and MAE throughout the out-of-sample window. Furthermore, we argue that such a combination yields a powerful approach due to the low residual correlation of the models that compose the ensemble.

Next, we also argue that the nowcasts are better at the beginning of the term structure when only a few conventional macroeconomic series are disposable. In such cases, the inclusion of alternative data sources can induce an even lower MSE and MAE when compared to the benchmark. However, the gains fade as time passes, and close to the GDP release, the benchmark is hard to beat, as all information is encoded on the market expectation from agents.

Then, we explore the variables which have been selected, depending on the point in which we analyze the nowcasting term structure. We argue that alternative data is more selected over initial predictions than usual data.

However, as time passes by, conventional data sources that become available embed highly informative data. We also discuss the choice of modeling the average or median in terms of usual or quantile regression. We argue that quantile regression generates a less volatile term structure of predictions, while usual regression has better explanatory power. Then, we suggest that combining LASSO and AdaLASSO with average and median modeling yields a "general" model, which is a best-performer, as it reduces estimation noise and tends to be more robust.

Finally, we provide an initial analysis of the effects of combining both conventional and alternative data sources. All results suggest that alternative data can encode valuable information that can be used to enhance the quality of forecasts. Our results reiterate the current findings of the nowcasting literature that uses non-conventional data sources as predictors, such as Bybee et al. (2020), Thorsrud (2016), Ellingsen et al. (2020) and Saiz et al. (2021). Also, our approach explicitly decomposes the gains of alternative data adoption and explains the higher benefits of such inclusion whenever usual macroeconomic indicators are not available.

However, we also note that these results are particular to the Brazilian economy, and there is plenty of space for additional research. A natural extension of this paper is analyzing whether the information component from alternative data sources, as also noted by Saiz et al. (2021) for the Euro area, is disseminated in economies with more indicators, as is the case of the United States. We also suggest including and testing additional nonlinear specifications as in our preliminary studies, nonlinear models such as Random Forests did not display good nowcasting capabilities.

Tables

Table 2.1: Summary of News Collected

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Folha	598	5329	6012	5235	15476	30628	24048
Estadão	55298	94182	90182	92639	61469	54209	31085
Valor	-	-	-	65545	51057	45509	45048
Sum	55896	99511	96194	163419	128002	130346	100181

	2016	**2017**	**2018**	**2019**	**2020**	**2021**	**Total**
Folha	22073	18479	19853	16375	15278	13988	194944
Estadão	35758	40258	34957	32405	36337	31459	692866
Valor	47010	46156	47001	46059	60773	64969	525309
Sum	104841	104893	101811	94839	112388	110416	1413119

[a] Evolution of the total news for the collected newspapers.

[b] Note that Valor Econômico only has a history of articles after 2012. Also, note that in 2009, Folha articles covered only the final months of this year. However, from 2010 onwards, Folha and Valor provide well approximations for the overall number of news.

Table 2.2: Descriptive Statistics for Weekly-Aggregated News

	Folha	**Estadão**	**Valor**	**Sum**
count	639	673	519	676
mean	302	1025	998	2067
std	156	500	216	558
min	12	244	58	126
25%	149	639	877	1819
50%	303	767	961	2037
75%	389	1376	1147	2323
max	903	2691	1709	4219

[a] Descriptive statistics for articles for each newspaper and in total every week.

[b] Note that "count" represents the total number of weeks collected for each newspaper.

Table 2.3: Google Trends Terms

Economy	Investment	Company
economia	bolsa	petrobrás
brasil economia	investimento	vale
pib	petróleo	itaú
serviços	gás	ambev
agricultura	ações	bradesco
agropecuária	títulos	weg
indústria	tesouro	santander
desemprego	fundos	banco do brasil
carteira de trabalho	bancos	jbs
fgts	crédito	oi
bolsa família	bolsa de valores	csn
comércio	bitcoin	gerdau
varejo	índice	suzano
inflação	previdência	cemig
ipca		sabesp
fiscal	Politics	magazine luiza
tributação	lula	telefônica
dívida	dilma	vivo
imposto	bolsonaro	copel
imposto de renda	temer	
ir	trump	Energy
taxa	marina	gasolina
taxa selic	serra	alcool
juros	ciro	diesel
poupança	presidente	energia
cdb	governador	hidrelétrica
dólar	prefeito	
euro	moro	Risk
câmbio	corrupção	risco
exportação	obama	crise
importação		vírus
safra		vacinação

[a] List of all Google Trends search terms collected from Google Trends API and used on the double weighting procedure to become predictors in the Nowcasting estimation.

Table 2.4: Hard Data Variables

Financial and Commodities	Inflation Rates and Cost Indexes	Fiscal Policy and Tax Revenues
Cboe Volatility Index	FGV Brazil General Prices IGP10 MoM	Brazil Central Govt Primary Ba
BRAZIL CDS USD SR 5Y D14	FGV Brazil General Prices IGP-	Brazil Central Government Prim
BRAZIL CDS USD SR 1Y D14	FGV Brazil General Prices IGP MoM	Brazil Federal Tax Income Nomi
BRAZIL CDS USD SR 10Y D14	FGV Brazil General IGPM YoY	Brazil Federal Public Debt
Brazil Commodities Index Compo	FGV Brazil CPI IPC-DI MoM	Brazil Total Federal Revenue
Generic 1st 'CO' Future	FGV Brazil IGP-M Construction	Brazil Public Primary Budget R
Generic 1st 'CL' Future	Brazil Basic Food Basket Sao P	Brazil Public Primary Budget
BRAZIL IBOVESPA INDEX	Brazil CPI IPCS Weekly	Brazil Public Primary Budget %
	Brazil CPI IPCA YoY	Brazil Public Nominal Interest
Activity	Brazil CPI IPCA MoM	Brazil Public Nominal Budget R
Anfavea Brazil Vehicle Sales	Brazil CPI INPC YoY	Brazil Public Net Debt
Anfavea Brazil Vehicle Sales L	Brazil CPI INPC MoM	Brazil Public Net Debt % of GD
Anfavea Brazil Vehicle Product	Brazil CPI IGPM Weekly Preview	Brazil General Government Gros
Anfavea Brazil Vehicle Exports	Brazil CPI Fipe YoY	
Brazil Retail Sales Volume Mon	Brazil CPI Fipe Weekly	Interest Rates and Monetary Policy
Brazil Retail Sales Volume MoM	Brazil CPI Fipe MoM	BRL SW BMF DI FUT 42DY
CNI Brazil Manufacture Industr	Brazil Dieese Sao Paulo Cost o	BRL SW BMF DI FUT 21DY
Brazil Auto Sales Subtotal	Brazil FGV Consumer General Price Index Second	Brazil Selic Target Rate
Brazil Amplified Retail S YoY	Brazil FGV Consumer General Price Index First	Brazil BNDES Long Term Interes
Brazil ABCR Total Traffic Flow	Brazil Producer Price Index YoY	Brazil Money Supply M4 MoM
Brazil Economic Activity GDP Y	Brazil Producer Price Index MoM	Brazil Money Supply M1 MoM
Brazil Economic Activity GDP M	Brazil IPCA-15 CPI Extended YoY	Brazil Monetary Base MoM
Brazil Industrial Production T	Brazil IPCA-15 CPI Extended MoM	
Brazil Industrial Production N		External Accounts
Brazil GDP YoY 1995=100	Unemployment and Employment	Brazil Current Account Monthly
Brazil GDP Qtrly Accumulated 4	IBGE Brazil Unemployment Six Metro Areas	Brazil Current Account Last 12
Brazil GDP QoQ SA 1995=100	IBGE Brazil Unemployment Rate	Brazil Trade Balance Weekly Ba
	IBGE Brazil Unemployment R 30	Brazil Trade Balance FOB Impor
Credit and Loans	Brazil CAGED Government Regist	Brazil Trade Balance FOB Expor
ACSP Brazil Default Loans Spc		Brazil Trade Balance FOB Balan
ACSP Brazil Default Loans Cons	Confidence Index	Brazil Net Intl Reserves
ACSP Brazil Bankruptcy Delinqu	Brazil CNI Industrial Confiden	Brazil Intl Reserves in Cash
Brazil Financial Total Outstan	Brazil CNI Consumer Confidence	Brazil International Reserves
Brazil Financial System Loans	Brazil FGV Consumer Confidence	Brazil International Daily Res
Brazil Financial Private Syste		Brazil Foreign Direct Investme
Brazil Personal Loans More Tha		

[a] List of all macroeconomic series collected from Bloomberg and used on the Nowcasting estimation.

Table 2.5: Information Flow Example

Reference Date	Release Date	Forecast Date	Brazil GDP YoY 1995=100	News and Media-Attention	Brazil CDS USD SR 5Y D14	Focus Expectation GDP YoY	Google Trends Searches	Brazil CPI IPCA YoY	Brazil Industrial Production
2017Q1	6/1/2017	3/19/2017	N	N	N	N	N	N	N
2017Q1	6/1/2017	3/26/2017	N	N	N	N	N	N	N
2017Q1	6/1/2017	4/2/2017	N	N+B	N+B	N+B	N	N	N
2017Q1	6/1/2017	4/9/2017	N	N+B	N+B	N+B	N+B	N+B	N
2017Q1	6/1/2017	4/16/2017	N	N+B	N+B	N+B	N+B	N+B	N
2017Q1	6/1/2017	4/23/2017	N	N+B	N+B	N+B	N+B	N+B	N
2017Q1	6/1/2017	4/30/2017	N	N+B	N+B	N+B	N+B	N+B	N
2017Q1	6/1/2017	5/7/2017	N	N+B	N+B	N+B	N+B	N+B	N+B
2017Q1	6/1/2017	5/14/2017	N	N+B	N+B	N+B	N+B	N+B	N+B
2017Q1	6/1/2017	5/21/2017	N	N+B	N+B	N+B	N+B	N+B	N+B
2017Q1	6/1/2017	5/28/2017	N	N+B	N+B	N+B	N+B	N+B	N+B
2017Q1	6/1/2017	5/31/2017	N	N+B	N+B	N+B	N+B	N+B	N+B
2017Q2	9/1/2017	6/11/2017	N	N	N	N	N	N	N
2017Q2	9/1/2017	6/18/2017	N	N	N	N	N	N	N
2017Q2	9/1/2017	6/25/2017	N	N	N	N	N	N	N
2017Q2	9/1/2017	7/2/2017	N	N+B	N+B	N	N	N	N
...	...	...	...	...	...	...	...	...	...
2017Q2	9/1/2017	8/31/2017	N	N+B	N+B	N+B	N+B	N+B	N+B

[a] Information Flow for selected variables for 2017Q1 and 2017Q2. Note that in the nowcasting model coexists three different dates: reference date, which indicates the period where the indicator refers, release date considering when the indicator was available, and forecast date considering when the forecast has been made. Note that the nowcast is only calculated when the last release is available, so the release date is who dictates the initial forecast date.

[b] The word "N" denotes whether this variable is used to nowcast the current state of the economy or if there is a latest release "B" that, in addiction to the indicator from the reference period, may be used to backcast the target variable.

Table 2.6: Correlation of Residuals

Correlation						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	0.1564	0.1564	0.1564	0.1564	0.1564	0.1564
LM	-0.2145	-0.2145	-0.2145	-0.2145	-0.2145	-0.2145
LA	**0.3325**	**0.2900**	**0.3663**	**0.3285**	**0.3519**	**0.2307**
AL	**0.4024**	**0.4882**	**0.4927**	**0.4605**	**0.5093**	**0.4608**
LA-QR	**0.2542**	**0.2745**	**0.3815**	**0.3413**	**0.3294**	**0.1491**
AL-QR	**0.4309**	**0.4357**	**0.4783**	**0.4768**	**0.4056**	**0.1748**
E-LA	0.7733	0.7708	0.7950	0.7799	0.7799	0.6676
E-AL	0.8054	0.8515	0.8514	0.8385	0.8652	0.8316
E-LA-QR	0.7427	0.7610	0.7959	0.7809	0.7397	0.5705
E-AL-QR	0.8288	0.8311	0.8509	0.8399	0.8057	0.6017
E-G	0.8031	0.8168	0.8359	0.8245	0.8110	0.6837

[a] AR1 denotes the first-order autoregressive model. LM denotes the linear model over market expectations. LA and AL denotes the LASSO and AdaLASSO models estimates using quadratic loss-function. LA-QR and AL-QR denotes the LASSO and AdaLASSO models estimates using quantile regression. All models starting with "E" denotes an ensemble between the model and market expectations from focus (e.g. E-LA is the ensemble with LASSO estimates; E-LA-QR is the ensemble with LASSO quantile regression). E-G denotes an ensemble of all shrinkage models estimated.

[b] All metrics have been normalized by market expectations from Focus (benchmark).

[c] Note that shrinkage models display a low (positive) correlation with market expectations (in bold).

Table 2.7: Nowcast Evaluation: MSE and MAE (Normalized)

Mean Squared Error - MSE

	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	42.6683	42.6683	42.6683	42.6683	42.6683	42.6683
LM	41.6059	41.6059	41.6059	41.6059	41.6059	41.6059
LA	1.4955	1.3256	1.3917	1.3869	1.5104	2.6420
AL	1.4129	1.1553	1.2068	1.2251	1.0506	1.3373
LA-QR	1.4914	1.3707	1.4815	1.4421	1.9968	3.7308
AL-QR	1.2193	1.2064	1.1806	1.3019	1.4421	3.3692
E-LA	0.8256	0.7474	0.8155	0.7905	0.8450	1.1007
E-AL	0.8432	0.8008	0.8228	0.8121	**0.7743**	**0.8518**
E-LA-QR	**0.7789**	0.7526	0.8536	0.8167	0.9834	1.3334
E-AL-QR	0.7935	0.7906	0.8037	0.8483	0.8554	1.2581
E-G	**0.7789**	**0.7490**	**0.7989**	**0.7880**	0.8309	1.0513

Mean Absolute Error - MAE

	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	4.7538	4.7538	4.7538	4.7538	4.7538	4.7538
LM	4.4497	4.4497	4.4497	4.4497	4.4497	4.4497
LA	1.1598	1.0438	1.1614	1.1282	1.1933	1.4845
AL	1.0112	0.9674	1.0236	1.0104	0.9567	1.0793
LA-QR	1.1214	1.0673	1.2289	1.1384	1.3330	1.7726
AL-QR	1.0085	0.9792	1.0662	1.0968	1.1782	1.6694
E-LA	0.9170	**0.8516**	0.9244	0.9021	0.8850	1.0260
E-AL	0.9240	0.8978	0.9170	0.9099	0.8811	**0.9094**
E-LA-QR	**0.8766**	0.8676	0.9335	0.9162	0.9429	1.1407
E-AL-QR	0.9056	0.9074	0.9108	0.9343	0.9002	1.0904
E-G	0.8952	0.8694	**0.9058**	**0.9013**	**0.8803**	1.0044

[a] AR1 denotes the first-order autoregressive model. LM denotes the linear model over market expectations. LA and AL denotes the LASSO and AdaLASSO models estimates using quadratic loss-function. LA-QR and AL-QR denotes the LASSO and AdaLASSO models estimates using quantile regression. All models starting with "E" denotes an ensemble between the model and market expectations from focus (e.g. E-LA is the ensemble with LASSO estimates; E-LA-QR is the ensemble with LASSO quantile regression). E-G denotes an ensemble of all shrinkage models estimated.

[b] All metrics have been normalized by market expectations from Focus (benchmark). Columns represent different data, while rows represent different specifications. The specification with the best result is presented in bold for each column.

Table 2.8: Diebold Mariano Test: p-values

Mean Squared Error - MSE						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
LA	0.005*	0.14	0.009*	0.005*	0.011*	0*
AL	0.18	0.40	0.19	0.16	0.76	0.028*
LA-QR	0.03*	0.10	0.001*	0.003*	0.001*	0*
AL-QR	0.29	0.38	0.18	0.021*	0.018*	0*
E-LA	0.019*	0*	0.007*	0.003*	0.06	0.35
E-AL	0.004*	0.001*	0.004*	0.001*	0*	0.038*
E-LA-QR	0.001*	0*	0.038*	0.006*	0.88	0.008*
E-AL-QR	0*	0*	0.002*	0.011*	0.06	0.035*
E-G	0*	0*	0.002*	0.001*	0.023*	0.60
Mean Absolute Error - MAE						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
LA	0.017*	0.48	0.01*	0.039*	0.003*	0*
AL	0.86	0.53	0.67	0.85	0.40	0.21
LA-QR	0.06	0.27	0*	0.031*	0*	0*
AL-QR	0.88	0.71	0.24	0.08	0.004*	0*
E-LA	0.05	0*	0.06	0.012*	0.006*	0.59
E-AL	0.009*	0*	0.008*	0.003*	0*	0.018*
E-LA-QR	0.002*	0*	0.10	0.027*	0.23	0.009*
E-AL-QR	0.002*	0.002*	0.01*	0.049*	0.011*	0.08
E-G	0.002*	0*	0.009*	0.004*	0.002*	0.92

[a] AR1 denotes the first-order autoregressive model. LM denotes the linear model over market expectations. LA and AL denotes the LASSO and AdaLASSO models estimates using quadratic loss-function. LA-QR and AL-QR denotes the LASSO and AdaLASSO models estimates using quantile regression. All models starting with "E" denotes an ensemble between the model and market expectations from focus (e.g. E-LA is the ensemble with LASSO estimates; E-LA-QR is the ensemble with LASSO quantile regression). E-G denotes an ensemble of all shrinkage models estimated.

[b] The table presents the p-values associated with a Diebold Mariano test, comparing forecasts of shrinkage and ensemble models with the market expectations from Focus (benchmark model).

[c] *denotes a p-value lower than 0.05.

Table 2.9: First Nowcast Evaluation: MSE and MAE (Normalized)

Mean Squared Error - MSE						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	19.1668	19.1668	19.1668	19.1668	19.1668	19.1668
LM	18.6660	18.6660	18.6660	18.6660	18.6660	18.6660
LA	1.7701	2.0522	1.7762	1.6849	1.4714	3.2276
AL	2.5489	1.5786	1.6332	1.8162	1.7895	2.0857
LA-QR	2.1801	2.1831	1.8546	1.9910	1.4473	3.2603
AL-QR	1.9330	2.1798	1.8293	1.8335	1.6230	3.1159
E-LA	**0.6872**	0.7189	0.7343	**0.6936**	**0.6953**	1.2686
E-AL	0.8376	**0.6440**	**0.6723**	0.7552	0.7476	**0.9474**
E-LA-QR	0.7480	0.7704	0.7646	0.8301	0.7114	1.2108
E-AL-QR	0.7124	0.7839	0.7551	0.8506	0.7232	1.1400
E-G	0.7289	0.7179	0.7254	0.7649	0.7016	1.1106
Mean Absolute Error - MAE						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	3.1521	3.1521	3.1521	3.1521	3.1521	3.1521
LM	2.9702	2.9702	2.9702	2.9702	2.9702	2.9702
LA	1.2313	1.2034	1.2494	1.2825	1.2676	1.8224
AL	1.2301	1.0031	1.1403	1.2914	1.1944	1.5357
LA-QR	1.2657	1.2929	1.3432	1.4300	1.2756	1.8638
AL-QR	1.1562	1.1816	1.3317	1.3719	1.2949	1.7845
E-LA	**0.8884**	0.8742	0.9100	**0.8468**	0.8935	1.1361
E-AL	0.9592	**0.8218**	**0.8636**	0.8833	0.9064	**1.0160**
E-LA-QR	0.9025	0.9117	0.9314	0.9738	**0.8836**	1.1334
E-AL-QR	0.8930	0.8997	0.9291	0.9427	0.8854	1.0962
E-G	0.9079	0.8689	0.9037	0.9088	0.8922	1.0863

[a] AR1 denotes the first-order autoregressive model. LM denotes the linear model over market expectations. LA and AL denotes the LASSO and AdaLASSO models estimates using quadratic loss-function. LA-QR and AL-QR denotes the LASSO and AdaLASSO models estimates using quantile regression. All models starting with "E" denotes an ensemble between the model and market expectations from focus (e.g. E-LA is the ensemble with LASSO estimates; E-LA-QR is the ensemble with LASSO quantile regression). E-G denotes an ensemble of all shrinkage models estimated.

[b] All metrics have been normalized by market expectations from Focus (benchmark). Columns represent different data, while rows represent different specifications. The specification with the best result is presented in bold for each column.

Table 2.10: Last Nowcast Evaluation: MSE and MAE (Normalized)

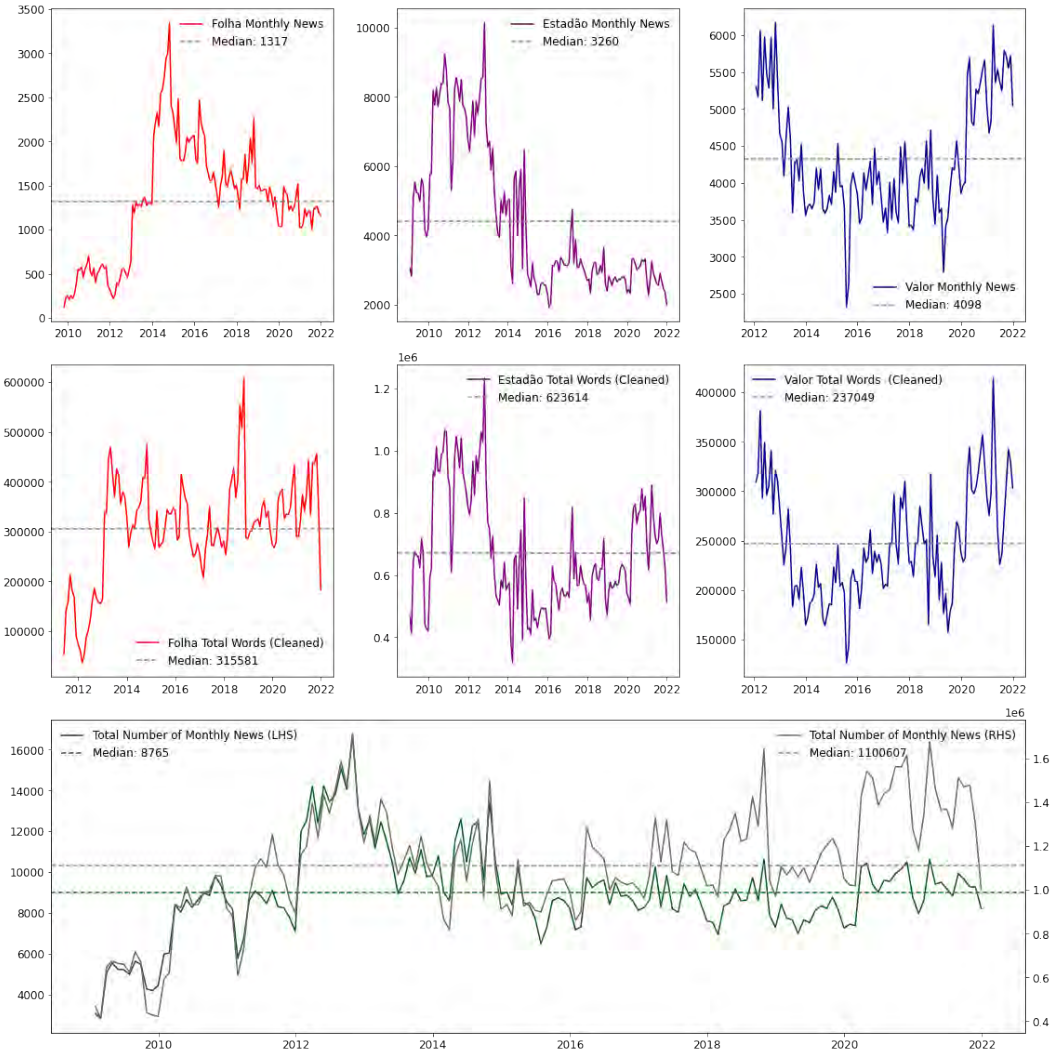
Mean Squared Error - MSE						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	151.0661	151.0661	151.0661	151.0661	151.0661	151.0661
LM	147.1192	147.1192	147.1192	147.1192	147.1192	147.1192
LA	2.7648	1.5543	1.9842	2.5388	3.3416	4.1380
AL	3.1031	1.3579	2.1232	2.5242	1.9607	2.1826
LA-QR	2.4220	2.1381	3.1589	2.3405	8.0465	8.3234
AL-QR	1.9025	1.8640	1.9515	1.8627	2.4911	6.4791
E-LA	1.4979	**1.0284**	**1.1956**	1.4372	1.5375	1.5493
E-AL	1.4799	1.0418	1.3322	1.5033	1.3808	**1.4015**
E-LA-QR	1.3031	1.2657	1.6398	1.3817	3.0182	2.5335
E-AL-QR	**1.2267**	1.2197	1.2504	**1.2632**	**1.3542**	2.0886
E-G	1.2413	1.0772	1.2638	1.3100	1.6357	1.5819
Mean Absolute Error - MAE						
	ALL	PCA/GT	PCA/ATT	PCA	GT	HARD
Focus	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AR1	8.1492	8.1492	8.1492	8.1492	8.1492	8.1492
LM	7.6789	7.6789	7.6789	7.6789	7.6789	7.6789
LA	1.4216	1.0578	1.4058	1.4026	1.5580	1.6971
AL	1.3571	1.0706	1.2154	1.3746	1.2505	1.2257
LA-QR	1.2795	1.3026	1.6646	1.3327	2.1122	2.2566
AL-QR	1.1845	1.0781	1.3917	1.2591	1.6039	2.0600
E-LA	1.1078	**0.8843**	1.1048	1.1221	1.0834	1.1863
E-AL	1.1075	0.9584	**1.0474**	1.1245	1.1045	**1.0396**
E-LA-QR	**0.9455**	0.9877	1.1688	1.1109	1.2676	1.3734
E-AL-QR	0.9999	1.0069	1.0936	1.1069	1.1063	1.2885
E-G	0.9790	0.9337	1.0514	**1.0953**	**1.0657**	1.1543

[a] AR1 denotes the first-order autoregressive model. LM denotes the linear model over market expectations. LA and AL denotes the LASSO and AdaLASSO models estimates using quadratic loss-function. LA-QR and AL-QR denotes the LASSO and AdaLASSO models estimates using quantile regression. All models starting with "E" denotes an ensemble between the model and market expectations from focus (e.g. E-LA is the ensemble with LASSO estimates; E-LA-QR is the ensemble with LASSO quantile regression). E-G denotes an ensemble of all shrinkage models estimated.

[b] All metrics have been normalized by market expectations from Focus (benchmark). Columns represent different data, while rows represent different specifications. The specification with the best result is presented in bold for each column.

Figures

Figure 2.1: Evolution of the Number of News and Words



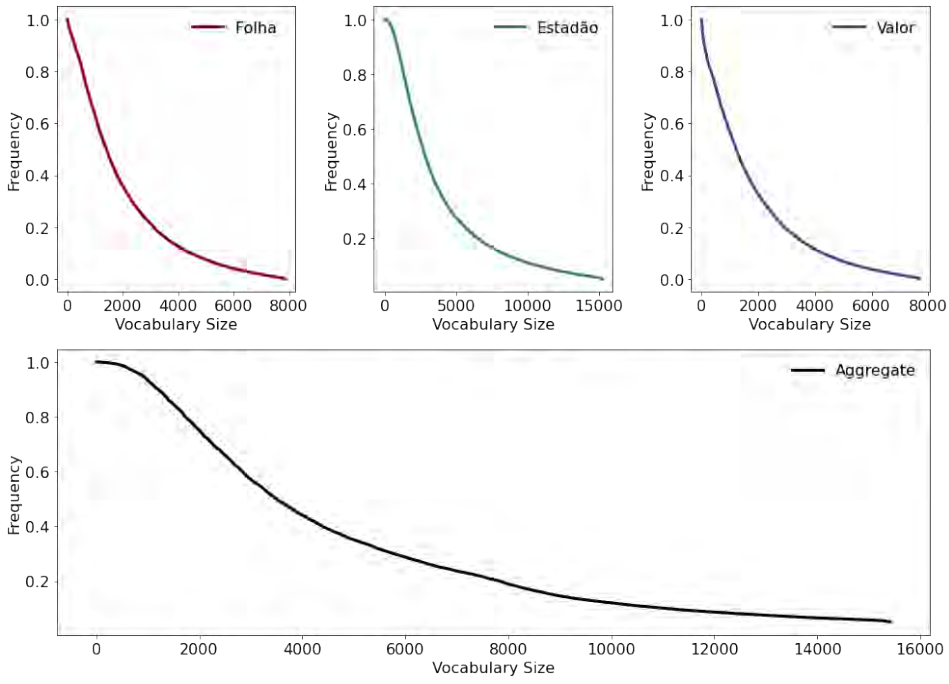
Evolution of the number of news and total (cleaned) words for January 1, 2009 up to December 31, 2021. Note how news and words evolve similar.

Figure 2.2: Example of Cleaning Algorithm over an Estadão News (Portuguese)

grafton , wisconsin - quinhentas doses da vacina contra a covid-19 tiveram que ser descartadas porque não foram devidamente refrigeradas em wisconsin , nos estados unidos . de acordo com a rede de hospitais aurora medical center , os imunizantes foram aparentemente estragadas de forma deliberada por um funcionário . no sábado , o hospital havia revelado que as doses foram acidentalmente deixadas em temperatura ambiente durante a noite por um funcionário da unidade de grafton . no entanto , nesta quarta-feira , 30 , o aurora divulgou uma nota afirmando que o funcionário envolvido `` reconheceu que as vacinas foram retiradas intencionalmente da geladeira '' . o comunicado diz ainda que o funcionário foi demitido e o assunto foi entregue às autoridades para uma investigação mais aprofundada . o depoimento não menciona o possível motivo dessa ação , e os executivos do sistema de saúde não responderam no momento às mensagens que lhes foram enviadas em busca de mais informações . " continuamos a acreditar que a vacinação é a nossa saída para a pandemia . estamos mais do que frustrados com o fato de o comportamento desse indivíduo atrasar a vacinação de mais de 500 pessoas " , disse a nota . o aurora medical center se recusou a fornecer informações adicionais , mas disse que daria mais detalhes na quinta-feira./ap

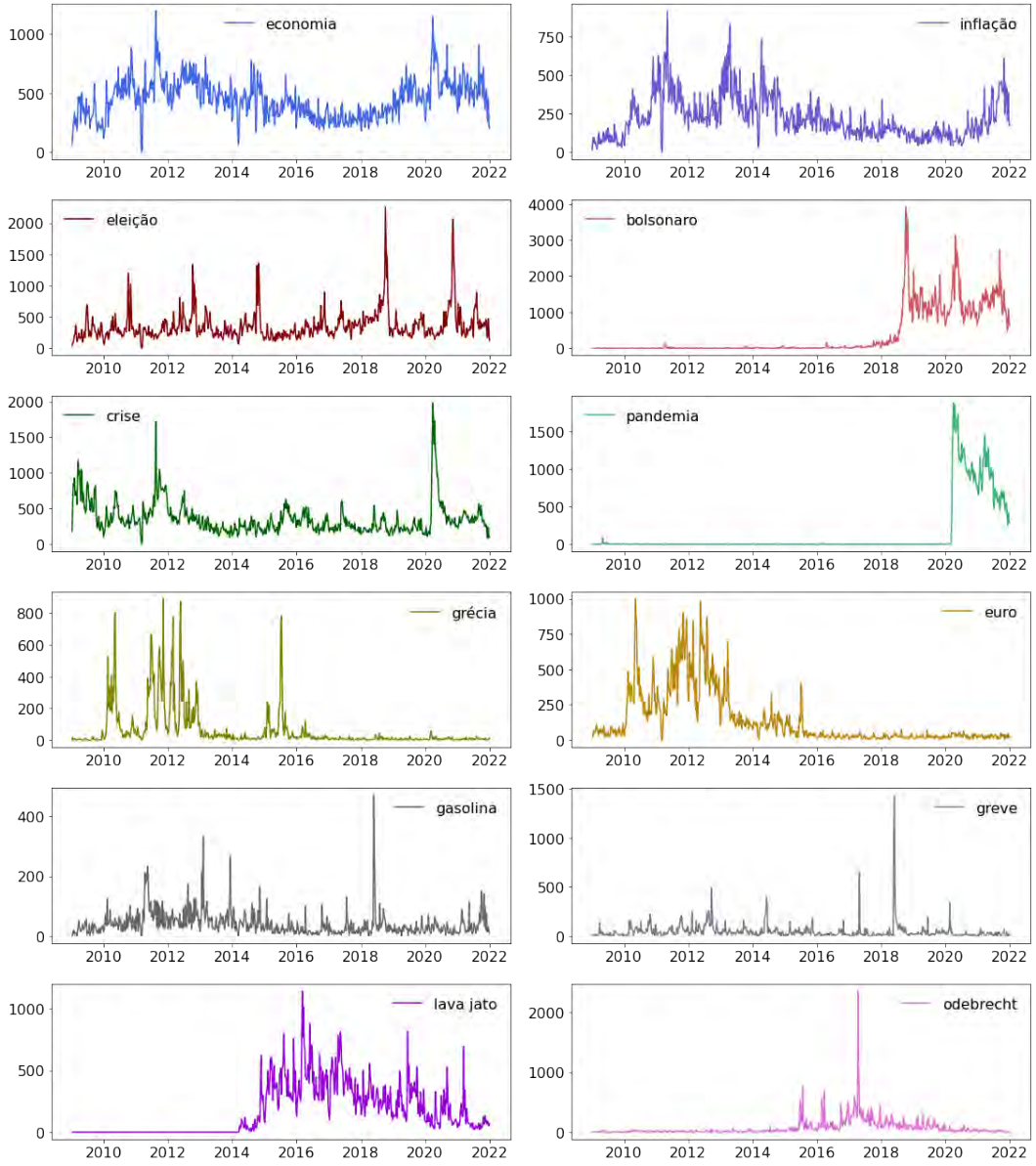
Cleaning and pre-processing of textual data from a particular article from Estadão. We mark the punctuation extraction in blue together with the number and single-letter removal. We highlight the stopwords, rare words, and temporal markers removed in red. In green, we point out the words that should be lemma-reduced. In black, we highlight words that should not be ex-ante modified.

Figure 2.3: Unique Terms Counting: Zipf's Law



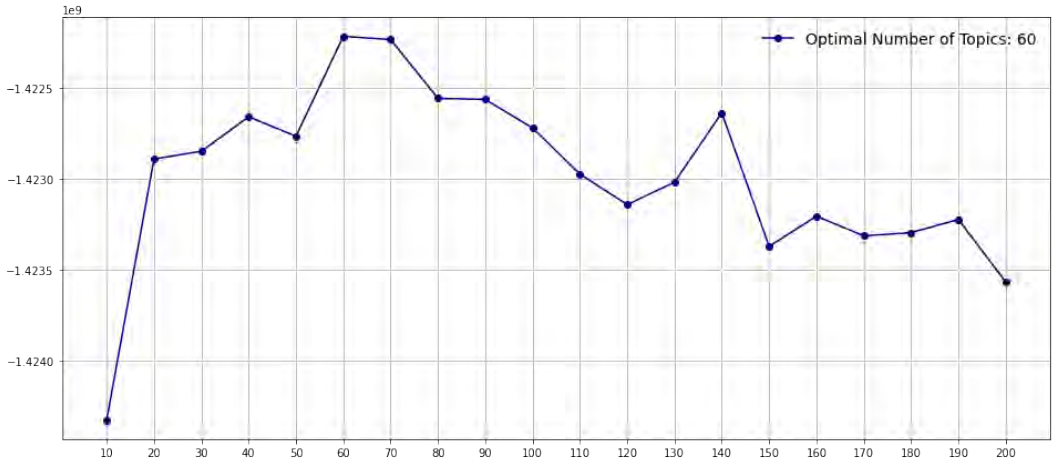
Zip's law reconstitution. The ordering of unique terms for their total appearances on the overall sample (normalized) shows the decay of counts of words. Note that the implied histogram reveals the highly right-skewed distribution of unique terms, i.e. many terms appear less than 20% on the overall sample size. Such observation provides a graphical representation of Zipf's Law: few terms occurs in almost all articles, while many terms occur in few articles.

Figure 2.4: Selected N-grams Evolution



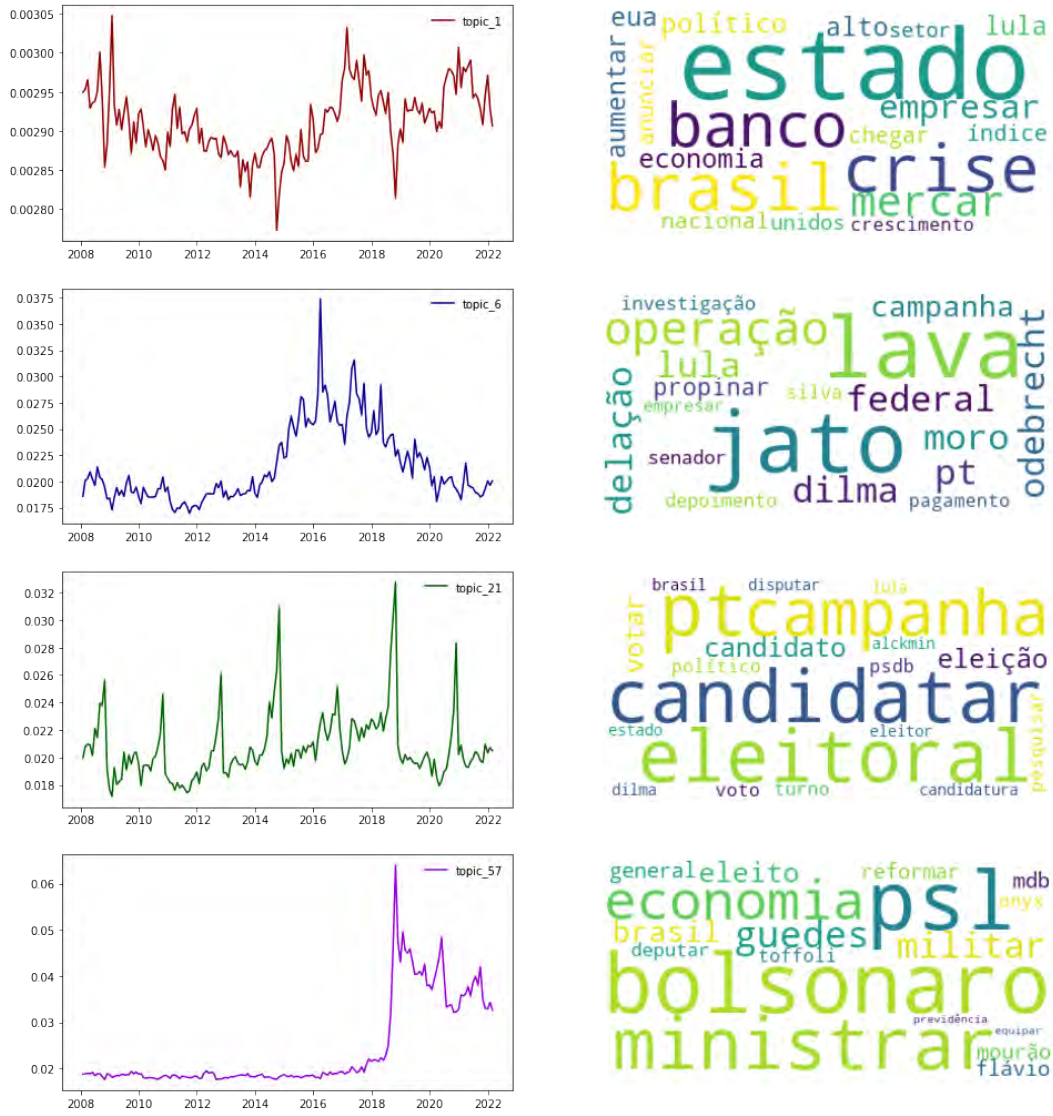
Specific examples for N-grams counting. Each line display similar terms for economy, politics and international affairs.

Figure 2.5: Data-Driven Optimal Number of Topics



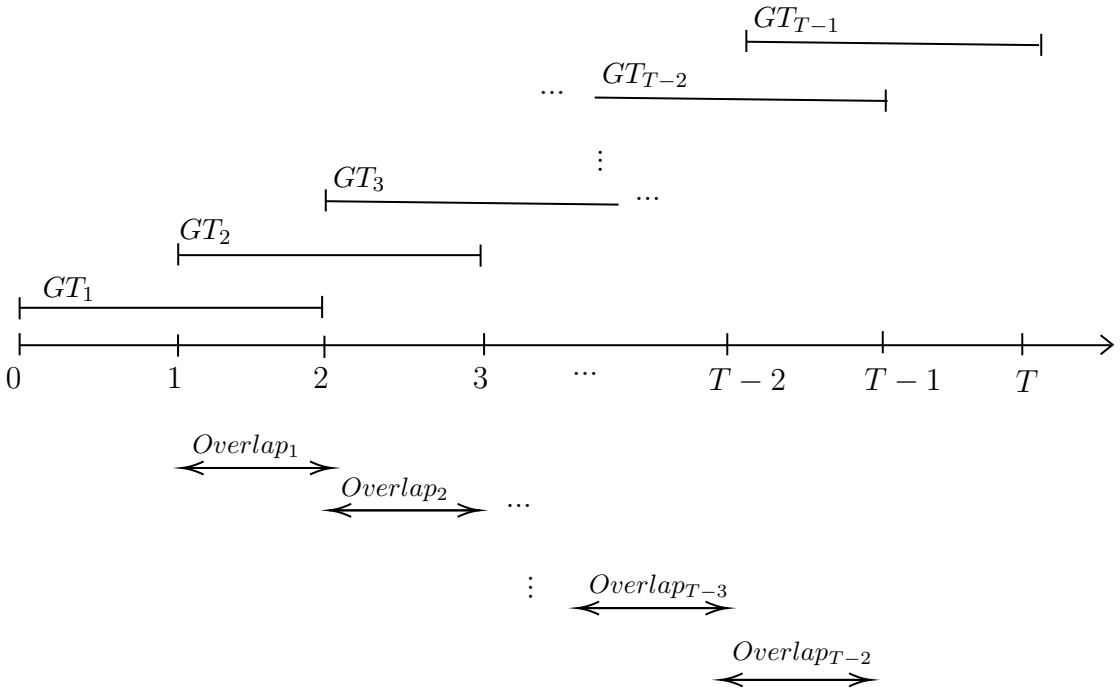
Optimal number of topics $K = 60$ that maximizes the score function. Note that the overall sample estimation may not consider the possibility of new vocabulary or topic of attention; however, it can be helpful to understand the overall media attention during the in-sample period. On the other side, we expect media attention groups to be relatively stable over time, which does not generate the need to keep reestimating the attention groups.

Figure 2.6: Selected Media Attention Evolution



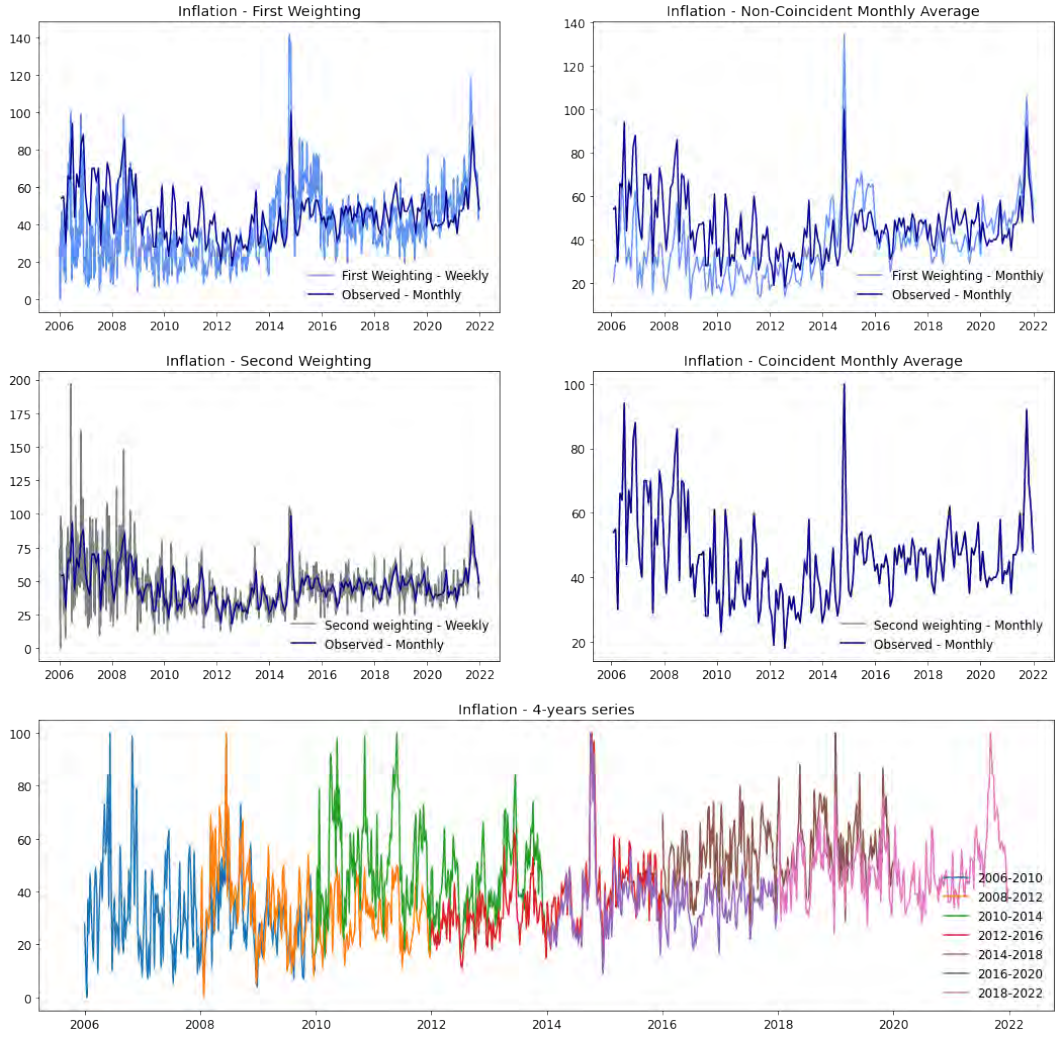
Evolution for selected media attention groups (LHS plot) together with its twenty most common terms (RHS word-cloud). Note that based on the word clouds, one can proceed with a manual inspection to check the coherence of media attention groups (e.g. Bolsonaro group only surges by 2018, as it was not well known on the media prior to the presidential election).

Figure 2.7: Google Trends Transformation: Weekly Series



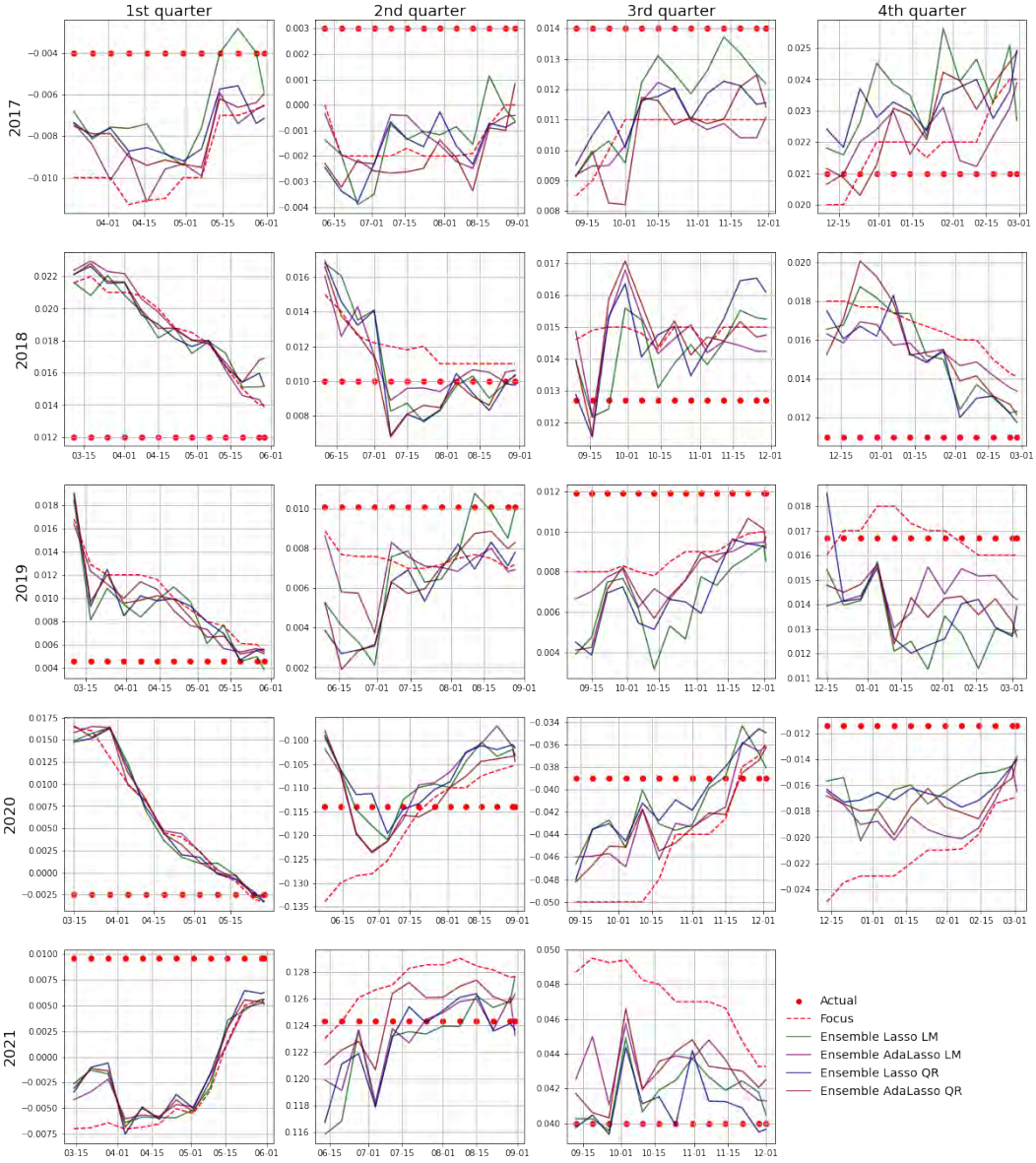
Scheme representing the enchainment idea used to reconstitute longer horizons weekly Google Trends series. The idea is to use short weekly series (about four years of observations) with overlaps (about two years of observations). The overlap is used to adjust the series to generate a more extended time series. However, simply enchaining the series does not ensure that the series behaviour coincides with the monthly observations. We thus use the latter as an adjustment weighting to replicate weekly observations for the observed monthly series.

Figure 2.8: Google Trends Transformation: Inflation Example



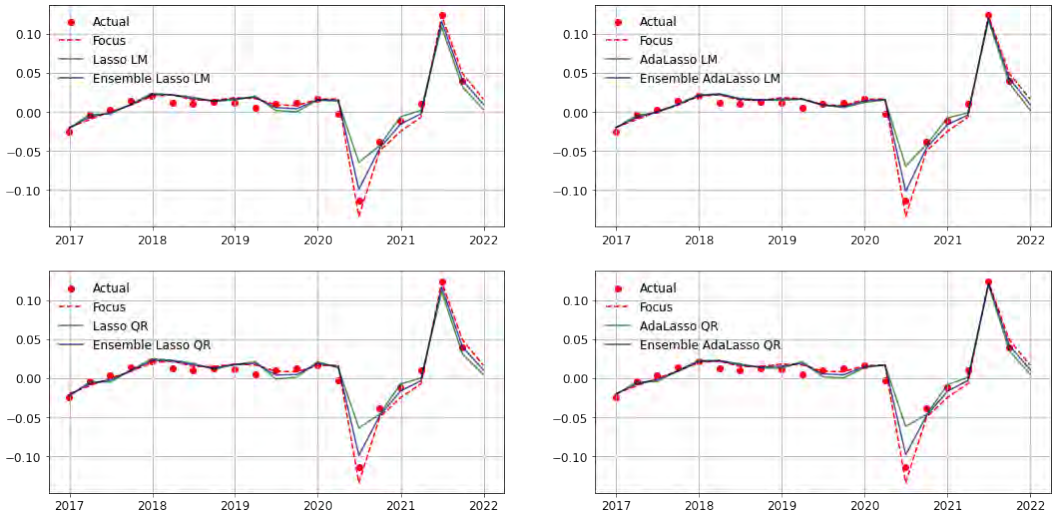
Reconstitution of weekly Google Trends series for "Inflation" search-term. Note that some distortions are still evident after the first weighting procedure, as the monthly average of the weekly series does not coincide with the monthly observed series. However, after the second weighting, the weekly series becomes centred over the monthly observed series, preserving its inter-monthly behaviour.

Figure 2.9: Nowcasting Results for PCA/GT: Quarter-by-Quarter



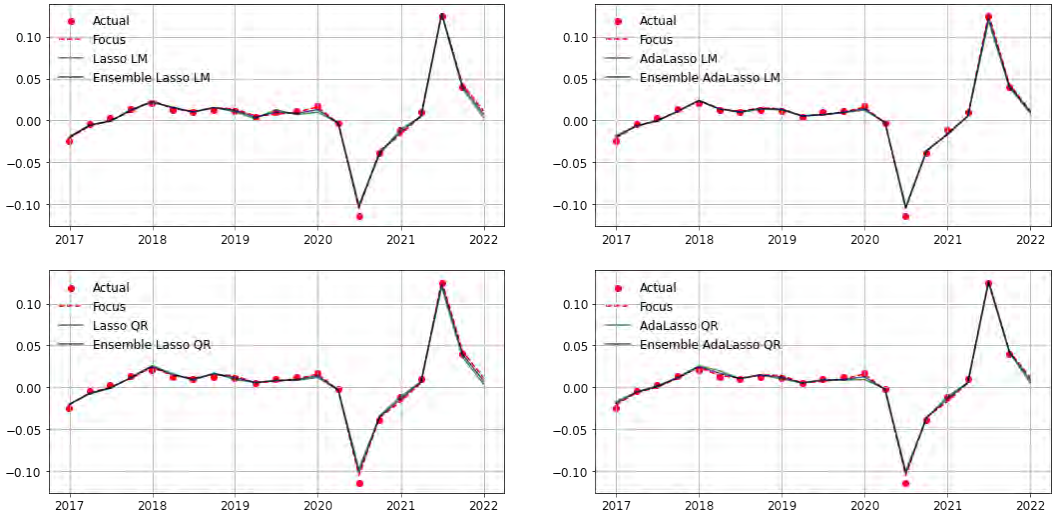
Nowcasting term-structure for each quarter starting in 2017Q1 to 2021Q3 for the best-performers ensemble specification. Note that years are presented over the rows, while quarters are presented over the columns.

Figure 2.10: Nowcasting Results for PCA/GT: First Week Prediction



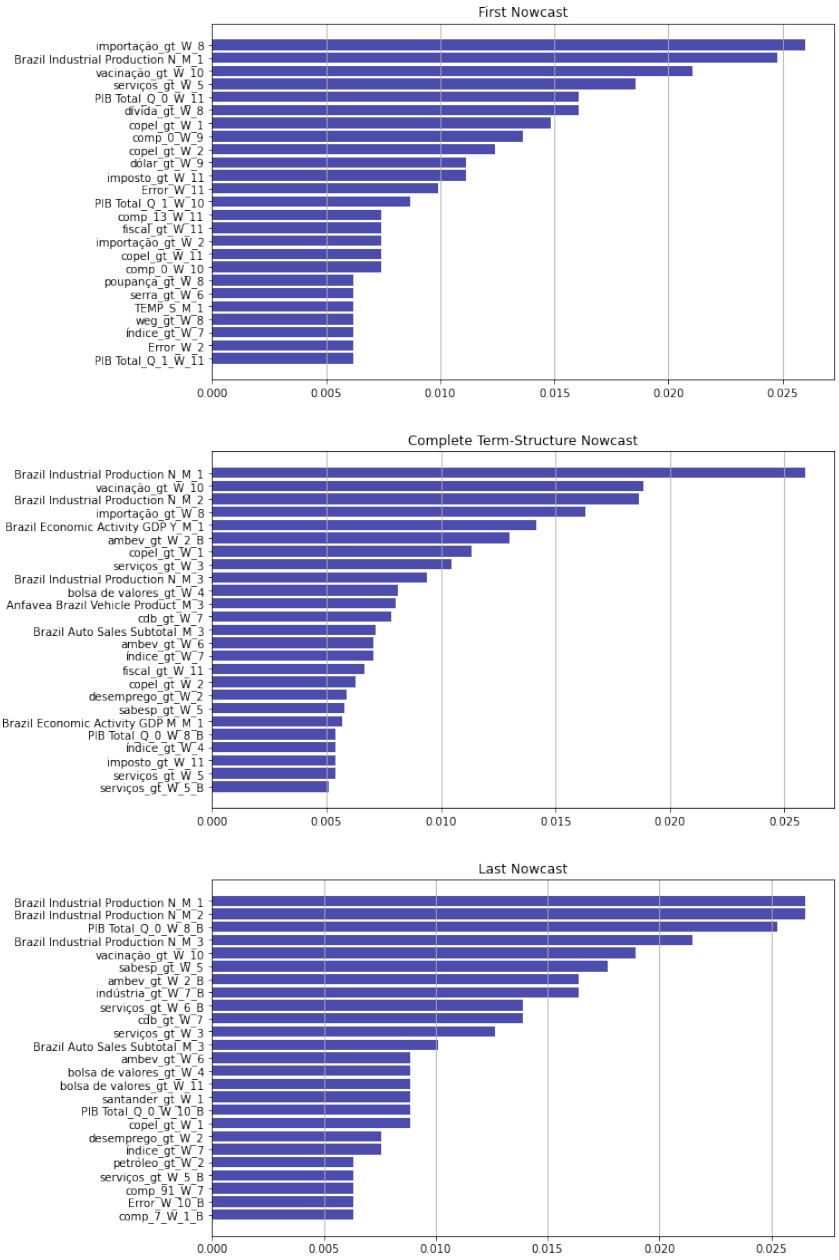
Nowcasting zoom to the first week of the nowcast through all quarters from 2017Q1 to 2021Q3 for both shrinkage and ensemble specifications.

Figure 2.11: Nowcasting Results for PCA/GT: Last Week Prediction



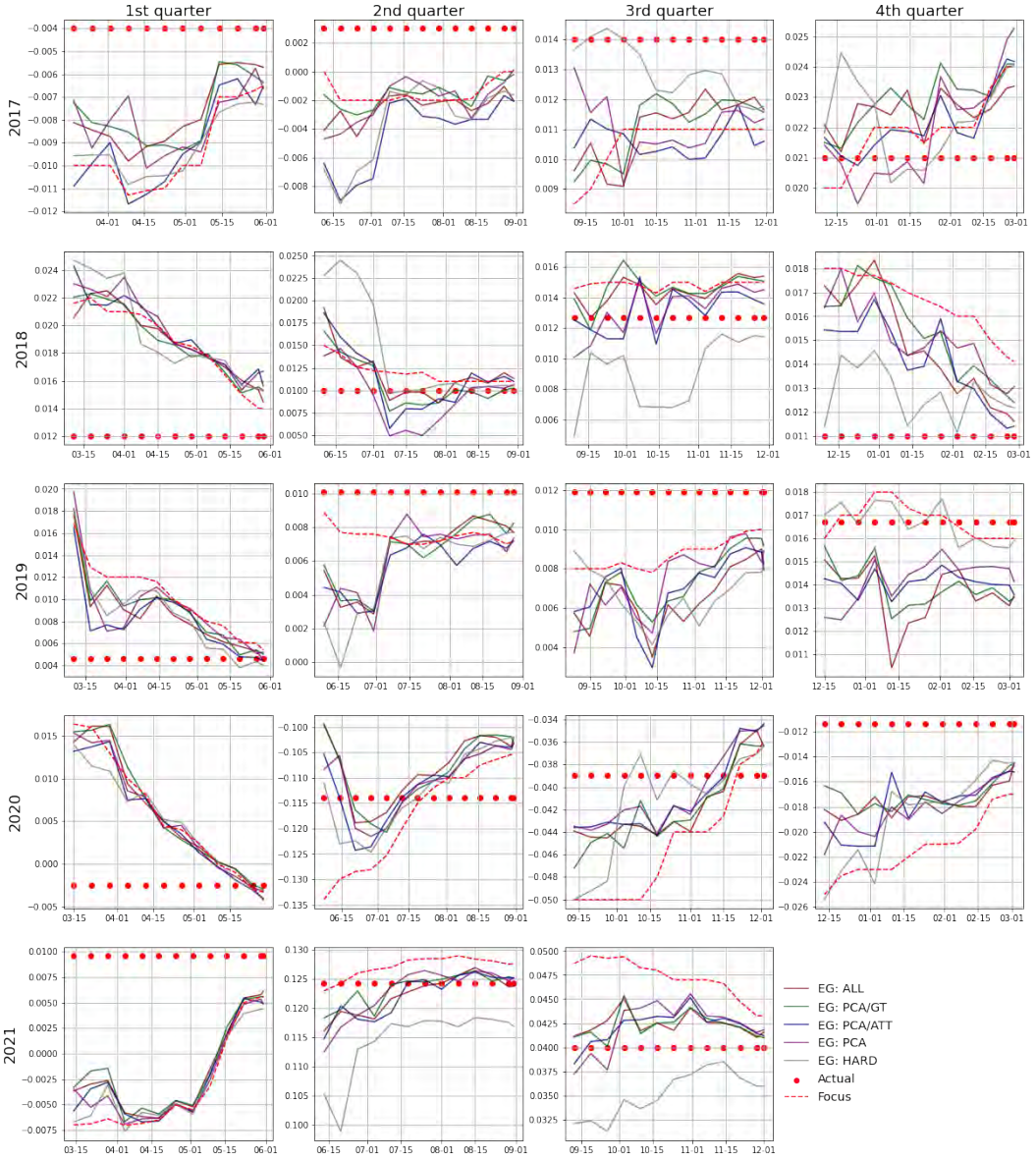
Nowcasting zoom to the last week of the nowcast through all quarters from 2017Q1 to 2021Q3 for both shrinkage and ensemble specifications.

Figure 2.12: Nowcasting Results for PCA/GT: Variables Selected in LASSO QR



Top 25 most selected variables from the LASSO QR regression based on the period covering 2017Q1 to 2021Q3. Note that "W" denotes a weekly variable, while "M" denotes a monthly variable. Also, "B" denotes a backcasting variable. Finally, the number after "W" or "M" denotes the number of weeks passed after the first released variable, e.g. "W_8" means that eight weeks have passed since the first number covering the current quarter.

Figure 2.13: Nowcasting Results General Model: Quarter-by-Quarter



Nowcasting term structure for each quarter starting in 2017Q1 to 2021Q3 for the general ensemble specification for all data combinations. Note that years are presented over the rows, while quarters are presented over the columns.

Bibliography

ARNAL, R. P.; CONESA, D.; ALVAREZ-NAPAGAO, S.; SUZUMURA, T.; CATALÀ, M.; ALVAREZ, E. ; GARCIA-GASULLA, D.. **Private sources of mobility data under covid-19**. arXiv preprint arXiv:2007.07095, 2020.

BAI, J.; GHYSELS, E. ; WRIGHT, J. H.. **State space models and midas regressions**. Econometric Reviews, 32(7):779–813, 2013.

BAKER, S. R.; BLOOM, N. ; DAVIS, S. J.. **Measuring economic policy uncertainty**. The quarterly journal of economics, 131(4):1593–1636, 2016.

BARBOZA, L. A.; VÁSQUEZ, P.; MERY, G.; SANCHEZ, F.; GARCÍA, Y. E.; CALVO, J. G.; RIVAS, T. ; SALAS, D.. **The role of mobility and sanitary measures on covid-19 in costa rica, march through july 2020**. arXiv preprint arXiv:2103.08732, 2021.

BATTY, M.; MURCIO, R.; IACOPINI, I.; VANHOOF, M. ; MILTON, R.. **London in lockdown: Mobility in the pandemic city**. In: COVID-19 PANDEMIC, GEOSPATIAL INFORMATION, AND COMMUNITY RESILIENCE, p. 229–244. CRC Press, 2021.

BAYAT, N.; MORRIN, C.; WANG, Y. ; MISRA, V.. **Synthetic control, synthetic interventions, and covid-19 spread: Exploring the impact of lockdown measures and herd immunity**. arXiv preprint arXiv:2009.09987, 2020.

BAŃBURA, M.; GIANNONE, D. ; REICHLIN, L.. **Nowcasting**. 2010.

BAŃBURA, M.; RÜNSTLER, G.. **A look into the factor model black box: publication lags and the role of hard and soft data in forecasting gdp**. International Journal of Forecasting, 27(2):333–346, 2011.

BENÍTEZ, M. A.; VELASCO, C.; SEQUEIRA, A. R.; HENRÍQUEZ, J.; MENEZES, F. M. ; PAOLUCCI, F.. **Responses to covid-19 in five latin american countries**. Health policy and technology, 9(4):525–559, 2020.

BIRD, S.; KLEIN, E. ; LOPER, E.. **Natural language processing with Python: analyzing text with the natural language toolkit**. " O'Reilly Media, Inc.", 2009.

BLEI, D. M.; NG, A. Y. ; JORDAN, M. I.. **Latent dirichlet allocation**. Journal of machine Learning research, 3(Jan):993–1022, 2003.

BOK, B.; CARATELLI, D.; GIANNONE, D.; SBORDONE, A. M. ; TAMBALOTTI, A.. **Macroeconomic nowcasting and forecasting with big data**. Annual Review of Economics, 10:615–643, 2018.

BORN, B.; DIETRICH, A. M. ; MÜLLER, G. J.. **The lockdown effect: A counterfactual for sweden**. Plos one, 16(4):e0249732, 2021.

BORUP, D.; RAPACH, D. ; SCHÜTTE, E. C. M.. **Mixed-frequency machine learning: Now-and backcasting weekly initial claims with daily internet search-volume data**. Available at SSRN 3690832, 2021.

BYBEE, L.; KELLY, B. T.; MANELA, A. ; XIU, D.. **The structure of economic news**. Technical report, National Bureau of Economic Research, 2020.

CARNEIRO, C. B.; FERREIRA, I. H.; MEDEIROS, M. C.; PIRES, H. F. ; ZILBERMAN, E.. **Lockdown effects in us states: an artificial counterfactual approach**. arXiv preprint arXiv:2009.13484, 2020.

CARRIÈRE-SWALLOW, Y.; LABBÉ, F.. **Nowcasting with google trends in an emerging market**. Journal of Forecasting, 32(4):289–298, 2013.

CHAGAS, E. T.; BARROS, P. H.; CARDOSO-PEREIRA, I.; PONTE, I. V.; XIMENES, P.; FIGUEIREDO, F.; MURAI, F.; COUTO DA SILVA, A. P.; ALMEIDA, J. M.; LOUREIRO, A. A. ; OTHERS. **Effects of population mobility on the covid-19 spread in brazil**. PloS one, 16(12):e0260610, 2021.

CHOI, H.; VARIAN, H.. **Predicting initial claims for unemployment benefits**. Google Inc, 1:1–5, 2009.

CHOI, H.; VARIAN, H.. **Predicting the present with google trends**. Economic record, 88:2–9, 2012.

CINTRA, H.; FONTINELE, F.. **Estimative of real number of infections by covid-19 in brazil and possible scenarios**. Infectious Disease Modelling, 5:720–736, 2020.

DEB, P.; FURCERI, D.; OSTRY, J. D. ; TAWK, N.. **The economic effects of covid-19 containment measures**. 2020.

DIEBOLD, F. X.; MARIANO, R. S.. **Comparing predictive accuracy**. Journal of Business & economic statistics, 20(1):134–144, 2002.

DRIKVANDI, R.; LAWAL, O.. **Sparse principal component analysis for natural language processing**. Annals of data science, p. 1–17, 2020.

EISENSTEIN, J.. **Natural language processing**, 2018.

ELLINGSEN, J.; LARSEN, V. H. ; THORSRUD, L. A.. **News media vs. fred-md for macroeconomic forecasting**. 2020.

ELWERT, F.. **Graphical causal models**. In: HANDBOOK OF CAUSAL ANALYSIS FOR SOCIAL RESEARCH, p. 245–273. Springer, 2013.

ENGLE, S.; STROMME, J. ; ZHOU, A.. **Staying at home: mobility effects of covid-19**. Available at SSRN 3565703, 2020.

EVANS, M.. **Where are we now? real-time estimates of the macro economy**, 2005.

FORONI, C.; MARCELLINO, M.. **A comparison of mixed frequency approaches for nowcasting euro area macroeconomic aggregates**. International Journal of Forecasting, 30(3):554–568, 2014.

FORONI, C.; MARCELLINO, M. G.. **A survey of econometric methods for mixed-frequency data**. Available at SSRN 2268912, 2013.

FORONI, C.; MARCELLINO, M. ; SCHUMACHER, C.. **Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials**. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(1):57–82, 2015.

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, 33(3):679–693, 2017.

GAUTHIER, G.. **On the use of two-way fixed effects models for policy evaluation during pandemics**. arXiv preprint arXiv:2106.10949, 2021.

GENTZKOW, M.; KELLY, B. ; TADDY, M.. **Text as data**. Journal of Economic Literature, 57(3):535–74, 2019.

GERLEE, P.; KARLSSON, J.; FRITZELL, I.; BREZICKA, T.; SPRECO, A.; TIMPKA, T.; JÖUD, A. ; LUNDH, T.. **Predicting regional covid-19 hospital admissions in sweden using mobility data**. arXiv preprint arXiv:2101.00823, 2021.

GIANNONE, D.; REICHLIN, L. ; SMALL, D.. **Nowcasting: The real-time informational content of macroeconomic data**. Journal of Monetary Economics, 55(4):665–676, 2008.

GOLDSTEIN, P.; YEYATI, E. L. ; SARTORIO, L.. **Lockdown fatigue: The diminishing effects of quarantines on the spread of covid-19**. 2021.

HANSEN, S.; MCMAHON, M.. **Shocking language: Understanding the macroeconomic effects of central bank communication**. Journal of International Economics, 99:S114–S133, 2016.

HANSEN, S.; MCMAHON, M. ; PRAT, A.. **Transparency and deliberation within the fomc: a computational linguistics approach**. The Quarterly Journal of Economics, 133(2):801–870, 2018.

HONNIBAL, M.; MONTANI, I.; VAN LANDEGHEM, S. ; BOYD, A.. **spaCy: Industrial-strength Natural Language Processing in Python**, 2020.

HUANG, D.. **How effective is social distancing?** Available at SSRN 3680321, 2020.

JANIAK, A.; MACHADO, C. ; TURÉN, J.. **Covid-19 contagion, economic activity and business reopening protocols**. Journal of economic behavior & organization, 182:264–284, 2021.

JANSEN, W. J.; JIN, X. ; DE WINTER, J. M.. **Forecasting and nowcasting real gdp: Comparing statistical models and subjective forecasts**. International Journal of Forecasting, 32(2):411–436, 2016.

KE, Z. T.; KELLY, B. T. ; XIU, D.. **Predicting returns with text data**. Technical report, National Bureau of Economic Research, 2019.

KELLY, B. T.; MANELA, A. ; MOREIRA, A.. **Text selection**. Technical report, National Bureau of Economic Research, 2019.

KONG, E.; PRINZ, D.. **Disentangling policy effects using proxy data: Which shutdown policies affected unemployment during the covid-19 pandemic?** Journal of Public Economics, 189:104257, 2020.

KONTOANGELOS, K.; ECONOMOU, M. ; PAPAGEORGIOU, C.. **Mental health effects of covid-19 pandemia: a review of clinical and psychological traits**. Psychiatry investigation, 17(6):491, 2020.

LIU, L.; MOON, H. R. ; SCHORFHEIDE, F.. **Panel forecasts of country-level covid-19 infections**. Journal of econometrics, 220(1):2–22, 2021.

LOVELL, M. C.. **Tests of the rational expectations hypothesis**. The American Economic Review, 76(1):110–124, 1986.

MANELA, A.; MOREIRA, A.. **News implied volatility and disaster concerns**. Journal of Financial Economics, 123(1):137–162, 2017.

MCCRACKEN, M. W.; NG, S.. **Fred-md: A monthly database for macroeconomic research**. Journal of Business & Economic Statistics, 34(4):574–589, 2016.

MCCRACKEN, M.; NG, S.. **Fred-qd: A quarterly database for macroeconomic research**. Technical report, National Bureau of Economic Research, 2020.

MEDEIROS, M.; STREET, A.; VALLADÃO, D.; VASCONCELOS, G. ; ZILBERMAN, E.. **Short-term covid-19 forecast for latecomers**. arXiv preprint arXiv:2004.07977, 2020.

MEINSHAUSEN, N.; BUHLMANN, P.. **Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso**. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich, 2004.

NOUVELLET, P.; BHATIA, S.; CORI, A.; AINSLIE, K. E.; BAGUELIN, M.; BHATT, S.; BOONYASIRI, A.; BRAZEAU, N. F.; CATTARINO, L.; COOPER, L. V. ; OTHERS. **Reduction in mobility and covid-19 transmission**. Nature communications, 12(1):1–9, 2021.

PEARL, J.. **Causal diagrams for empirical research**. Biometrika, 82(4):669–688, 1995.

PEARL, J.. **Causality**. Cambridge university press, 2009.

RESENDE, M.; MACIEL, M.. **Social distancing and covid-19: Some evidence at the municipality level in brazil**. Available at SSRN 3881417, 2021.

RÜNSTLER, G.; BAŃBURA, M. ; OTHERS. **A look into the factor model black box: publication lags and the role of hard and soft data in forecasting gdp**. Technical report, 2007.

SAIZ, L.; ASHWIN, J.; KALAMARA, E. ; OTHERS. **Nowcasting euro area gdp with news sentiment: a tale of two crises**. Technical report, 2021.

SCHERBINA, A.. **Could the united states benefit from a lockdown? a cost-benefit analysis**. A Cost-Benefit Analysis (January 12, 2021), 2021.

SCHWABE, A.; PERSSON, J. ; FEUERRIEGEL, S.. **Predicting covid-19 spread from large-scale mobility data**. arXiv preprint arXiv:2106.00356, 2021.

SMITH, P.. **Google's midas touch: Predicting uk unemployment with internet search data**. Journal of Forecasting, 35(3):263–284, 2016.

STANNARD, T.; STEVEN, G.; MCDONALD, C. ; OTHERS. **Economic impacts of covid-19 containment measures**. Technical report, Reserve Bank of New Zealand Wellington, 2020.

THORSRUD, L. A.. **Nowcasting using news topics. big data versus big bank**. 2016.

TIBSHIRANI, R.. **Regression shrinkage and selection via the lasso**. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

VESPE, M.; MINORA, U.; IACUS, S. M.; SPYRATOS, S.; SERMI, F.; FONTANA, M.; CIUFFO, B. ; CHRISTIDIS, P.. **Mobility and economic impact of covid-19 restrictions in italy using mobile network operator data**. arXiv preprint arXiv:2106.00460, 2021.

WOLOSZKO, N.. **Tracking activity in real time with google trends**. 2020.

ZHAO, P.; YU, B.. **On model selection consistency of lasso**. The Journal of Machine Learning Research, 7:2541–2563, 2006.

ZOU, H.. **The adaptive lasso and its oracle properties**. Journal of the American statistical association, 101(476):1418–1429, 2006.

ZOU, H.; HASTIE, T. ; TIBSHIRANI, R.. **On the "degrees of freedom" of the lasso**. The Annals of Statistics, 35(5):2173–2192, 2007.

A
Appendix to Chapter 1

Appendix A.1 Keywords and Categories

To generate our soft data controls, we need to specify the Google Trends search terms chosen by individuals or the keywords used to constitute the News-Index, i.e., we need to specify the dictionary of terms. The objective is to be concise and precise while selecting the terms in order to capture a general overview of the Covid-19 situation at the individual level that may affect both mobility and number of cases and deaths, according to the DAG proposed in Figures 1.5 and 1.6. We also focus on generating compatibility within search terms and keywords for both types of controls.

There are two main categories for Google Trends and News that encode general effects (g) and behavioral effects (b). To capture the general effects (g) of the Covid-19 pandemic, we selected words that refer to Covid-19 evolution in general. This general effects topic has been chosen to estimate the number of infections and deaths over time. To capture the effects of the behavioral channel (b), we selected: (i) Prevention related topics; (ii) Fake-news about the pandemic; (ii) Vaccination evolution. The first category should embed individual behavior (not observed by hard data indicators) that may affect the infection evolution throughout time. The second topic has the objective of allowing the direct impact of fake-news dissemination over Covid-19 spread. The third topic is the vaccination campaign to control the infection evolution correctly. We provide all terms used for the Google Trends series constitution (GT-series) and the News-index generation (N-index) in detail below:

1. General Covid-19 terms: covid, pandemia, coronavirus, covid-19, mortes covid, morrer de covid, covid o que fazer, covid como proceder, pegar covid, transmissão covid, covid mata, covid contagioso, covid transmite, contagio covid, sintomas covid, morte de covid, casos covid.

2. Fake News related terms: kit-covid, hidroxicloroquina, cloroquina, azitromicina, gripezinha, ivermectina, remedio covid, tratamento covid.

3. Vaccination related terms: vacinação covid, vacinas covid, pfizer, astrazeneca, janssen, butantan, coronavac, moderna, biontech, oxford, fiocruz, sputnik v.

4. Prevention related terms: mascara, lavas as mãos, alcool em gel, isolamento, distanciamento, quarentena, lockdown, confinamento, ficar em casa, toque de recolher, toque de restrição, restrições, circulação.

As described in Section 1.3, as behavior is a non-observable variable, the absence of such terms as proxies would bias the effects of mobility on cases (deaths). The usage of general Covid-19 related terms has the objective of capturing potential inertial effects via lags of the dependent variable. As we opt to estimate a fixed-effects model, including lags would violate the strict exogeneity condition. However, these general terms act as proxies for such lagged effects and therefore are helpful while controlling for inertial effects.

In collecting Google searches, the Google Trends API for R supplies the evolution (in percentage) for a specified period for each search given term. To create the News-Index, we recur to a similar approach as in Baker et al. (2016), collecting G1 news regarding Covid-19 and counting the topic appearances based on the presence or absence of the term in the title of the news. This way, we solely focus on news that focuses on the given topic directly[1].

Formally, as the sets of search words and keywords used to create GT-series and N-index are equal, define the indicator $g \in \{\boldsymbol{GT}, \boldsymbol{N}\}$. Recall that we have a panel data (at state level and weekly frequency) of each topic. Therefore, each observation is indexed by a state marker $s \in \{1, \dots S\}$ and a time stamp $t \in \{t_0, \cdots, T\}$. For each gt and n-series of controls, we have a vector $\boldsymbol{g_{(s,t)}} = (g_{1,(s,t)}, g_{2,(s,t)}, g_{3,(s,t)}, g_{4,(s,t)})$ that represents the four above categories. For each category $g_{i,(s,t)}$, for $i \in \{1, 2, 3, 4\}$, there are n_i associated search words time-series, denoted $\boldsymbol{g_{i,w}}$, that are also observed weekly and at state-level, i.e. $g_{i,w,(s,t)}$. We thus generate the indexes in the following manner:

$$g_{i,(s,t)} = \sum_{w=1}^{n_i} g_{i,w,(s,t)} \tag{A-1}$$

for each $i \in \{1, 2, 3, 4\}$, $s \in \{1, \dots S\}$ and $t \in \{t_0, \cdots, T\}$.

As an example, consider News regarding solely prevention related terms ($i = 4$), i.e. $N_{4,(s,t)}$. In this case, we have $n_4 = 13$ and w denotes a certain topic (e.g. $w =$ mascara). Therefore, the $N_{4,\text{mascara},(s,t)}$ denotes the number of counts of "mascara", that belongs to prevention related terms, over each state j and each week t. Finally, the News-Index for prevention related terms is given by:

[1]Differently, we could have counted the direct and indirect appearances of a given topic based on its inclusion on overall text elements. However, by focusing on its appearance on the title, we restrict the number of news to consider only directly referring topics.

$$N_{4,(s,t)} = \sum_{w=1}^{13} N_{4,w,(s,t)}$$

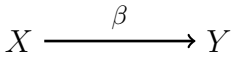
Appendix A.2 Introduction to Direct Acyclic Graphs

The usage of DAGs in causal inference have roots in Pearl (1995) and Pearl (2009), providing causal interpretation based on variables relations. Based on our estimation, the principal objective is to identify the causal effect of mobility on Covid-19 cases and deaths. We aim to isolate causal from noncausal associations. As described by Elwert (2013), DAGs are a powerful tool to identify what control variables we should include and which we should not include to "achieve identification".

The main idea of using DAGs consists in generating a clear and objective graph that should encode the central causal relations that we aim to describe in our model. At a glance, we should interpret a DAG-based on three elements, following Elwert (2013): (i) Variables, that are represented in nodes; (ii) Arrows, suggesting possible direct causal impacts; (iii) Missing arrows, encoding "sharp assumptions" about the absence of causality effects.

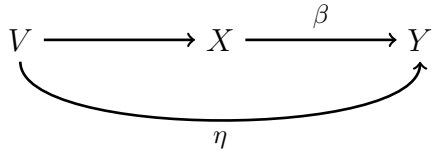
Ideally, if we were able to observe mobility in a way that it is not affected by external elements, i.e. exogenously (e.g. a randomized version of mobility), we would retrieve the causal effect of mobility on Covid-19 variables in a direct way. This representation would imply in the Figure A.1 Graph:

Figure A.1: DAG Representing Causal Chain Between Mobility and Cases (Deaths)



As we do not observe this "artificial" measure of mobility, we cannot retrieve a causal effect merely by regressing those two variables solely. Therefore, we should include potential control variables related to mobility measures to remove the omission bias. For example, take the vaccination as a potential control: this variable affects mobility and the number of cases and deaths. Therefore, its inclusion as a regressor (estimating η) is necessary to identify the causal effect of mobility correctly, i.e., without vaccination; we would bias the estimation of β through the channel of the η relation. This result on the Figure A.2 DAG:

Figure A.2: DAG Representing Causal Chain Between Vaccination, Mobility and Cases (Deaths)



By proceeding in the same fashion, we identify potential helpful controls in our identification strategies, such as internet searches and news related to Covid-19. The result is the complete DAG that we present in Section 1.3.