n. 31

Texto para Discussão

Fundação João Pinheiro

MODEL-BASED SINGLE-MONTH UNEMPLOYMENT ESTIMATES FROM THE BRAZILIAN LABOUR FORCE SURVEY INCORPORATING GOOGLE TRENDS DATA

César Soares Gonçalves
Luna Hidalgo
Denise Britz do Nascimento Silva
Jan Van den Brakel
Paulo Roberto Betbeder Gonçalves

Belo Horizonte
2025

FAPEMIG

FJP FUNDAÇÃO JOÃO PINHEIRO

GOVERNO DE MINAS
AQUI O TREM PROSPERA.

MINAS GERAIS
Governador
Romeu Zema Neto
Vice-Governador
Mateus Simões

SECRETARIA DE ESTADO DE PLANEJAMENTO E GESTÃO
Secretária
Silvia Caroline Listgarten Dias

FUNDAÇÃO JOÃO PINHEIRO
Presidente
Luciana Lopes Nominato Braga
Vice-Presidente
Mônica Moreira Esteves Bernardi

FICHA TÉCNICA

Elaboração
Caio César Soares Gonçalves
Luna Hidalgo
Denise Britz do Nascimento Silva
Jan Van den Brakel
Paulo Roberto Betbeder Gonçalves

Normalização
Graziella Napoli da Terra Caldeira

Capa
Aline de Faria Pereira

TEXTO PARA DISCUSSÃO

São textos que visam divulgar trabalhos preliminares. Possuem o objetivo de compartilhar ideias e obter comentários, críticas e sugestões.

FUNDACÃO JOÃO PINHEIRO
Alameda das Acácias, 70
Bairro São Luiz - Pampulha
Belo Horizonte - Minas Gerais
CEP 31.275-150
Telefones: (31) 3448.9711
www.fjp.mg.gov,br
E-mail: comunicacao@fjp.mg.gov.br.

Gonçalves, Caio César Soares.
Model-based single-month unemployment estimates from the Brazilian Labour Force Survey incorporating Google Trends data/ Caio César Soares Gonçalves, Luna Idalgo, Denise Britz do Nascimento Silva, Jan Van den Brakel, Paulo Roberto Betbeder Gonçalves. – Belo Horizonte : FJP, 2025.

82 p. : il. (Texto para discussão. Fundação João Pinheiro; n. 31)

G635m

1. Mercado de trabalho – Brasil. 2. Mão de obra – Brasil. I. Hidalgo, Luna. II. Silva, Denise Britz do Nascimento. III. Brakel, Jan Van den. IV. Gonçalves, Paulo Roberto Betbeder. V. Título. VI. Série.

CDU 331.6(81)

Ficha catalográfica elaborada por Ana Paula da Silva CRB-6 2390

PREFACE

Over the last decade, the global increase in internet use has provided multifaceted opportunities, mainly due to the abundance and variety of alternative data sources that have become available. The so-called big data is currently recognised as a potential data source for producing statistics and monitoring social and economic phenomena.

According to the Fundamental Principles of Official Statistics, "data for statistical purposes may be drawn from all types of sources," and the National Statistical Offices must be committed to pursuing the best ones to produce information, confirming the importance of contemplating alternative data sources besides traditional sample surveys and censuses. Hence, examining the possible role of big data in various contexts can contribute to a more consistent integration of diverse inputs within the scope of official statistics.

The Brazilian Labour Force Survey (BLFS) is the main source of official statistics on the labour market. This working paper aims to explore methods for producing unemployment figures based on a statistical modelling approach that incorporates big data sources, such as Google Trends. The modelling procedure for survey and Google Trends data is presented in detail, followed by a comprehensive description and new proposals for selecting Google Trends terms whose trends are related to the time series of unemployment estimates. In addition, results are discussed in view of their value for the production of official statistics.

This research represents the culmination of an innovative collaborative effort that emerged from the growing recognition of the need to modernize the production of official statistics in Brazil. The project originated from a partnership between the National School of Statistical Sciences (ENCE) and the Coordination of Household Sample Surveys (Coordination of Household Sample Surveys - COPAD) within the Division of Surveys (Diretoria de Pesquisas - DPE) of the Brazilian Institute of Geography and Statistics (IBGE). This collaboration was further enriched by the participation of Jan Van den Brakel, a distinguished professor at Maastricht University and senior statistician at the Central Bureau of Statistics (CBS) of the Netherlands, who brought invaluable international expertise to this initiative.

The genesis of this collaboration can be traced back to 2019, when Professor Van den Brakel visited ENCE to deliver a specialized course on big data and official statistics. His presentation sparked considerable interest among Brazilian researchers and statisticians, prompting ongoing discussions on the potential application of advanced model-based estimation techniques to enhance the quality and timeliness of labour market statistics in Brazil. A fruitful partnership was established bringing together the mutual benefits of combining Brazilian expertise in survey methodology with Dutch innovations in

statistical modelling and big data integration. Central to this collaborative effort was the involvement of Caio César Soares Gonçalves, a doctoral student at ENCE and researcher affiliated with the João Pinheiro Foundation (FJP) in Minas Gerais His dual affiliation offered a unique perspective that connected academic research with practical applications in regional statistics production. Under the supervision of Professor Denise Britz do Nascimento Silva (ENCE), Caio's doctoral research served as the basis for exploring the integration of Google Trends data with traditional survey-based unemployment estimates. The project team also included Luna Hidalgo from DPE/COPAD whose extensive expertise in the methodology and operational aspects of the Brazilian Labour Force Survey was crucial to developing and ensuring feasibility of the proposed innovations.

The collaborative nature of this project reflects a broader trend in official statistics towards international cooperation and knowledge sharing. The partnership with the Netherlands' CBS was particularly valuable given their innovative work in model-based labour statistics production since 2010. The Dutch experience provided a proven framework that could be adapted to the Brazilian context, while accounting for the unique characteristics of the Brazilian labour market and survey infrastructure.

The research agenda was driven by several pressing needs identified by COPAD and the broader statistical community in Brazil. These included the demand for more frequent publication of state-level unemployment estimates, the need for seasonal adjustment of unemployment series, and the growing interest in incorporating alternative data sources to enhance the precision and timeliness of official statistics. The Covid-19 pandemic further highlighted these needs, as policymakers required more timely and granular labour market statistics to respond to rapidly changing economic conditions.

The methodological approach developed through this collaboration represents a significant advancement in the application of model-based estimation techniques to official statistics in Latin America. By combining state-space modelling with big data integration, the research team established a framework that could potentially be applied to other statistical domains and adapted by other national statistical offices throughout the region.

The project's emphasis on experimental statistics reflects a cautious yet forward-looking approach to innovation in official statistics production. By initially presenting the results as experimental rather than official statistics, the team acknowledged the importance of conducting a thorough validation and stakeholder consultation before potentially incorporating these methods into regular statistical production processes.

This working paper thus represents not only a technical contribution to the literature on model-based estimation and big data integration, but also a testament to the value of international

collaboration in advancing statistical methodology. The partnership between Brazilian and Dutch institutions demonstrates how knowledge sharing can drive innovation and improve the quality of official statistics, ultimately benefiting policymakers and society by providing more accurate and timely information about labour market conditions.

ABSTRACT

This paper investigates the potential of incorporating Google Trends data into model-based unemployment estimates from the Brazilian Labour Force Survey (BLFS) to improve the precision and timeliness of official statistics. The study explores multivariate time series models that combine traditional survey data with big data sources, specifically Google search queries related to job seeking behaviour. The research addresses the growing demand for more frequent and precise labour market indicators, particularly at the state level and for specific demographic groups such as young people. The methodology employs state-space models and dynamic factor analysis to integrate unemployment statistics from the BLFS with Google Trends series. Variable selection techniques, including penalized regression elastic net and time series clustering with dynamic time warping distance, are used to identify relevant Google search terms. The analysis covers the period from January 2012 to December 2021, focusing on national estimates and two selected states: Minas Gerais (largest sample) and Roraima (smallest sample). Results demonstrate that incorporating Google Trends data can enhance the quality of unemployment estimates, particularly for areas with smaller sample sizes. The model-based approach demonstrates potential for producing single-month estimates and nowcast indicators, addressing the need for more timely labour market statistics. This research contributes to the literature on multi-source statistics and provides insights for national statistical offices seeking to leverage big data for improving official statistics production in developing countries.

FJP
FUNDAÇÃO
JOÃO PINHEIRO

TABLE OF CONTENT

1 INTRODUCTION

The growing global use of the internet offers a wide range of opportunities. The web is used for communication, work, education, shopping, entertainment, and many other purposes. People's behaviour on networks can reveal valuable information for understanding current reality, recognising social changes, and making predictions. For example, search activities can pinpoint people's preferences, such as the type of news they seek, entertainment choices, consumption intentions, travel destination planning, and searches for new job opportunities.

The amount and variety of alternative data sources, such as big data, are rapidly growing and consolidating as potential data sources for producing statistics and monitoring social and economic phenomena. According to Pfeffermann (2015), the use of big data is one of the most intriguing challenges for the production of statistical information. Furthermore, Hand (2018) highlighted the need to investigate how to improve the use and analysis of administrative and big data to replace current approaches or overcome the challenges of combining different data sources. Hence, examining the possible role of big data in different contexts can contribute to a more consistent integration of these data sources to produce official statistics.

In addition, one of the Fundamental Principles of Official Statistics refers to the commitment to pursue the best source to produce information (United Nations, 2014, 2015). This confirms the importance of contemplating alternative data sources besides traditional sample surveys and censuses. One way to conduct this investigation is to explore methods for producing indicators based on statistical modelling, referred to as the model-based approach, which permits the combination of different data sources.

De Waal, Van Delden, and Scholtus (2020) mentioned initiatives of national statistical institutes (NSIs), mainly in Europe, to produce multi-source instead of single-source statistics. Several advantages can be mentioned when combining data sources such as survey data with administrative and big data. Multi-source statistics can provide policymakers and society with more timely and detailed statistics. Furthermore, it reduces the cost of data collection and processing and lessens the burden on respondents.

Since some data sources may be available even before an official statistic is ready to be published, predicting a statistical indicator with extremely tight time frames may be possible. The so-called nowcast estimates are claimed to provide a precise early indicator of the phenomenon of interest.

Thus, this paper investigates whether model-based estimates that combine data from different sources can be used to expand the scope of statistical outputs already provided by regular surveys in Brazil. The two data sources investigated in the present paper are unemployment statistics from the Brazilian Labour Force Survey (BLFS) and job search queries from Google Trends.

The BLFS[1], called Continuous National Household Sample Survey (PNADC), is the main source of official statistics on the Brazilian labour market. Its monthly publication currently refers only to national unemployment figures. It is based on information compiled from rolling three-month data (also called rolling quarterly data). On the other hand, the states' results are published based on quarterly survey data aggregated by calendar quarter (Instituto Brasileiro de Geografia e Estatística, 2022). Therefore, there is a demand to expand the production of state-level monthly estimates. Considering that the BLFS sample was not designed to provide single-month estimates with adequate precision at the state-level, methods based on statistical modelling could improve the quality of these required estimates. A model-based procedure has already been used, for example, by Statistics Netherlands since 2010 (Van den Brakel; Krieg, 2009, 2015), for the production of local labour market statistics in the United States (the Local Area Unemployment Statistics - LAUS program) (United States, 2018; Pfeffermann; Tiller, 2006), and for monthly experimental labour market statistics in the United Kingdom (Office for National Statistics, 2019; Elliott; Zong, 2019). It is also noteworthy that the need to produce single-month estimates to monitor the labour market has increased, especially after the Covid-19 outbreak.

Since Google is the most widely used browser globally, Google Inc. launched Google Trends in 2006, as a specific tool to monitor search queries in Google Search. This explorer analyses the Google web searches and reports a scaled time series of an individual query search (e.g., job, CV) or a specific topic (e.g., employment). Google Trends also provides data disaggregated by location (countries and states) and different time frequencies (daily, weekly or monthly) (Google, 2022).

The hypothesis is that the series represented by the level of queries for words related to seeking a job in the Google search tool presents similar behaviour to the unemployment series measured by the BLFS. The similarity of the trends can be exploited in the modelling procedure through a multivariate time series model, which would potentially increase the quality of the statistics, especially when the sample size is small, such as in Brazilian states or among specific population groups, like young people. People from this age group tend to be the most active internet users, so they are most likely to look for a job using Google.

[1] The Brazilian survey is called Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) (Instituto Brasileiro de Geografia e Estatística, 2022).

The use of Google Trends (GT) series in a multivariate time series models poses challenges, such as the dimensionality issue. Variable selection techniques, like penalised regression elastic net (Hui; Hastie, 2005), have been presented in the literature to address this issue. In addition, this paper proposes using time series clustering with dynamic time warping (DTW) distance (Aghabozorgi; Shirkhorshidi; Wah, 2015) to target the predictors (GT series). Another strategy tested is to fit bivariate state-space models for the unemployment estimates using each Google Trends series to identify cases of correlated trend evolution. After the selection stage, the models are estimated in two steps (Doz; Giannone; Reichlin, 2011), using principal component analysis and a dynamic factor model. Furthermore, all models take into account the sampling errors, as presented by Schiavoni *et al.* (2020).

Therefore, this paper aims to investigate the potential of producing precise estimates of monthly unemployment figures based on multivariate time series models that integrate survey data with big data. In addition, it seeks to determine whether combining information from Google Trends series related to job searches with the unemployment figures can improve the precision of these estimates for states and population groups, such as young people. Furthermore, the potential to produce a nowcast estimate for one month ahead is also explored. The analysis period spans from January 2012 until December 2021. The study focuses on national estimates and two selected states, those with the largest (Minas Gerais) and the smallest (Roraima) BLFS sample sizes.

This paper contributes to the literature by testing and discussing the usefulness of incorporating big data in the production of official statistics to improve its quality, relevance, accuracy, and timeliness (OECD, 2011, p. 7-9). This use of multi-source information also supports the debate on best practices for utilizing big data, especially the Google Trends database, to improve precision and to deliver nowcast indicators in the context of a large Latin American country based on a sample survey such as the BLFS.

The remainder of the report is organised as follows. Section 2 reviews the use and applications of Google Trends queries series. Section 3 details the Brazilian Labour Force Survey features, and Section 4 describes Google Trends data and procedures to select the job search related words. Model and estimation procedures are presented in Section 5, while Section 6 displays the results and compares the models developed at the national and state levels, as well as for the young population group. Section 7 contains the final remarks.

2 LITERATURE REVIEW

Official statistics are based on data collected and processed by national and regional government statistical offices and associated government agencies. These institutions use censuses, sample surveys, administrative records, or various combinations of these as data sources (Pfeffermann, 2015). Administrative data sources have been studied for some time, with a focus on improving their quality (Hand, 2018). Regarding big data, several challenges still need to be overcome (Pfeffermann, 2015), and it is worth noting the current status of statistics production and the opportunities to incorporate new data sources.

National statistical bureaus traditionally use design-based estimates (i.e., weighted according to the probability distribution generated by the survey sampling design), following the precepts of classical sampling theory. The Horvitz-Thompson and the generalised regression estimators are the most used ones, as presented by Cochran (1977) and Särndal, Swensson, and Wretman (1992). However, these estimators have desirable properties that fail when the sample size is small. This is one of the main disadvantages of design-based estimates, leaving room for model-based estimation procedures.

The combination of sampling and time series methods has been the target of several studies, including those by Scott and Smith (1974) and Scott, Smith, and Jones (1977), who applied signal extraction theory. Later, papers by Binder and Dick (1989), Pfeffermann (1991) and Tiller (1992) introduced state-space models for time series analysis of repeated surveys. They demonstrated the relevance of considering sampling errors as a latent component of design-based estimates when modelling time series from sample surveys. The literature advanced to enable the identification of time series models for the unobserved sampling error process according to data availability, as presented by Pfeffermann, Feder, and Signorelli (1998).

In addition, several aspects related to the sample survey features were examined: rotation group bias (Bailar, 1975); discontinuities and survey redesign (Van den Brakel; Krieg, 2015; Van den Brakel; Buelens; Boonstra, 2016); time series modelling of compositional data (Silva, 1996; Silva; Smith, 2001); and small area estimation (Rao; Mingyu, 1994; Datta *et al.*, 1999; Boonstra; Van den Brakel, 2022; Pfeffermann; Tiller, 2006; Krieg; Van den Brakel, 2012; Van den Brakel; Krieg, 2016). Durbin and Quenneville (1997) discuss benchmark procedures for multivariate structural time series models observed at different frequencies obtained by repeated sample surveys. More recently, Van den Brakel, Souren, and Krieg (2021) proposed introducing mechanisms to better capture the time series dynamics and accommodate abrupt changes in official statistics that occurred immediately after the Covid-19 outbreak.

Moreover, Harvey and Chia-Hui (2000) developed a bivariate model combining administrative data and sample survey estimates. A more recent data integration approach focuses on jointly modelling time series from repeated surveys and big data, as presented by Schiavoni *et al*. (2020).

The term "big data" is used to characterize a dataset with a hard-to-manage volume of data in a wide variety of formats that arrives at high speed and is often unstructured. However, there is no generally accepted definition of the term. Hand (2018), for example, used the term for data collected through an automatic system. According to Groves' (2011) terminology, big data is organic data in the sense that it is generated by unplanned data sources for pre-specified purposes, as opposed to designed data for statistical purposes.

As presented by Vichi and Hand (2019), there are numerous examples of this new data source, such as web scraping, administrative statistics, social media, and machine-generated data (Internet of Things - IoT). These examples are as varied as social media data, credit card transaction records, mobile phone call records, commercial website data, voluntarily provided geographical information, and search engine data, among others (Jianzheng *et al*., 2016). Among web search engine sources, the best known is the Google Trends tool.

2.1 Literature using Google Trends

Among the different possibilities of big data, Google Trends presents an index of searches by words or categories organised in time and by geographic areas. The initial studies using Google Trends time series, presented by Hyunyoung and Varian (2009a, 2009b, 2012), revealed their capacity to produce nowcast estimates of some economic indicators and to identify turning points in the series. The authors found that the query indices are correlated with unemployment claims, travel planning, consumer confidence, and economic activities such as retail, automotive, and home sales.

Other papers have also indicated the potential of the search query time series provided by Google Trends to detect, monitor, and forecast phenomena. Vosen and Schmidt (2011) created an indicator for private consumption based on query categories of goods and services. The forecasting experiments showed that the Google indicator outperformed two of the USA's most widely used survey-based indicators related to private consumption: the University of Michigan Consumer Sentiment Index and the Consumer Confidence Index. Woo and Owen (2018) also examined private consumption within the USA using consumption-related and news-related Google Trends query data. The forecasting models, including the Google Trends data, outperformed those that did not consider that source.

Applications exploiting Google Trends are varied. A complete list of the areas of study employing Google Trends query data is presented by Seung-Pyo, Hyoung and San (2017), who conducted a social network analysis of 657 documents (articles, conference papers, among others) published over ten years (2006-2016), since the introduction of Google Trends. These areas included IT, communications, medicine, health, business, and economics. Some examples are related to influenza epidemics (Ginsberg *et al*., 2009; Cook *et al*., 2011), suicide occurrence (Kristoufek; Moat; Preis, 2016), unemployment rate (Askitas; Zimmermann, 2009), automotive sales (Carrière-Swallow; Labbé, 2011), stock returns (Preis; Moat; Stanley, 2013; Salisu; Ogbonna; Adediran, 2020), hotel bookings (Yang; Bing; Haiyan, 2014; Rivera, 2016), energy (Hassani; Silva, 2016), recession (Tao *et al*., 2015), inflation expectations (Guzman, 2011), and exchange rates (Bulut, 2017; Takumi *et al*., 2021).

However, it is necessary to pay attention to the word selection step, representativeness, and possible spurious associations, as seen in the case of Google Flu Trends. The prediction of the spread of influenza was the first application using Google Trends data, according to Seung-Pyo, Hyoung and San (2017). Ginsberg *et al*. (2009) showed the possibility of tracing the spread of influenza one to two weeks before the Centers for Disease Control and Prevention (CDC) using Google flu search activities. This paper inspired the application to monitor other diseases such as dengue (Althouse, 2011) and zika (Yue, 2017; Morsy *et al*., 2018). Butler (2013) refuted the results of Ginsberg *et al*. (2009), indicating that the percentage of the US population with influenza-like illness was overestimated by more than twice the actual number, based on laboratory surveillance reports from all regions of the US. Another study focusing on this limitation was presented by Lazer *et al*. (2014). They generalised this problem to all applications and pointed out the need to discuss the nature of phenomena captured by search data and data obtained from social media platforms.

Other examples of Google Trends' weaknesses and challenges are mentioned in Lui, Metaxas, and Mustafaraj (2011). The authors conducted a study to analyse the winning chances of candidates in the 2008 and 2010 US Congressional elections using Google Trends data. The estimated model did not yield results that stood out from the traditional polling of the New York Times. Lui, Metaxas, and Mustafaraj (2011) explained that these results could be due to the amount of negative information about the candidates. However, solutions to overcome this limitation have been proposed in the literature to differentiate positive and negative news, as Vosen and Schmidt (2011) applied in the financial area.

Therefore, it is essential to identify which areas of Google Trends data are most useful and the best technique to extract information from them. Studies focused on labour market figures, such

as those related to unemployment, found positive results. The applications aim to improve the ability to forecast (and nowcast) labour market indicators.

2.2 References using Google Trends for unemployment prediction

The use of time series of job search queries from Google Trends to predict unemployment has already been investigated by several authors, for many countries. Askitas and Zimmermann (2009) found strong correlations between groups of keyword searches and unemployment rates. They applied an error-correction model to monthly German data, and the results indicated the potential for utilising Google Trends but acknowledged the need for further investigation. Other studies confirmed this result by testing the forecast accuracy, such as D'Amuri (2009), based on Italian quarterly data; D'Amuri and Marcucci (2010), which applied the model to monthly US unemployment rates; and Suhoy (2009), focused on the Israeli case. It is also interesting to note how these articles implemented the search queries and which time series models were chosen. D'Amuri (2009) used a single term, "job offers" ("offerte di lavoro"), and modelled the unemployment rate with an autoregressive integrated moving average (ARIMA) model including other exploratory variables besides the Google Trends series, such as the Employment Expectations Index and the Industrial Production Index. D'Amuri and Marcucci (2010) also selected an ARIMA structure to model the unemployment rate and considered Google Trends (keyword "jobs") and initial claims (US unemployment insurance) as exogenous variables. Suhoy (2009) investigated other economic indicators besides the unemployment rate and utilised categories of words. Suhoy (2009) used a single series of the query index of the human resources (recruitment and staffing) category to project the monthly unemployment rate. The Israeli Labour Force Survey provided the quarterly unemployment rate series while the job opening ratio, published by the Ministry of Industry, Trade and Labour, was included as auxiliary information. This data modelled with a state-space and ARIMA approach to interpolate and produce a monthly unemployment rate series.

Along with the discussion of various model types and keywords, or categories of words, there are papers addressing several issues related to the use of Google Trends series. It is relevant to highlight the following aspects and challenges to produce unemployment statistics that are covered in the literature regarding the use of Google Trends series: a) the identification of turning points in nowcast attempts; b) job searching online segmented by specific population groups; c) experiences from different countries and geographic-related job searches for the production of regional indicators; d) combination of time series sampled at different frequencies; and finally , e) dimension reduction

techniques and strategies to deal with high dimensionality, considering the most relevant terms/words and targeting procedures. Each of these topics is described below.

One of the main targets of short-term analysis is to detect sudden changes in a time series. D'Amuri and Marcucci (2017) reported that the model incorporating Google Trends outperformed most of the other tested models. They emphasised the positive point that this model predicted well the turning point observed at the beginning of the last great recession (December 2007 to June 2009 in the US). Anvik and Gjelstad (2010) also supported the idea that the Google Trends series contain useful information to predict short-term unemployment changes. Hence, there is potential to use them for nowcasting and predicting turning points, or at least to provide a first sign of a change in trend.

Another notable discussion found in the Google Trends related literature is the internet activity of different population groups. Young people are usually the most active internet users and, consequently, are more willing to look for a job through this channel. Therefore, some studies have investigated the use of the Google Trends series for a specific population group, relating this information to users' characteristics. Fondeur and Karamé (2013) tested the Google Trends series to predict French youth unemployment (claimant count). The authors used a bivariate state-space model, and the incorporation of Google Trends series enhanced the unemployment forecast. Naccarato *et al*. (2018) investigated the possibility of using the keyword "job offer" ("offerte di lavoro") to enhance forecasts of youth unemployment in Italy. A monthly ARIMA model (with just the unemployment rate time series) and a vector autoregression (VAR) model (with the unemployment rate and the Google Trends keyword time series) were examined, and the results indicated a reduction in the forecast error when the Google Trends series was taken into account. Another application considering a specific population group, Canadians between 25 and 44 years old, was reported by Dilmaghani (2019) to predict the unemployment rate.

In Brazil, Shikida *et al*. (2012) used the category "jobs" of Google Trends in ARIMA models and tested whether it could reduce the prediction error of the unemployment rate obtained from the Brazilian Monthly Employment Survey. The results showed that the inclusion of Google Trends in the model did not lead to notable benefits. Alabrella (2017) also did not detect any improvement in the predictive power when the Google Trends series was incorporated in a model integrating estimates from the current Brazilian Labour Force Survey (BLFS) and principal components obtained from the analysis of the 58 Google Trends series.

In addition to the studies based on the Brazilian unemployment rates series, international experiences using unemployment insurance data as an auxiliary series have been reported for several countries: Germany (Askitas; Zimmermann, 2009), Italy (D'Amuri, 2009; Naccarato *et al*., 2018), United

States (D'Amuri; Marcucci, 2010; Hyunyoung; Varian, 2012), Israel (Suhoy, 2009), France (Fondeur; Karamé, 2013), and Netherlands (Schiavoni *et al*., 2020). Similar studies have been carried out for other countries such as Norway (Anvik; Gjelstad, 2010), the UK (Smith, 2016), Canada (Dilmaghani, 2019), China (Zhi, 2014), and Romania (Simionescu, 2020). Barreira, Godinho, and Melo (2013) considered countries in southwestern Europe (Portugal, Spain, France, and Italy) and explored the existing diversity related to job searching on the internet, examining both geographical and cultural aspects, as well as language considerations. D'Amuri and Marcucci (2017) examined employment growth predictability for the 50 US States plus the District of Columbia, whereas Borup and Schütte (2022) carried out state-level analysis for US states, tuning the search procedure by state. Simionescu (2020) also applied a regional perspective but chose the same three words for all 42 Romanian counties in a panel data model. Simionescu and Cifuentes-Faura (2022) elaborated unemployment predictions for Portugal, Spain, and regions with two search terms (unemployment and job offers) in their respective languages. The set of words with different terms is a feature that can be explored since residents of distinct regions may use alternative terms to search for jobs online or names of job search services operating in specific geographic areas.

Another aspect addressed in the literature refers to the treatment of different time frequencies. Google Trends provides daily, weekly, and monthly data, which can be synthesised at the same frequency as the variable of interest. In addition, one can jointly model the mixed frequencies as performed by Smith (2016) and Maas (2020) for the US unemployment rate via a mixed data sampling (MIDAS) regression model. Another example was reported by Schiavoni *et al*. (2020), who tested weekly and monthly frequency series in principal component analysis to obtain factors and used a monthly dynamic factor model to re-estimate the corresponding factors.

The final aspect discussed in the existing literature refers to the application of high-dimensional models instead of univariate or low-dimensional models, such as ARIMA, bivariate state-space, or VAR models (Fondeur; Karamé, 2013; Naccarato *et al*., 2018). In addition, Schiavoni *et al*. (2020), Smeekes and Wijler (2021), Wijler (2021), and Borup and Schütte (2022) explored this aspect by considering a set of words and composed models to deal with this high dimensionality. The advantages of using a set of search words instead of a general category or a unique word are varied: the inclusion of different terms compensates the word selection problems, regional and demographic language characteristics are taken into account for job searches and capture different behaviour of search words across time (Borup; Schütte, 2022). In relation to word selection, Schiavoni *et al*. (2020) and Borup and Schütte (2022) employed a regularisation method (Elastic Net) to filter out irrelevant series as a procedure to target the predictors according to the study of Bai and Ng (2008). Schiavoni *et*

al. (2020) utilised this as a pre-selection method for Google Trends, which is then combined with principal component analysis in the next step, as part of the dynamic factor model. Borup and Schütte (2022) tested the pre-selected Google Trends in regression methods such as random forest, bagging and complete subset regression.

A dynamic factor model, in combination with principal component analysis, is based on the assumption that a large set of auxiliary series contributes to the target series of interest. Penalised regression, on the other hand, is based on the assumption that a small part of the auxiliary series, out of a large set of potential series, contributes to the target series of interest. Dingdong *et al*. (2021) introduced the penalised regression with the inferred seasonality module (PRISM) method, a two-stage estimation procedure to forecast the unemployment initial claim series. The first stage is the seasonal decomposition to produce the estimated seasonally adjusted component. The second stage involves a penalised regression for the variable of interest, including exogenous time series. Smeekes and Wijler (2021) and Wijler (2021) extended the classical single-equation error correction model, proposing the single-equation penalised error correction selector (SPECS) for modelling a large number of cointegrated series. The authors illustrated the method with Google Trends series related to job searches from the Netherlands. It is important to note all quoted works, except Schiavoni *et al*. (2020), did not take into account the sampling error process for cases in which the time series are obtained from repeated sample surveys. For a review of additional papers that have investigated unemployment prediction using Google Trends, see Simionescu and Cifuentes-Faura (2022, sec. 3).

Considering the spectrum of available literature on the use of the Google Trends series, the choice of words (whether a single word or category) and the choice of one or more factors to represent the essence of the series stands out. Additionally, the frequency of the Google Trends series, the statistical procedure for handing mixed frequencies, and the estimation method vary according to the proposed objective. These methodological aspects are presented in Section 5.

3 THE BRASILIAN LABOUR FORCE SURVEY

The Brazilian Labour Force Survey (BLFS) is the largest household survey conducted by the Brazilian Institute of Geography and Statistics (IBGE) and produces the official unemployment figures for the country and its administrative divisions (five regions, further divided into 26 states and the Federal District). The BLFS has a stratified two-stage cluster design and is a rotating panel survey with five rotation groups and partially overlapping samples of secondary sampling units (households). The primary sampling units (PSU) and census enumeration areas are stratified based on geographical regions and statistical criteria (household income according to the most recent census data). In each PSU, 14 households are selected, and all household members are interviewed (Freitas *et al*., 2007). However, only those aged 14 or older are considered for employment purposes. The rotation scheme is 1-2(5), in which a household is visited in one month, is out of the sample for two months, and this cycle is repeated five times (Instituto Brasileiro de Geografia e Estatística, 2022). Therefore, there is an overlap of PSUs once every three months.

Since the start of the survey, demand has been growing for the publication of monthly indicators (using a single-month sample) for various geographical levels, particularly the states. Due to the BLFS sampling design, national statistics are released monthly (based on rolling 3-month data, referred to as rolling quarters) and quarterly (using quarterly data). Subnational figures for regions and states are also published quarterly. The data used in this paper refers to the figures published after a major revision of the entire series released at the end of November 2021. The revised estimates were calculated based on calibration by age and sex and the variance calculation method implemented by bootstrap. For more details on this revision, see Instituto Brasileiro de Geografia e Estatística (2021).

Previous to the Covid-19 pandemic, the BLFS collection mode was CAPI (computer-assisted personal interviewing). Since the end of March 2020, when states began to adopt social isolation measures in accordance with municipal government protocols, the CAPI technique was entirely replaced by CATI (computer-assisted telephone interviewing) due to the health crisis. During the second quarter of 2021, the CAPI mode was reinstated. The effects of this change on data collection mode could not be contemplated in the models due to the difficulty of distinguishing survey mode effects from the lockdown's impact on unemployment since both occurred simultaneously (Gonçalves *et al*., 2022).

4 GOOGLE TRENDS JOB SEARCH DATA

Google Trends has provided a series of word queries since January 2004 for several countries and, in some cases, the searches can be specific to states or provinces. The data are also grouped into categories such as health, employment, sports, travel, etc., using a natural language classification engine. However, Google Trends does not report the exact search volume. The measure is normalised on a scale from 0 to 100, where the maximum (100) represents the query's highest point considering a specific time set by the user. This procedure is adopted because the data are samples of Google searches, which are not publicly available (Google, 2022). Therefore, it is essential to identify the period extracted, as the series may exhibit relevant differences with the insertion inclusion of new time points.

There are some relevant details to mention about the Google Trends data. The data comprises popular terms, so words searched by only a few people are excluded. Another important point is the exclusion of repeated queries made by the same user in a short period to avoid overestimating the corresponding measure for the searched term. Finally, apostrophes and other special characters are also excluded to improve the accuracy of the series, according to Google (2022).

The words related to job search activities were selected according to the following steps. First, an initial group of the most expected words was constructed. Words like vacancies, jobs, openings, opportunities, resumes, and names of recruiting firms were included in this list. Terms already used in the Brazilian literature were also considered (Shikida *et al*., 2012; Alabrella, 2017). All words were searched in Brazilian Portuguese[2]. This initial set of words was expanded using the option provided by Google Trends titled "related searches". Each single search word generates a list of 50 associated queries. The aim was to find the maximum number of unique terms used for job searches. Schiavoni *et al*. (2020) and Borup and Schütte (2022) also utilised Google algorithms to identify additional terms semantically related to the job search queries initially provided.

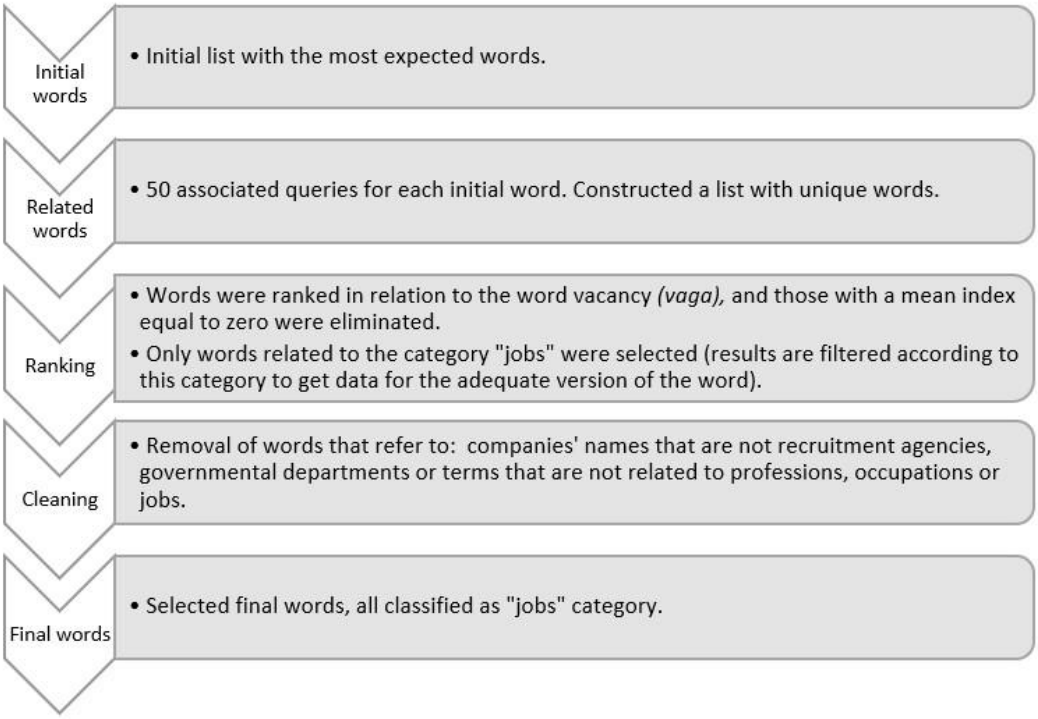
Second, a search ranking of the terms was created to eliminate the lesser-used words over time. It was a relative ranking to the word "vaga" (vacancy), considered the most relevant one. Google Trends does not provide the exact number of queries, but it allows the construction of a series of up to five words, with the index being built relatively between them. In this way, making pairs with the word "vaga", the series of each term was extracted. In this step and the following ones, the search filter within the Google Trends "jobs" category was applied to capture the queries specifically related to job searching. Since the Google Trends index is a normalised measure on a 0-100 scale, the tool

[2] The list of initial words is reported in Appendix A.1.

indicated cases with values smaller than one but different from zero. In these instances, the indices were considered to be zero. The mean of the index was calculated from January 2012 to December 2021. These summary measurements were used to rank and remove words with a zero mean.

Lastly, an additional procedure was implemented to eliminate words deemed “inadequate”, and the following types of queries were excluded: a) queries specifying professions, such as “motorista” (driver), or those including specific years, such as “salário mínimo 2019” (minimum monthly wage of 2019); b) queries using company names, like “Vale” (a large mining company), without explicitly referencing job opportunities, except in the case of job agency names; and c) isolated terms such as names of government departments like “Ministério do Trabalho” (Ministry of Labour) or referring to laws. After excluding these words, the final list of monthly Google Trends series terms related to job search in Brazil and for selected states was ready for use. Figure 1 summarises these word selection steps. Data are freely downloadable and were extracted with the gtrendsR package (Massicotte, 2022).

Figure 1: Steps for word selection



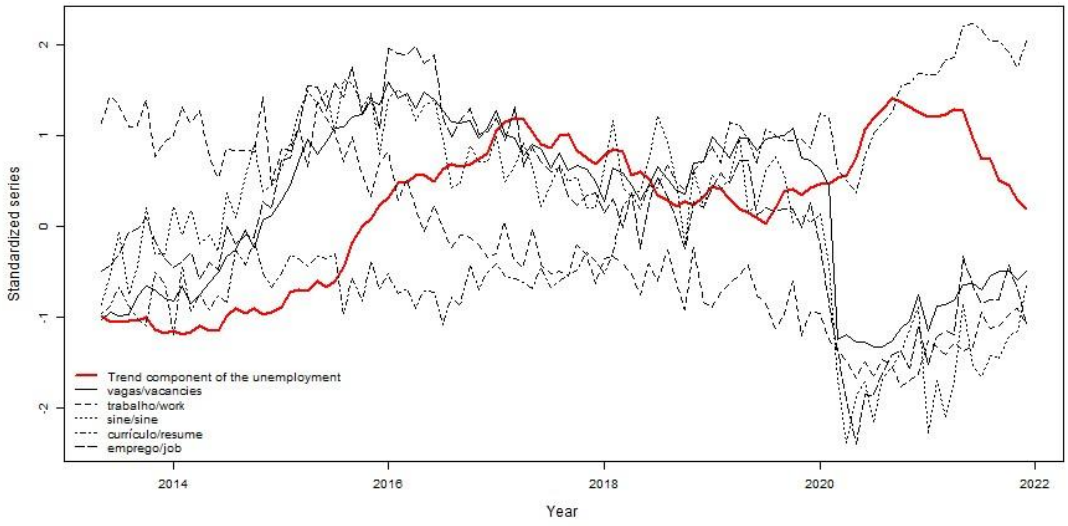
Source: Elaborated by the authors.

Graphic 1 shows the trajectory of the Google Trend indexes for some words related to job search consulted in the state of Minas Gerais, such as “vacancies”, “work”, “resume”, “job”, and “SINE”, which is the acronym in Portuguese for National Employment System, a governmental

recruitment service[3], together with the single month employment trend. Some of these words showed a drop in searches during the initial months of the pandemic, despite an increase in unemployed people in this period, since the country was facing a lockdown.

The selected Google Trends series showed similar behaviour to the total unemployed people measured by the BLFS (red line in Graphic 1). This similarity of trajectories could be explored in the modelling procedure and potentially improve the quality of the statistical figures.

Graphic 1: Single-month unemployment trend estimates and selected Google search terms – Minas Gerais – May 2013-Dec. 2021



Source: Brazilian Labour Force Survey Series (IBGE); Google Trends (2025). Elaborated by the authors.
Note: A basic structural model (BSM) estimated the single-month unemployment trend.

[3] The list of Google search terms selected in the last stage is reported in Appendix A.2.

5 TIME SERIES MODELS

Structural time series models are a class of models that decompose an observed time series into a set of unobserved components such as trend, seasonal, cyclic and regression components. See, e.g. Harvey (1989) or Durbin and Koopman (2012) for an introduction to structural time series modelling. An interesting aspect of the multivariate structural time series models is the possibility of defining common components for several series (Harvey, 1989, p. 9). Series might be related over time, even if the levels of the series are different. In such cases, more parsimonious models can be obtained by defining common unobserved trend components, resulting in so-called seemingly unrelated time series (SUTSE) models. These models are particularly beneficial for producing more timely and precise nowcasts for parameters of interest.

Applying these models to series obtained from repeated surveys generally requires additional components to model the sampling error (Pfeffermann, 1991; Durbin; Quenneville, 1997; Binder; Dick, 1989; Van den Brakel; Krieg, 2015; Schiavoni *et al*., 2020). In this paper, SUTSE models are used to combine target series from repeated surveys with related series, such as Google trends, to make more precise and timely nowcasts of survey target indicators. Thus, Google trends series are used as auxiliary information without assuming causal relationships with the target series of the repeated survey.

An alternative approach to incorporate an auxiliary series in the model is to extend the time series model for the survey series by adding a regression component for the auxiliary series. The major drawback of this approach is that the auxiliary series will partially explain the trend and seasonal effects. This hampers the estimation of a trend for the target variable. This approach is therefore not considered in this paper.

The model proposed by Schiavoni *et al*. (2020) combines a signal extraction model for unemployment estimates from a repeated sample survey (Equation 1) with the Google Trends series, which is modelled by a factor model (Equation 12). It is a type of SUTSE model. The two models are detailed separately in the following sections, and their combined formulation is presented next.

5.1 Modelling the survey data

A signal extraction model is defined for the BLFS estimates to account for the intrinsic sampling error component due to the survey sampling process. Let $\hat{y}_t$ denote the design-based estimate for unemployment (total of unemployed people) in month t, which is decomposed as the unknown population parameter θ_t and the sampling error e_t:

$$\hat{y}_t = \theta_t + e_t. \tag{1}$$

The parameter θ_t in Equation 1 is the signal and represents the true unknown unemployment value, which in turn can be decomposed into three unobserved components - the trend, seasonal, and irregular components:

$$\theta_t = L_t + S_t + I_t, \quad I_t \sim \mathcal{N}(0, \sigma_I^2) \tag{2}$$

where L_t denotes the level, S_t a seasonal effect and I_t the unexplained variation of the population parameter. The trend is modelled with the smooth trend model (Durbin and Koopman, 2012), which is defined as

$$L_t = L_{t-1} + R_{t-1}, \tag{3}$$
$$R_t = R_{t-1} + \eta_{R,t}, \quad \eta_{R,t} \sim \mathcal{N}(0, \sigma_R^2), \ \operatorname{cov}(\eta_{R,t}, \eta_{R,t'}) = 0, \quad \forall t \neq t'. \tag{4}$$

In Equation 3, L_t denotes the trend level, and R_t represents slope that can be interpreted as a change in the trend level. The smooth trend model only includes a stochastic term $\eta_{R,t}$ in the slope equation (4), which is an independent white noise series with a time-constant variance. This implies, by construction, that R_t is $I(1)$, while L_t is $I(2)$[4].

A trigonometric model is assumed (Durbin and Koopman, 2012) for the seasonal component S_t . It can be expressed via six frequencies of the monthly seasonal series:

$$S_t = \sum_{\iota=1}^{\frac{s}{2}=6} S_{\iota,t}, \tag{5}$$

$$\begin{pmatrix} S_{\iota,t} \\ S_{\iota,t}^* \end{pmatrix} = \begin{bmatrix} \cos\left(\frac{\pi\iota}{6}\right) & \sin\left(\frac{\pi\iota}{6}\right) \\ -\sin\left(\frac{\pi\iota}{6}\right) & \cos\left(\frac{\pi\iota}{6}\right) \end{bmatrix} \begin{pmatrix} S_{\iota,t-1} \\ S_{\iota,t-1}^* \end{pmatrix} + \begin{pmatrix} \eta_{S,\iota,t} \\ \eta_{S,\iota,t}^* \end{pmatrix}, \tag{6}$$

$$\begin{pmatrix} \eta_{S,\iota,t} \\ \eta_{S,\iota,t}^* \end{pmatrix} \sim \mathcal{N}\left(0, \sigma_S^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \quad \text{where } \iota = 1, \dots, 6, \tag{7}$$

$$cov(\eta_{S,\iota,t}, \eta_{S,\iota,t'}) = 0, \quad \forall t \neq t', \tag{8}$$

$$cov(\eta_{S,\iota,t}^*, \eta_{S,\iota,t'}^*) = 0, \quad \forall t \neq t'. \tag{9}$$

The last component of Equation 2 is the irregular component (I_t), which is treated as an uncorrelated zero-mean disturbance with variance σ_I^2.

As proposed by Binder and Dick (1989), the sampling error e_t in (1) is scaled with the standard error of the input series to account for heteroscedasticity that arises from varying sample sizes of the survey:

[4] A series is integrated with order d if it is stationary after differentiating it d times. Thus, if the series is $I(0)$, it is stationary, while if it is $I(1)$, the series is stationary in the first difference (Harvey, 1989, p. 429).

$$e_t = \hat{c}_t \tilde{e}_t, \quad \hat{c}_t = \sqrt{\text{var}(\hat{y}_t)} \tag{10}$$

According to the rotating panel design of the BLFS, households are interviewed quarterly. As a result, the sampling error at time t is correlated with the sampling error at time $t-3$. This autocorrelation can be modelled with an AR(3) model, as in Van den Brakel and Krieg (2009), with only one non-zero coefficient ϕ (at time $t-3$):

$$\tilde{e}_t = \phi \tilde{e}_{t-3} + \eta_{\tilde{e},t}, \qquad \eta_{\tilde{e}} \sim \mathcal{N}(0, \sigma^2_{\tilde{e}}). \tag{11}$$

The AR(3) process identification was confirmed by estimating the autocorrelation function and partial autocorrelation function following Smith (1978) but based on BLFS microdata from overlapping PSUs. The Yule-Walker equations were used to estimate the coefficient ϕ. Hence, the autocorrelation is assessed externally to the state-space model framework. Note that the variance of the scaled sampling error in (11) is $\text{var}(\tilde{e}_t) = \sigma^2_{\tilde{e}}/(1-\phi^2)$. If the maximum likelihood estimate $\hat{\sigma}^2_{\tilde{e}} \approx \left(1-\hat{\phi}^2\right)$, then $\widehat{var}(\tilde{e}_t) \approx 1$ and $\widehat{var}(\text{e}_t) \approx \widehat{var}(\hat{y}_t)$.

5.2 Modelling Google Trends data

As Schiavoni *et al*. (2020) presented, modelling big data involves reducing the high dimensionality of the Google Trends series to a few common factors using principal components and factor analysis. Let $\mathbf{x}_t$ denote an $n^* \times 1$ vector of the Google Trends series, which can be expressed as a factor model. Furthermore, let $\mathbf{f}_t$ denote an $r \times 1$ vector containing the common factors of $\mathbf{x}_t$, i.e.:

$$\mathbf{x}_t = \mathbf{\Lambda} \mathbf{f}_t + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}). \tag{12}$$

The factor loading matrix $\mathbf{\Lambda}$ has dimension $n^* \times r$, and $\boldsymbol{\xi}_t$ are the idiosyncratic components represented in a vector of size $n^* \times 1$ with mean zero and an $n^* \times n^*$ diagonal variance matrix $\mathbf{\Psi}$.

As presented by Durbin and Koopman (2012, sec. 3.7), the application of factor analysis in time series means that the time dependence of measurements is taken into account by replacing the serial independence assumption of f_t by a model that accounts for serial dependence. Here, the factors are modelled as a random walk process, i.e.

$$\mathbf{f}_t = \mathbf{f}_{t-1} + \mathbf{u}_t \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r). \tag{13}$$

with $\mathbf{u}_t$ innovations. The model for $\mathbf{x}_t$ that incorporates a dynamic process for the factors is referred to in the literature as a dynamic factor model, which is composed by Equations 12 and 13. For more details about principal component and factor analysis for time series, see Tsay (2013, chap. 6), Stock and Watson (2017), Wei (2019, chap. 4-5) and Doz and Fuleky (2020).

The matrices $\mathbf{\Lambda}$ and the variance matrices in (12) can be estimated by maximum likelihood procedures. To guarantee the identifiability of the dynamic factor model, restrictions must be imposed. In this case, the covariance matrix of the common factors is taken to be an identity matrix $\mathbf{I}_r$ (Harvey, 1989, p. 451), meaning that the factors are uncorrelated, which is in line with the fact that the common factors are orthogonal. The variance decomposition was examined to determine the number of factors, and the explanation criterion of around 80% was adopted as a threshold.

The estimation procedure for the model expressed in Equations 12 and 13 followed the two-step strategy proposed by Doz, Giannone, and Reichlin (2011) and Giannone, Reichlin, and Small (2008). As presented by Bai (2003), the unobserved factors can be consistently estimated by principal components. In the preparatory stage, the factors are estimated by principal component analysis (PCA). The first step itself consists of estimating the parameters ($\mathbf{\Lambda}$ and $\mathbf{\Psi}$ of the model (Equation 12) by ordinary least squares (OLS) regression with the principal components extracted from $\mathbf{x}_t$. The estimated matrices $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{\Psi}}$ are then plugged into the following step.

In the second step, using the estimated values of $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{\Psi}}$ from the previous step, the Kalman smoother is applied to extract the factors, re-estimating $\mathbf{f}_t$ as a state variable of the dynamic factor model in Equations 12 and 13. This description of the two-step estimation procedure refers to the model specified for the Google Trends series. However, it is still necessary to incorporate the series of the total unemployed people that will be jointly estimated. Therefore, as detailed in Subsection 5.4, the dynamic factor model must be combined with the one specified for the survey data. Thus, the reported second step is implemented in the combined model using the R packages nowcasting (Valk; Mattos; Ferreira, 2019) and dlm (Petris, 2010).

The consistency of the two-step estimator has been proven for the stationary framework (Doz; Giannone; Reichlin, 2011) and the non-stationary case (Barigozzi; Luciani, 2017). Equation 13 for $\mathbf{f}_t$ implies the assumption that the factors are $I(1)$. On the other hand, R_t in Equation 4 is also $I(1)$, and to verify the co-integration of the series, the factors and the change in unemployment must have the same order of integration.

Therefore, some procedures are required to treat and select the Google Trends series. First, the Google Trends series can present considerable noise, with some points identified as outliers. For example, many search terms experienced a substantial decline in March 2020 due to the Covid-19 pandemic. Tests to detect outliers were performed, and those identified were replaced by linear interpolation using procedures available in the forecast package (Hyndman; Khandakar, 2008). In addition, the series must be $I(1)$ since the factors must be linked to the trend component of the BLFS. Therefore, the seasonal components were subsequently removed from the Google Trends series using

the seas package (Toews; Whitfield; Allen, 2007). In addition, augmented Dickey-Fuller (ADF) stationarity tests were carried out to determine the order of integration of the series via the urca package (Pfaff, 2008). It included the constant term plus trend and the Bayesian information criterion (BIC) to select the lag length. This test was adopted given that outliers were removed. Otherwise, other tests could be used, as presented by Reisen *et al*. (2017). Google Trend series $I(1)$ were selected as potential auxiliary variables in the dynamic factor model. Also, a test was conducted on the panel time series in which the null hypothesis is that all series had a unit root as proposed by Palm, Smeekes, and Urbain (2011) and implemented in the package bootUR (Smeekes; Wilms, 2023). Finally, cases of highly correlated series were identified due to slight differences in query terms, such as “vaga de emprego” and “vaga emprego” (job vacancy). In cases like these, where the differences in strings were due to a preposition or the order of words, only one corresponding series was retained.

Furthermore, Bai and Ng (2008) recommended the use of a “targeting the predictors” procedure, which means a pre-selection variable stage before conducting the PCA to improve forecasts. The authors argued that having more data to extract factors is not always better. According to Bai and Ng (2008, p. 306), “when the data are too noisy, we can be better off throwing away some data even though they are available”. This is because adding a series without impact on the factors deteriorates the estimation of the factors and loadings in PCA. It happens because the PCA assigns a non-zero weight when calculating the estimated factor as a weighted mean (Schiavoni *et al*., 2020). In addition, Zamprogno *et al*. (2020) also proposed a preselection step of series before conducting PCA by adjusting a multivariate linear model, as the presence of correlated series increases the variability of the principal components. Since the classic PCA was used only as a starting point for the first step of the estimation procedure, the adjustment of a linear model was not adopted before fitting the dynamic factor model. However, three different strategies were implemented to select the series related to the variable of interest, described in the following section.

5.3 Targeting the predictors

The motivation for this analytical step is to improve the factor model by incorporating only those series that can contribute to the factors in the PCA analysis. The inclusion of a step as “targeting the predictors” was introduced by Bai and Ng (2008) and Schiavoni *et al*. (2020) applied an elastic net specifically for SUTSE modelling. This paper tests two other strategies in addition to the proposed elastic net. The first uses time series clustering to identify a group of series with a shape similar to the slope R_t of the unemployment series. The other fits bivariate structural models to check series that potentially would have common trends with the unemployment series.

5.3.1 Elastic net

Bai and Ng (2008) proposed the use of the elastic net, as introduced by Hui and Hastie (2005). The method consists of a penalised regression that sets the coefficients of irrelevant series to a value close to or equal to zero. In this case, the strategy is to estimate a regression of the first difference of the slope $\Delta\hat{R}_t^y$ with the first difference of the selected series $I(1)$ standardised to mean zero and variance one. The first difference in the slope is obtained by estimating a univariate structural model for the unemployment series. The series with non-zero coefficients in the elastic net are selected for inclusion in the PCA estimation. The tuning parameters of the elastic net are selected by a grid search to minimise the root mean squared error (RMSE) and significant correlation coefficient estimated in the dynamic factor model. This estimation was carried out using the glmnet package (Friedman; Hastie; Tibshirani, 2010).

5.3.2 Clustering of Time Series

Bai and Ng (2008) found that using a targeting procedure is more effective than using all available auxiliary variables directly in PCA. More specifically, they concluded that using thresholding methods, such as least absolute shrinkage selection operator (LASSO), least angle regression (LARS), or elastic net (EN), was effective in selecting predictors for forecasting economic time series. Besides using EN, this study proposes testing a clustering procedure as an alternative method.

Cluster analysis can reduce large datasets by identifying groups that are not known a priori, have internal cohesion, and are mutually heterogeneous (Mingoti, 2005). This constitutes an unsupervised learning tool focused on describing patterns and associations between variables; it does not distinguish between response and predictor variables (Morettin; Singer, 2022). Thus, this strategy can handle the series of total unemployment (first difference of slope) and the time series from Google Trends (also in first difference).

As in any cluster analysis, it is essential to consider a measure of distance. In the present case, the distance measure must be suitable for time series. Various research fields seek to recognise patterns within time series, such as speech recognition, finance with stock prices, genetic sequencing and brain activity in medicine, among others (Berndt; Clifford, 1994; Aghabozorgi; Shirkhorshidi; Wah, 2015). Dynamic time warping (DTW) distance is a type of shape-based time-series clustering that takes into account distances between different points in time subject to specific constraints. One of the initial studies of DTW was carried out by Hiroaki and Seibi (1978) for speech recognition. The idea is to calculate the distance from a point t of a series to another series. However, in addition to considering

the distance at the same time t (which would be a Euclidean distance), times other than t are also considered, for example, $t-2, t-1, t+1, t+2$. The objective is to find an optimal path that minimises these distances. Once this path has been identified for each pair of series, the similarity matrix is established, and the standard cluster analysis procedures for cross-section data is performed. Aghabozorgi, Shirkhorshidi, and Wah (2015) highlighted that DTW provides elastic measures, which permit distances of one-to-many and one-to-one, in addition to effectively handling drifted time series. Also, it is one of the most popular similarity measures in time-series clustering and is more accurate than the Euclidean distance.

Consider two times series $\Delta\hat{R}_t^y$ and $\Delta\mathbf{x}_{t\prime}$ both with size T arranged to form a T-by- T' plan, where each grid point (t, t') corresponds to an alignment between the elements $\Delta\hat{R}_t^y$and $\Delta\mathbf{x}_{t\prime}$. Therefore, let $t = 1, \ldots, T$ index the $\Delta\hat{R}_t^y$ elements whereas $t' = 1, \ldots, T'$ is used for those in $\Delta\mathbf{x}_{t\prime}$. The non-negative function $d(t, t')$ represents a local dissimilarity function commonly assumed to be a Euclidean distance, and it is applied in each point each grid point (t, t'). $\mathbf{D}$ denotes the cross-distance matrix between $\Delta\hat{R}_t^y$ and $\Delta\mathbf{x}_{t\prime}$, also called the local cost matrix (LCM).

Using the grid resulting from $d(t, t')$, it is possible to minimise the alignment between $\Delta\hat{R}_t^y$ and $\Delta\mathbf{x}_{t\prime}$ by iteratively applying the LCM ($\mathbf{D}$). It starts at $d(1,1)$ and finishes at $d(T, T')$. Within $\varphi = (1,1), \ldots, (T, T')$, $\varphi(\kappa)$ is the optimum path defined by the warping curve. It is possible to compute the average accumulated distance to obtain a summary measure:

$$d_\varphi\left(\Delta\hat{R}_t^y, \Delta\mathbf{x}_{t\prime}\right) = \sum_{\kappa=1}^{T^*} \frac{d\left(\varphi(\kappa)\right) m_\varphi(\kappa)}{M_\varphi} \tag{14}$$

where m_φ is the weighting coefficient and M_φ is the corresponding normalisation constant to ensure that the accumulated distortions will be comparable along different paths. Therefore, the idea behind DTW is to find the optimal alignment $\varphi(\kappa)$ that minimises the average accumulated distance such that:

$$DTW\left(\Delta\hat{R}_t^y, \Delta\mathbf{x}_{t\prime}\right) = \min_\varphi d_\varphi\left(\Delta\hat{R}_t^y, \Delta\mathbf{x}_{t\prime}\right) \tag{15}$$

Giorgino (2009) mentioned that Equation 15 can be solved by dynamic programming. See Rabiner and Biing-Hwang (1993, sec. 4.7), Sardá-Espinosa (2019, p. 5-6) and Giorgino (2009, p. 2-3) for more details.

The time series clustering was implemented using the package dtwclust (Sardá-Espinosa, 2022). The similarity matrix was constructed with a set of optimal accumulative distances. The hierarchical algorithm using complete linkage (farthest neighbour method) as the similarity measure between groups was used to delimit the clusters. Different numbers of clusters were tested and combined with the significant correlation coefficient of the dynamic factor model. The Google Trend series classified in the same group of the unemployment figures were considered for the next stage.

5.3.3 Bivariate structural model

In order to include only relevant Google Trend series in the PCA procedure, one can select only Google Trend series that are individually related to the series of interest. Therefore, this paper proposes to jointly model each Google Trends series with the series of interest (the unemployment series) in a bivariate structural time series model as a procedure to select the relevant Google Trends series. Bivariate structural models can be fitted to determine whether slope disturbances of a particular Google Trends series and the unemployment series are correlated. The Google Trends series that show significant correlation are considered in the next estimation step. Different cutoffs of correlation values were tested and combined with significant correlation coefficients from dynamic factor models. The corresponding bivariate model can be expressed as follows:

$$\begin{pmatrix} \hat{y}_t \\ x_t \end{pmatrix} = \begin{pmatrix} \theta_t^y \\ \theta_t^x \end{pmatrix} + \begin{pmatrix} e_t \\ 0 \end{pmatrix} \tag{16}$$

where θ_t^y follows the same specification as in Equation 2 and $\theta_t^x = L_t^x + I_t^x$, with L_t^x the smooth trend model, which is similarly defined as for y in Equations 3 and 4. Model 16 allows for a non-zero correlation between the slope disturbance terms of L_t^x and L_t^y. Google Trends series with a correlation that is significantly different from zero are selected for PCA. It is understood that θ_t^x does not contain the seasonal component, since Google Trends are seasonally adjusted during data preparation. Also, series selected using elastic net or time series clustering procedures can still be submitted to the present strategy.

5.4 Modelling series from different data sources

Models for time series from different sources have been examined by Harvey and Chia-Hui (2000), who jointly modelled the UK series of unemployment and benefit claims. Section 5.1 described the issues involved when modelling time series from repeated sample surveys, whereas Section 5.2 focused on the challenges of modelling a broad set of series and corresponding techniques for dimension reduction. Since this study aims to produce monthly estimates of the total number of unemployed individuals, the survey data and Google Trends series are modelled together, allowing for the interaction between possible common components and borrowing strength between series. Thus, the models are combined and written as a signal extraction model:

$$\mathbf{z}_t = \begin{pmatrix} \hat{y}_t \\ \mathbf{x}_t \end{pmatrix} = \begin{pmatrix} \theta_t \\ \widehat{\mathbf{\Lambda}}\mathbf{f}_t \end{pmatrix} + \begin{pmatrix} e_t \\ \boldsymbol{\xi}_t \end{pmatrix} \tag{17}$$

The state-space formulation[5] is given by:

$$\mathbf{z}_t = \mathbf{H}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \tag{18}$$

$$\boldsymbol{\alpha}_t = \mathbf{G}\boldsymbol{\alpha}_{t-1} + \mathbf{W}\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \qquad t = 1, \ldots, T \tag{19}$$

The link across the series is expressed via the correlation between the slope disturbance term $\eta_{R,t}$ with the factors' disturbance term $u_{i,t}$ with $i = 1, \cdots, r$. The covariance matrix includes the term $\text{cov}(\eta_{R,t}, u_{i,t}) = \rho_{R,f_i} \cdot \sigma_{R,t}$ since $\sigma_{f_i,t} = 1$ with ρ_{R,f_i} indicating the correlation between the slope disturbance term of the unemployment series and the factor disturbance of the selected Google Trends series. The model assuming non-zero correlation combines the strength of two series (survey and big data), which can improve the precision of the estimates for the unemployment figures. In summary, as pointed out by Durbin and Koopman (2012, sec. 3.10.3), the goal is to use auxiliary information $\mathbf{x}_t$ to improve the precision of the estimates $\hat{y}_t$ via a model-based approach.

The variances and covariances of the disturbance terms of the state components are estimated by maximum likelihood. Then, a Kalman filter is applied to obtain optimal estimates for the unobserved components in the state vector and their corresponding variances. For details about Kalman filtering, see Harvey (1989, sec. 3.1) and Durbin and Koopman (2012, sec. 4.3). The package dlm (Petris, 2010) was used to estimate the unobserved components and corresponding hyperparameters.

The state vector encompasses the level, slope, seasonal components and sampling errors of the unemployment series and Google Trends series factors, as presented in Equation 22. Filtered estimates of the structural components' figures, as well as the signal and precision estimates presented as coefficients of variation (CV), are the main results of the modelling procedure (model-based estimates).

Another relevant aspect is the production of nowcast estimates. The Google Trends series are available one month before the single-month unemployment estimates. Following Harvey (1989), the Kalman filter forecasting step can be summarised as the addition of a missing information element at the end of the series. For the specific case of this paper, the problem can be understood as a case of delayed observation, where the deferred data are treated as missing information. Harvey (1989, sec. 3.4.7) mentioned options to deal with missing data, one is the use of smoothing equations to estimate the missing observation. In the univariate case, the updated estimate is considered a prediction, which implies the Kalman gain equals zero. However, in the context of the SUTSE framework, related series can be used to obtain preliminary estimates that correspond to nowcast

[5] A description of the matrices is presented in Appendix A.3.

estimates. According to Giannone, Reichlin, and Small (2008), nowcasting refers to forecasting for the current period or, as defined by Banbura, Giannone, and Reichlin (2010), is the prediction of the present. The Kalman updating expressions — presented in Harvey (1989) — can be rewritten to handle temporary missing data in the latest time period (T).

For model evaluation and selection, measures of in- and out-of-sample estimation accuracy are calculated based on relative mean squared error (RMSE) and relative mean squared forecast error (RMSFE) for the trend, slope, and signal, based on the univariate model for unemployment series as the baseline. In the case of MSFE, the last six observations are considered as out-of-sample, and the information set includes an additional single observation when generating successive forecasts, simulating the availability of data and the one-month delay in the unemployment totals. The model is re-estimated at every point in time, updating the estimates of hyperparameters and structural components. The set of selected Google Trend series is kept fixed.

In summary, six types of models, defined according to the strategies for targeting the predictors (elastic net, time series clustering and bivariate model), are examined:

a) UN: univariate model (baseline);

b) DFM EL: dynamic factor model with elastic net;

c) DFM CL: dynamic factor model with time series clustering;

d) DFM BI: dynamic factor model with bivariate structural model;

e) DFM EL+BI: dynamic factor model with elastic net combined with the bivariate structural model;

f) DFM CL+BI: dynamic factor model with time series clustering combined with bivariate structural model.

6 RESULTS

The models were fitted to the series of the total number of unemployed persons covering the period from January 2012 to December 2021 for Brazil and the state of Minas Gerais. In addition, the series of youth unemployment (people between 18 and 29 years old) was also analysed. There were not enough Google Trends series integrated of order one ($I(1)$) related to job searching for Roraima (as presented in Table 1), so the proposed models are inadequate for this case, and only Minas Gerais state-level series were analysed. Table 1 presents the number of words considered in each dataset and treatment stage, according to the process outlined in Figure 1.

Table 1: Number of words considered in the series selection and treatment stages before the targeting procedure

Stages	Brazil	Minas Gerais	Roraima
Initial words	82	82	82
Related words	1964	1246	176
Most searched words	527	560	112
After removing inadequate words	448	430	85
I(1)	259	116	5
Final words	222	94	2

Source: Elaborated by the authors.

The initial set of terms contained 82 words. This was expanded with single related terms, totalling nearly 2,000 words for Brazil, roughly 1,200 for Minas Gerais and only 170 for Roraima. After using the ranking procedure, most of these series were disregarded when measuring their relevance in relation to the word “vaga” (vacancy). The number of words was reduced by removing the terms that were not directly associated with job searching. Since only the $I(1)$ series can be incorporated into the proposed models, the number of series decreased even further. Additionally, after identifying those that were highly correlated, only one of the correlated series was retained. In this case, the correlation concerned the similarity of the search terms, including spelling, word order, and the use of prepositions. The final number of words was 222 for Brazil, 94 for Minas Gerais and only 2 for Roraima.

The next stage involved implementing strategies to identify and target the key predictors. The penalised regressions using elastic net (EN) revealed 12 series with significant coefficients for Brazil and 3 for Minas Gerais. When analysing youth unemployment, 20 series were selected for Brazil and 12 for Minas Gerais. The use of time series clustering (CL) resulted in 16 series in the same group of the unemployment slope for Brazil and three terms for Minas Gerais. In the context of youth

unemployment, there were four at the national level and for Minas Gerais. Finally, using bivariate structural models (BI), correlations between slope disturbances were significant in 12 cases for the Brazilian workforce and 9 for the youth group. In the case of Minas Gerais, 12 were selected for the working age population[6] and 17 for youth unemployment. When elastic net and time series clustering were combined with bivariate structural models (EN+BI and CL+BI), the final number of words selected varied from 0 to 4. Tables 2 and 3 present the words selected in each case (in Portuguese).

The selected set for Brazil encompassed terms such as curriculum vitae, employment and vacancy with complements, along with specific job search services like “SINE” and “OLX”, among others. Additionally, the final list included some references to particular places. It is worth mentioning that among the three strategies, very few identical words were selected in more than one of them. For Minas Gerais, terms with complements such as employment, “SINE”, and vacancies also appeared. Furthermore, 47.1% of the words mentioned city names from the corresponding state or initials (“Contagem”, “BH”, and “Uberlandia”). The same conclusion applies to data from the young population. The Appendix A.4 presents the details of these results with the estimated coefficients, group clustering, and significance tests of the correlations of bivariate models.

[6] The working age population is defined as those aged 14 years or older.

Table 2: Selected words by targeting strategies - working age population - Brazil and Minas Gerais

Words	EN	CL	BI	EN+BI	CL+BI	Words	EN	CL	BI	EN+BI	CL+BI
Brazil											
Agência de emprego		X				O que é currículo			X		
Banda B empregos	X					Objetivo		X			
Blog do emprego DF		X			X	OLX Curitiba	X				
CIEE PE		X				OLX empregos Curitiba			X		
Classificados	X					OLX empregos RJ		X			
Como fazer currículo			X			OLX MS			X		
Contratando			X			OLX vagas de empregos		X			
Curriculo jovem aprendiz			X			Quero bolsa			X		
Currículo para primeiro emprego	X					SINE IDT			X		
Currículo pronto			X			Trabalhe conosco	X	X			
Curriculo word			X			Vagas		X			
Emprego OLX SP			X			Vagas de emprego Curitiba		X			
Emprego PE		X				Vagas de emprego na OLX		X			X
Empregos Campo Grande MS		X				vagas de emprego PE	X				
Estagios		X				Vagas de emprego SP		X			
Fazer currículo pelo celular			X			Vagas Indeed	X				
Infojobs entrar		X				Vagas RJ	X				
Meu primeiro emprego	X					Vagas SINE hoje	X				
Modelo currículo	X					Vagas SINE RJ	X				
Modelos currículo		X									
Minas Gerais											
Catho			X			SINE de contagem			X		
Concursos			X			SINE emprego			X		
Emprego Uberlandia	X					Vaga de	X				
Empregos Uberlandia			X			Vagas de emprego		X			
Fazer currículo		X				Vagas de emprego no SINE			X		
Procuro emprego		X				Vagas de emprego SINE			X		
SINE BH			X			Vagas de emprego SINE BH			X		
SINE contagem			X			Vagas SINE BH	X		X	X	
SINE de BH			X								

Source: Elaborated by the authors.

Table 3: Selected words by targeting strategies – youth population – Brazil and Minas Gerais

Words	EN	CL	BI	EN+BI	CL+BI	Words	EN	CL	BI	EN+BI	CL+BI
Brazil											
CIEE PE	X					Modelo currículo	X			X	
Classificados	X					O que é currículo			X		
Como fazer currículo			X			OLX Curitiba	X				
Contratando			X			OLX MS			X		
Currículo online pdf			X			SINE IDT		X			X
Currículo para primeiro emprego	X					Sucessor RH	X				
Currículo pronto			X			Trabalhe conosco	X				
Curriculo Word		X	X		X	Vaga de empregos	X			X	
Emprego BH	X					Vagas de emprego Fortaleza		X			X
Emprego Brasilia	X					Vagas de emprego indeed	X				
Emprego na OLX	X					Vagas de emprego PE	X				
Fazer currículo pdf			X			Vagas de empregos	X				
Fazer currículo pelo celular			X			Vagas emprego RJ	X				
Fazer um currículo	X					Vagas Rio	X				
Informe vagas PE	X					Vagas RJ	X				
Mais empregos		X				Vagas SINE hoje	X				
Minas Gerais											
Catho			X			Vaga de	X				
Catho empregos			X			Vagas	X				
Comunidade de emprego	X					Vagas contagem	X				
Concursos			X			Vagas de emprego no SINE	X		X	X	
Currículo para primeiro emprego		X				Vagas de emprego SINE	X		X	X	
Empregos Uberlandia	X		X	X		Vagas de emprego SINE BH			X		
SEST SENAT vagas	X					Vagas em BH	X				
SINE			X			Vagas emprego BH	X				
SINE BH			X			Vagas indeed BH	X				
SINE BH empregos		X				Vagas no SINE			X		
SINE Contagem			X			Vagas SINE		X	X		X
SINE de BH			X			Vagas SINE BH	X		X	X	
SINE de Contagem			X			Vagas SINE Contagem			X		
SINE emprego			X			Vagas.com.br		X			

Source: Elaborated by the authors.

Using the BLFS microdata from overlapping PSUs, the estimated value for the autoregressive parameter ϕ of Equation 11 was 0.41 for the working-age population unemployment at the national level and 0.38 for Minas Gerais. In the case of youth population data, $\hat{\phi}$ was also 0.41 for Brazil and 0.35 for Minas Gerais. These values were considered as known (fixed) in the modelling procedure. Also, the series of $\hat{y}_t$ and its standard error $\hat{c}_t$ were deemed exogenous and are the design-based estimates.

Table 4 presents the estimated hyperparameters of the models with one and two factors, as well as the different strategies used to target the predictors. The number of series selected by each targeting procedure is also indicated. In the case of EN, the selected series correspond to those that presented coefficients statistically different from zero. For CL, the series are those classified in the same group as the first difference of the unemployment series slope. For the bivariate model strategy (BI), the selected series showed significant correlations between the slope disturbance term of unemployment and the error term of the factors' equations. Furthermore, the last strategy (BI) was also combined with the other strategies, as mentioned before.

Table 4 also shows the results of the LR tests for correlation coefficients. The LR tests were performed to verify whether the hyperparameters of σ_R^2, σ_S^2 and σ_I^2 were significantly different from zero. Statistical evidence was found to reject the null hypothesis for nearly all cases. The only exception was the seasonality term. Therefore, the seasonal component was defined as deterministic (σ_S^2=0), and the other components as stochastic.

At the national level, the estimated correlation between the slope disturbances was high (>0.70, in absolute values) for the targeting strategy based on the elastic net with one factor and when using bivariate structural models with one and two factors. In this case, non-significant correlations are obtained using the clustering approach. Furthermore, there was no statistical evidence to reject the null hypothesis of zero correlations for most of Minas Gerais models. The only instance of significant correlation observed for this state occurred when applying the clustering model with two factors. This result suggests that the selected series displayed lower adherence at the state level, particularly concerning Minas Gerais.

In addition, six models that exploited the bivariate structural model, clustering and their combination as a targeting strategy for the Brazilian youth population (18 to 29 years old) showed a significant correlation with the slope disturbance term. This aligns with the expectation that young people are more likely to use the Google tool for job searches. None of the models presented significant correlations for Minas Gerais. For this reason, the remaining analysis is focused on models with significant values for the correlation. Table 5 presents the results of the models for youth unemployment.

Table 4: Estimated hyperparameters, by targeting strategy and number of factors – working age population unemployment – Brazil and Minas Gerais

Model	UN	DFM EN	DFM EN	DFM CL	DFM CL	DFM BI	DFM BI	DFM EN+BI	DFM EN+BI	DFM CL+BI	DFM CL+BI
Number of factors (r)		1	2	1	2	1	2	1	2	1	2
Brazil											
Number of series x		12	12	16	16	12	12	0[*]	0[*]	2[*]	2[*]
σ_R^2	7,728	15,534	27,805	12,391	9,979	16,812	28,630				
σ_I^2	1,514	2,127	4	2,709	2,803	3,145	3,742				
$\sigma_{\bar{e}}^2$	1.21	1.18	1.25	1.15	1.16	1.08	1.03				
ρ_{R,f_1}		0.78	0.79	0.68	0.34	0.77	0.76				
ρ_{R,f_2}			-0.39		0.46		-0.46				
$\rho_{R,f} = 0$ (p-value)		0.04	0.15	0.06	0.18	0.03	0.00				
Minas Gerais											
Number of series x		3	3	3	3	12	12	1[*]	1[*]	0[*]	0[*]
σ_R^2	60	157	333	114	768	81	80				
σ_I^2	1,614	1,633	1,812	1,594	1,625	1,728	1,724				
$\sigma_{\bar{e}}^2$	0.51	0.45	0.41	0.52	0.45	0.44	0.44				
ρ_{R,f_1}		0.75	0.17	-0.73	-0.67	0.45	0.44				
ρ_{R,f_2}			-0.91		-0.70		-0.06				
$\rho_{R,f} = 0$ (p-value)		0.14	0.06	0.31	0.02	0.18	0.41				

Source: Elaborated by the authors. (*) Unable to fit a model.

Table 5: Estimated hyperparameters, by targeting strategy and number of factors – youth unemployment – Brazil and Minas Gerais

Model	UN	DFM EN	DFM EN	DFM CL	DFM CL	DFM BI	DFM BI	DFM EN+BI	DFM EN+BI	DFM CL+BI	DFM CL+BI
Number of factors (r)		1	2	1	2	1	2	1	2	1	2
Brazil											
Number of series x		20	20	4	4	9	9	2[*]	2[*]	3	3
σ_R^2	1,073	2,525	2,255	5,112	8,655	3,099	6,996			4,945	6,010
σ_I^2	2,131	2,738	3,380	2,457	2,505	2,374	2,491			2,464	2,893
$\sigma_{\tilde{e}}^2$	1.34	1.31	1.25	1.16	1.15	1.25	1.18			1.17	1.15
ρ_{R,f_1}		0.82	0.75	-0.91	-0.74	0.83	0.61			0.90	0.73
ρ_{R,f_2}			0.28		-0.59		-0.72				-0.57
$\rho_{R,f} = 0$ (p-value)		0.17	0.18	0.00	0.01	0.02	0.01			0.01	0.00
Minas Gerais											
Number of series x		12	12	4	4	17	17	4	4	1[*]	1[*]
σ_R^2	11	13	39	19	70	13	18	13	24		
σ_I^2	600	609	577	592	522	606	611	612	616		
$\sigma_{\tilde{e}}^2$	0.66	0.64	0.64	0.67	0.69	0.64	0.63	0.64	0.63		
ρ_{R,f_1}		0.26	-0.05	-0.74	-0.38	0.26	0.30	0.35	0.08		
ρ_{R,f_2}			0.84		0.83		0.49		0.73		
$\rho_{R,f} = 0$ (p-value)		0.56	0.39	0.27	0.50	0.46	0.66	0.58	0.42		

Source: Elaborated by the authors. (*) Unable to fit a model.

Table 6 shows the results of relative mean squared error (RMSE) and relative mean squared forecast error (RMSFE) divided by the RMSE and RSMFE of the univariate model. As most values in Table 5 are smaller than one, it follows that most proposed models outperform the univariate model for the trend, slope, and signal state variable components. The best model in terms of MSE is the elastic net model with one factor. This model yields gains of approximately 4.9% for trend, 7.4% for slope, and 2.2% for signal, compared to the univariate model (UN). Regarding the MSFE, the EN also presented the best reductions compared to the UN model. In the case of Minas Gerais, the single selected model presented MSE gain values around 9.9% for the trend, 8.3% for the signal and, 11.4% for the trend and 10.6% for the signal in nowcasting accuracy. In most cases, models with two factors perform better than models with only one factor. Since the two factors are significantly different from zero jointly, both bring information about the unemployment figures.

Table 6: Accuracy measures for selected unobservable components – working age and youth population – Brazil and Minas Gerais

Domain	Model	r	MSE			MSFE		
			L	*R*	*θ*	*L*	*R*	*θ*
Brazil	DFM EN*	1	0.9514	0.9255	0.9784	0.9392	0.8985	0.9456
Brazil	DFM CL	1	0.9674	0.9647	1.0015	0.9661	0.9359	0.9707
Brazil	DFM CL	2	0.9535	0.9451	0.9705	0.9464	0.9256	0.9506
Minas Gerais	DFM CL*	2	0.9014	0.9804	0.9172	0.8860	0.9449	0.8935
Brazil (young people)	DFM CL	1	0.9122	1.1409	0.9029	0.9092	1.0600	0.9092
Brazil (young people)	DFM CL	2	0.8909	1.0987	0.8911	0.8782	0.9781	0.8813
Brazil (young people)	DFM BI	1	0.9383	0.9573	0.9427	0.9390	0.9585	0.9399
Brazil (young people)	DFM BI	2	0.8939	0.9953	0.9036	0.8870	0.9434	0.8903
Brazil (young people)	DFM CL+BI	1	0.9150	1.1458	0.9051	0.9132	1.0683	0.9129
Brazil (young people)	DFM CL+BI*	2	0.8946	1.0072	0.8993	0.8753	0.9232	0.8795

Source: Elaborated by the authors. (*) Indicates the selected model.

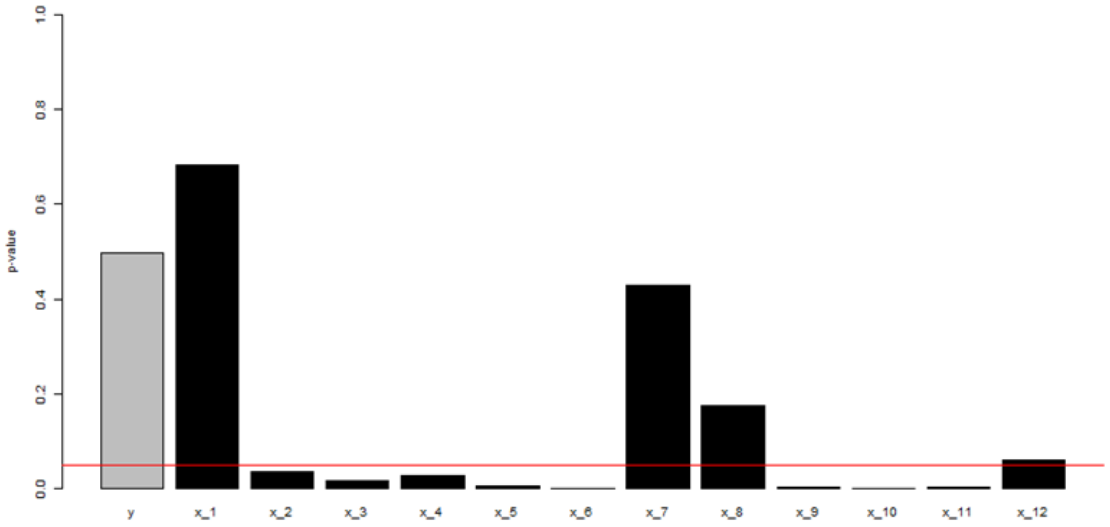
For the case of youth unemployment, the RMSE estimates of the model with CL targeting strategy, relative to the univariate model, were 0.8909 for trend, 1.0987 for slope, and 0.8911 for signal. This demonstrates that the univariate model had better accuracy for the slope. On the other hand, the proposed model surpassed it with gains of around 11% for the trend and signal. However, in the model combining the clustering and the bivariate model, the results were close, and the impact on the slope remained practically constant. For this reason, the model that combines the two strategies

was chosen. In this case, the RMSFE reached 0.8753 for trend, 0.9232 for slope, and 0.8795 for signal. This model differed from all others because it did not exhibit the highest gains exactly where the connection between the survey and Google Trends series was expressed (via the correlation of slope terms), as occured in all selected Brazilian models. As shown previously, the largest gains in RMSE and RMSFE were found in the slope and signal components for Minas Gerais and Brazilian youth unemployment (Table 6). Schiavoni *et al*. (2020) estimated the correlation between Dutch unemployment slope disturbances and factors from Google Trends words. They also detected superior accuracy gains for slope compared to signal and trend using the elastic net.

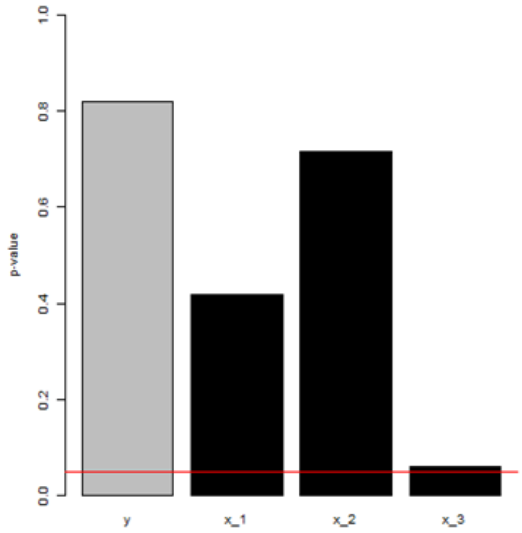
Regarding diagnostics, Graphic 2 presents the results of the individual normality tests for the standardised residuals of the chosen models.

Graphic 2: Shapiro-Wilk test p-values for standardised residuals of working age population and youth unemployment and limit of 5% significance - selected models

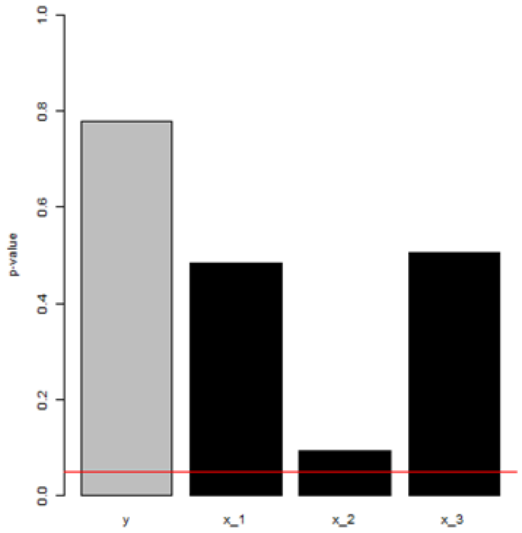
Working age population – Brazil (DFM EN)



Youth population – Brazil (DFM CL)



Working age population – Minas Gerais (DFM CL+BI)



Source: Elaborated by the authors.
Note: In grey are the p-values for the standardised residuals of the total unemployment model, and in black are the p-values for the standardised residuals of the Google Trends series.

The p-values indicated no evidence against the null hypothesis of normality of the residuals for the series of total unemployment and most of the idiosyncratic components of Google Trends at a 5% significance level (Graphic 2). This suggests that the model is correctly specified despite some Brazilian working-age population series not presenting a p-value greater than 5%. Schiavoni *et al*. (2020) claimed that an occurence of such results does not render the model's findings unfeasible. Furthermore, even dropping the normality assumption, ($\hat{\alpha}_t$) is still an optimal estimator within the class of all linear estimators since it minimised the MSE (Harvey, 1989, p. 105).

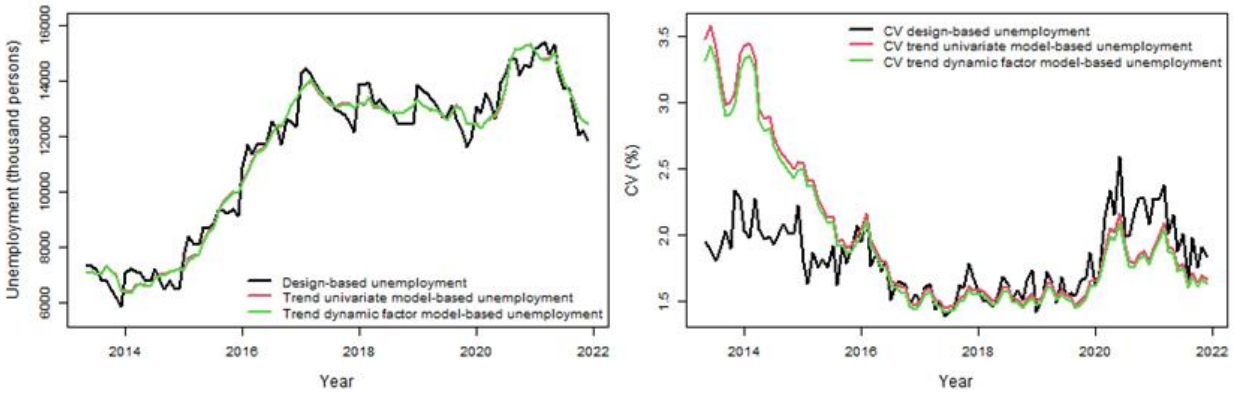
The estimates for the working age and youth population unemployment model are displayed in Graphic 2. The left panels of the Figures present a comparison of the filtered trend estimates obtained based on the univariate model and the selected dynamic factor model (DFM EN with one factor for Brazilian working age population, DFM CL + BI with two factors for Brazilian youth population, and DFM CL with two factors for Minas Gerais' working age population). Overall, the estimates from both time series models are quite similar throughout the entire period. It should be noted that the first 16 observations are excluded due to the Kalman filter initialisation process. The series of design-based estimates is also exhibited in the graph, demonstrating its noisier behaviour over the months.

The right panels in Graphic 3 present the coefficients of variation (CV) for the three estimates exhibited in the first graph. The design-based national estimates presented a good precision level, so the precision of model-based estimates only surpassed the design-based ones after accumulating information over more than 24 months. The loss in precision during the pandemic period is worth

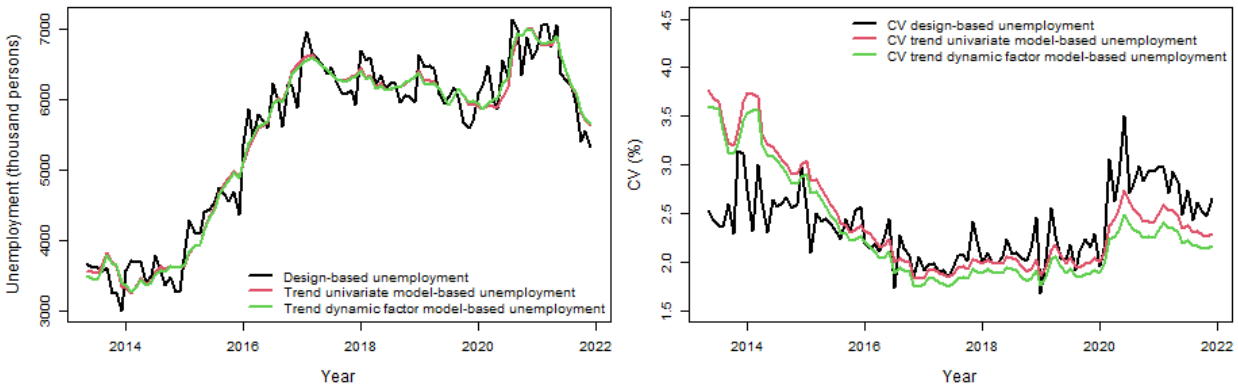
mentioning when the CV of design-based estimates exceeded 2% for Brazil. The differences between the univariate and dynamic factor models are not large. Nevertheless, there is an advantage to the models that incorporated the Google Trends data, particularly for the smaller domains, i.e., the youth population of Brazil and the working-age population of Minas Gerais.

Graphic 3: Design-based and model-based estimates for working age and youth population unemployment and selected Google Trends series – Brazil and Minas Gerais

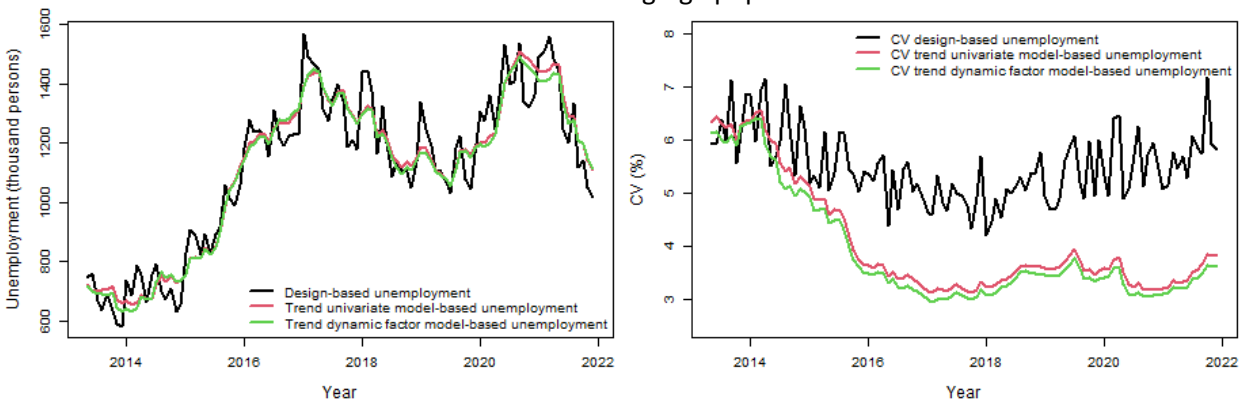
Brazil – working age population



Brazil – youth population



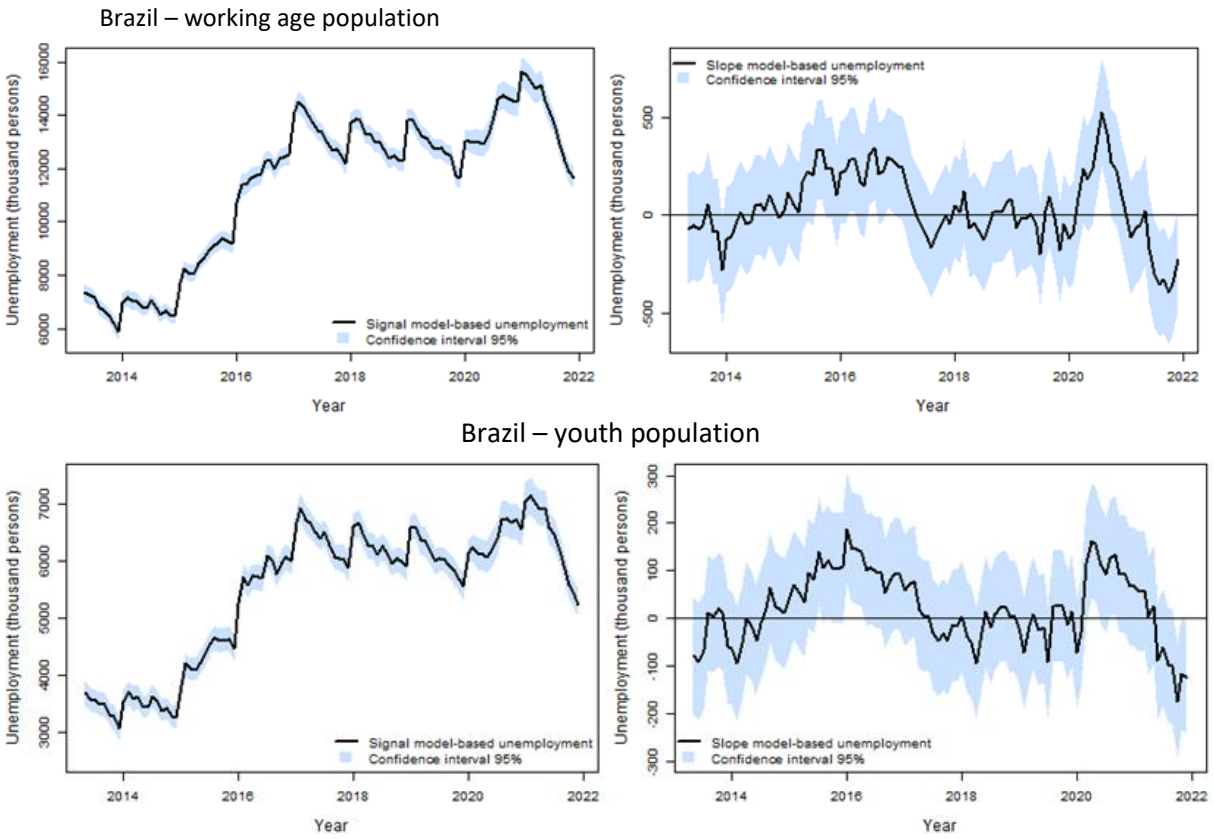
Minas Gerais – working-age population



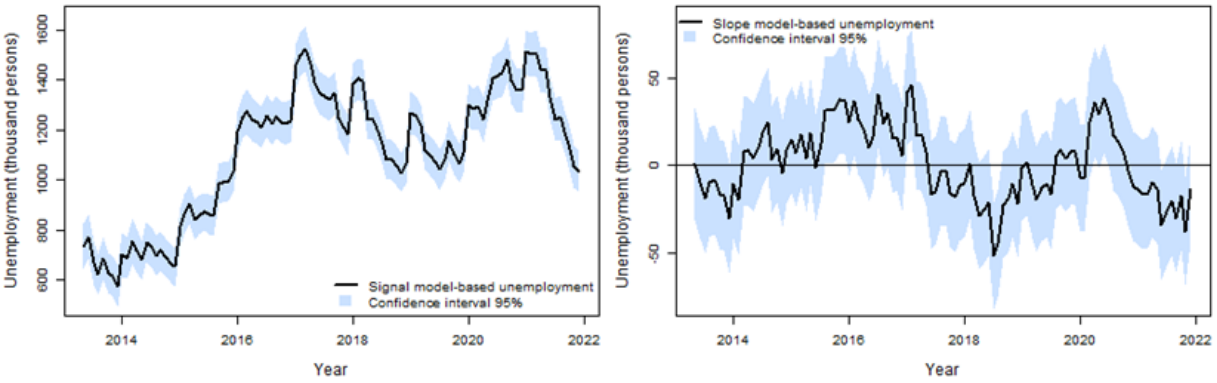
Source: Elaborated by the authors.

Comparison of the series derived from univariate and dynamic factor models indicated that the second presented lower CV values over time, resulting in smaller confidence intervals for signal and slope estimates. In Graphic 4, the filtered estimates for total unemployment obtained using the dynamic factor model for signal and slope, along with their 95% confidence intervals, are presented for the working age and youth population unemployment of Brazil and Minas Gerais. The slope estimates can also be interpreted as month-to-month changes in unemployment, indicating whether significant movements occurred in the short term. For some months, the 95% confidence interval does not include the value zero, providing evidence of significant month-to-month changes in unemployment figures at the national level and Minas Gerais state, for both working age and youth populations.

Graphic 4: Dynamic factor model signal and slope filtered estimates for working age and youth population unemployment, and corresponding 95% confidence intervals - Brazil and Minas Gerais



Minas Gerais – working age population

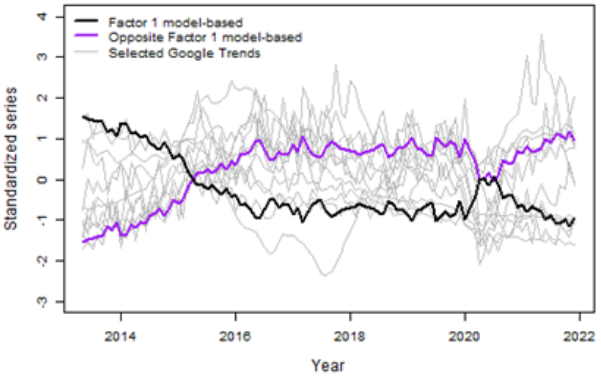


Source: Elaborated by the authors.

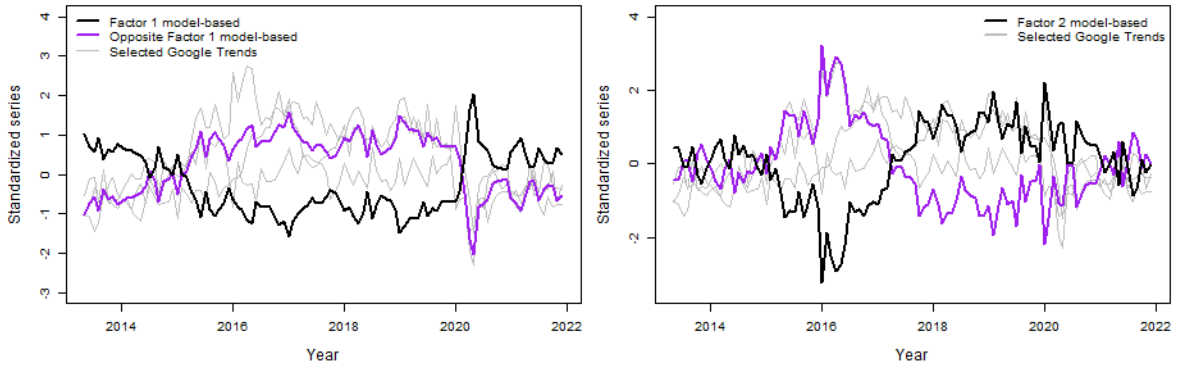
Graphic 5 shows standardised versions of the Google Trends factors estimated via the dynamic factor models. In addition, they present the selected Google Trends series incorporated into the multivariate model. In some cases, the factors are graphically represented by their opposite (-factor value) to highlight their association with the selected Google Trends series.

Graphic 5: Factor model-based estimates for working age and youth population unemployment and selected Google Trends series – Brazil and Minas Gerais

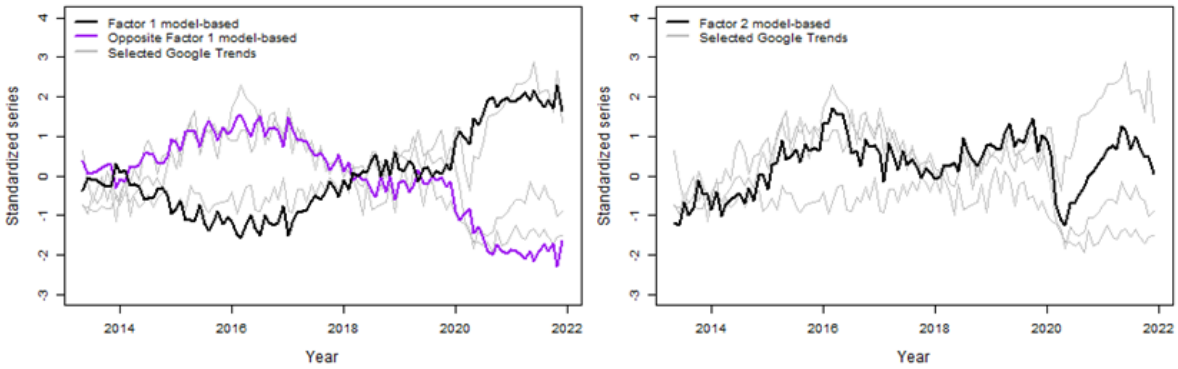
Brazil – working age population



Brazil – youth population



Minas Gerais – working-age population



Source: Elaborated by the authors.

Figure 2 presents a word cloud of selected terms that were taken into account when applying the targeting procedures for unemployment in Brazil. The size represents the ranking measure constructed comparatively with the word “vaga”, and the 12 terms used in the chosen dynamic factor model are those highlighted in red. Unexpectedly, popular search terms did not compose time series related to the evolution (trend variation) of unemployment in Brazil.

Figure 2: Selected terms represented by relative ranking

Source: Elaborated by the authors. Note: The word size represents a relative measure of search ranking, and the coloured highlight (in red) indicates the series incorporated in the dynamic factor model using a bivariate structural model to target the predictors.

7 CONCLUSIONS

This study investigated the potential for producing precise estimates of monthly unemployment figures by integrating data from repeated surveys with big data, as well as explored the feasibility of generating nowcast estimates. The results provide favourable evidence, presenting estimates with good precision, borrowing strength from Google Trends series and producing nowcast estimates with higher precision than the univariate models.

Gonçalves *et al*. (2022) had already indicated that univariate models are suitable for producing monthly unemployment statistics for Brazil and its states. Gonçalves (2023) provides evidence of significant short-term changes in the total unemployment indicator, confirming the need to produce single-month estimates for better monitor Brazilian labour market dynamics. It is also important to note that national and state-level dynamics differ, therefore state-level single-month estimates are also required. Furthermore, Gonçalves *et al*. (2022) have indicated that model-based estimation procedures incorporating claimant count auxiliary data do not significantly improve the precision of estimates in the Brazilian case. This finding contrasts with what is suggested international literature.

Some challenges were encountered in utilizing Google Trends series to take advantage of common trajectories. The identification of words that internet job seekers may use was derived from an initial list created based on the most anticipated searches, as indicated in the literature. However, a question remains whether all relevant words were considered. Therefore, the initial list of words was expanded with related terms. After that, a ranking in relation to a relevant word was performed to filter the most searched ones in the list. In the case of Brazil, the selected words showed national relevance, but for each state, more specific words, such as agency names or job search services, emerged. It is interesting to note that the words presented in the leading positions of the ranking were not those that were actually incorporated in the models. This showed that analysing only those most sought-after terms is not a valid strategy. Hence, carrying out studies with single and most searched words, as presented in the initial papers using Google Trends series (Hyunyoung; Varian, 2009a, 2009b, 2012) and others, e.g., D'Amuri (2009), D'Amuri and Marcucci (2010), Naccarato *et al*. (2018) and Simionescu (2020), is not advantageous, hence more effort has to be devoted to word selection.

The present study developed a sequence of procedures to select and handle the Google Trends series. However, another sequence of steps, such as a longer or shorter list of proposed initial words, could generate another selected series and, consequently, other factors. Additionally, an extension of the model to incorporate the lags of f_t could be tested, as well as observable components

as indicated in the Factor-Augmented Vector Autoregressive (FAVAR) approach (Bernanke; Boivin; Eliasz, 2005).

Alternative strategies for targeting the predictor were tested to solve this sensitivity challenge by introducing a step that selected only the words that contributed to the factors. The elastic net proved to be the best strategy for targeting the auxiliary series when modelling national unemployment. However, for Minas Gerais, using clustering as a targeting strategy resulted in the best model. For Brazilian youth unemployment, better results were found with the combination of clustering and bivariate strategy. Therefore, it cannot be concluded that there is one ultimate targeting strategy.

At the state level, only one model demonstrated the benefits of modelling the Google Trends series in conjunction with the unemployment series. Future studies are needed to further investigate the selected Google Trends series and their quality, as improvements in precision are essential at this level. The selection of words for Minas Gerais was carried out separately from the selection process for Brazil. In future studies, it may be beneficial to disregard the regional selection and use the same series for all geographical levels.

In the case of the young unemployed individuals, this group tends to use the internet more frequently. Therefore, it was expected that the youth unemployment series would align more closely with the behaviour of word searches on the Google Trends platform. However, the association was attained only for national-level statistics. Therefore, future studies could estimate unemployment for domains other than those defined only by age. It is noteworthy that studies of other nations focused on specific population groups (Fondeur; Karamé, 2013; Naccarato *et al*., 2018; Dilmaghani, 2019) have taken advantage of the Google Trends series for their proposed objectives.

After performing word selection and testing strategies that effectively reduced the number of factors based on relevant information, the possibility of producing nowcast estimates was verified. The discussion here focused on checking the viability of nowcasting rather than finding the best nowcast estimate for unemployment. In this case, the advantage was achieved by using the same modelling type to produce both the single-month unemployed and the nowcast estimates. In addition, it illustrated the procedure for obtaining an estimate in case it is not possible to collect the survey data in a given month. The results showed that Google Trends series can be useful for producing nowcast estimates at the national and state level with less error than a univariate model.

Finally, using the Google Trends series in the context of producing official statistics required a lengthy investigation. However, it is necessary to provide evidence and enrich the discussion on whether or not to use big data for official statistics. Despite positive signs in favour of its use, the

benefits were limited for the Brazilian setting. In addition, in the context of small areas or domains (geographical and other population groups), where the incorporation of an alternative data source could be even more relevant, the results did not reveal a noteworthy improvement.

ACKNOWLEDGEMENTS

The views expressed in this paper are those of the authors and do not reflect the policies of the Brazilian Institute of Geography and Statistics (IBGE), João Pinheiro Foundation or Statistics Netherlands.

REFERENCES

AGHABOZORGI, Saeed; SHIRKHORSHIDI, Ali Seyed; WAH, Teh Ying. Time-series clustering: a decade review. **Information Systems**, Amsterdam, v. 53, p. 16-38, 2015. Disponível em: http://doi.org/10.1016/j.is.2015.04.007. Acesso em: 14 out. 2025.

ALABRELLA, Natacha Perez. **Uso de dados de busca na internet na estimação de indicadores econômicos**. 2017. 41 f. Dissertação (Mestrado em Economia) – Insper Instituto de Ensino e Pesquisa, São Paulo, 2017. Disponível em: https://repositorio.insper.edu.br/handle/11224/2284. Acesso em: 14 out. 2025.

ALTHOUSE, Benjamin M.; YIH, YNG NG; CUMMINGS, Derek A. T. Prediction of dengue incidence using search query surveillance. **PLOS Neglected Tropical Diseases**, Califórnia, v. 5, p. 1-7, 2011. Disponível em: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0001258. Acesso em: 14 out. 2025.

ANVIK, Christian; GJELSTAD, Kristoffer. **Just Google it!**: forecasting Norwegian unemployment figures with web queries. 2010. 75 f. Dissertação (Mestrado em Ciências) – Norwegian School of Management, Oslo, 2010. Disponível em: https://biopen.bi.no/bi-xmlui/bitstream/handle/11250/95460/2010_11_CREAM_wp.pdf?sequence=1&isAllowed=y. Acesso em: 15 out. 2025.

ASKITAS, Nikolaos; ZIMMERMANN, Klaus F. Google econometrics and unemployment forecasting. **Applied Economics Quarterly**, [*s.l.*], v. 55, n. 2, p. 107-120, 2009. Disponível em: https://doi.org/10.3790/aeq.55.2.107. Acesso em: 15 out. 2025.

BAI, Jushan. Inferential theory for factor models of large dimensions. **Econometrica**, Hoboken, v. 71, n. 1, p. 135-171, 2003. Disponível em: http://doi.org/10.1111/1468-0262.00392. Acesso em: 15 out. 2025.

BAI, Jushan; NG, Serena. Forecasting economic time series using targeted predictors. **Journal of Econometrics**, Amsterdam, v. 146, n. 2, p. 304-317, 2008. Disponível em: http://doi.org/10.1016/j.jeconom.2008.08.010. Acesso em: 15 out. 2025.

BAILAR, Barbara A. The effects of rotation group bias on estimates from panel surveys. **Journal of the American Statistical Association**, London, v. 70, n. 349, p. 23-30, 1975. Disponível em: http://doi.org/10.2307/2285370. Acesso em: 15 out. 2025.

BANBURA, Marta; GIANNONE, Domenico; REICHLIN, Lucrezia. Nowcasting. **ECB Working Paper**, Frankfurt am Main, n. 1275, 2010. Disponível em: http://dx.doi.org/10.2139/ssrn.1717887. Acesso em: 15 out. 2025.

BARIGOZZI, Matteo; LUCIANI, Matteo. Common factors, trends, and cycles in large datasets. Washington: Board of Governors of the Federal Reserve System, 2017. (Discussion series, 2017-111). Disponível em: https://doi.org/10.17016/FEDS.2017.111. Acesso em: 15 out. 2025.

BARREIRA, Nuno; GODINHO, Pedro; MELO, Manuel. Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. **Netnomics**, New York, v. 14, n. 3, p. 129-165, 2013. Disponível em: http://doi.org/10.1007/ s11066-013-9082-8. Acesso em: 15 out. 2025.

BERNANKE, Ben S.; BOIVIN, Jean; ELIASZ, Piotr. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. **The Quarterly Journal of Economics**, Oxford, v. 120, n. 1, p. 387-422, 2005. Disponível em: https://academic.oup.com/qje/article-abstract/120/1/387/1931468?redirectedFrom=fulltext. Acesso em: 15 out. 2025.

BERNDT, Donald J.; CLIFFORD, James. Using dynamic time warping to find patterns in time series. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 3., 1994, Seattle WA. **Proceedings** […]. Seattle WA: AAAI Press, 1994. p. 359-370. Disponível em: https://dl.acm.org/doi/10.5555/3000850.3000887. Acesso em: 15 out. 2025.

BINDER, David A.; DICK, Peter. Modelling and estimation for repeated surveys. **Survey Methodology**, Ottawa, v. 15, n. 1, p. 29-45, 1989. Disponível em:https://www150.statcan.gc.ca/n1/pub/12-001-x/1989001/article/14579-eng.pdf. Acesso em: 15 out. 2025.

BOONSTRA, Harm Jan; VAN DEN BRAKEL, Jan A. Multilevel time series models for small area estimation at different frequencies and domain levels. **The Annals of Applied Statistics**, v. 16, n. 4, p. 2314-2338, 2022.

BORUP, Daniel; SCHÜTTE, Erik Christian Montes. In search of a job: forecasting employment growth using Google Trends. **Journal of Business and Economic Statistics**, London, v. 40, n. 1, p. 186-200, 2022. Disponível em: http://doi.org/10.1080/07350015.2020.1791133. Acesso em: 15 out. 2025.

BULUT, Levent. Google Trends and the forecasting performance of exchange rate models. **Journal of Forecasting**, Hoboken, v. 36, n. 3, p. 303-315, 2017. Disponível em: http://doi.org/10.1002/for.2500. Acesso em: 15 out. 2025.

BUTLER, Declan. When Google got flu wrong. **Nature**, New York, n. 494, p. 155-156, 2013. Disponível em: http://doi.org/10.1038/494155a. Acesso em: 15 out. 2025.

CARRIÈRE-SWALLOW, Yan; LABBÉ, Felipe. Nowcasting with Google Trends in an emerging market. **Journal of Forecasting**, Hoboken, v. 32, n. 4, p. 289-298, 2011. Disponível em: http://doi.org/10.1002/for.1252. Acesso em: 15 out. 2025.

COCHRAN, William G. **Sampling Techniques**. 3 rd ed. New York: John Wiley & Sons, 1977 (Wiley Series in Probability and Mathematical Statistics).

COOK, Samantha *et al*. Assessing Google flu trends performance in the United States during the 2009 influenza virus a (h1n1) pandemic. **PLoS ONE**, Califórnia, v. 6, e23610, 2011. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023610. Acesso em: 15 out. 2025.

D'AMURI, Francesco. **Predicting unemployment in short samples with internet job search query data**. [*S.l.*]: MPRA, 2009. (Working paper, 18403). Disponível em: https://mpra.ub.uni-muenchen.de/18403/. Acesso em: 15 out. 2025.

D'AMURI, Francesco; MARCUCCI, Juri. 'Google It!' forecasting the US unemployment rate with a Google job search index. **FEEM Working paper series**, Milan, n. 31, 2010. Disponível em: http://doi.org/10.2139/ssrn.1594132. Acesso em: 16 out. 2025.

D'AMURI, Francesco; MARCUCCI, Juri. The predictive power of google searches in forecasting us unemployment. **International Journal of Forecasting**, Amsterdam, v. 33, p. 801-816, 2017. Disponível em: http://doi.org/10.1016/j.ijforecast.2017.03.004. Acesso em: 16 out. 2025.

DATTA, G. *et al*. Hierarchical Bayes estimation of unemployment rates for the states of the U.S. **Journal of the American Statistical Association**, London, v. 94, n. 448, p. 1074–1082, 1999.

DE WAAL, Ton; VAN DELDEN, Arnout; SCHOLTUS, Sander. Multi-source statistics: basic situations and methods. **International Statistical Review**, Hoboken, v. 88, n. 1, p. 203-228, 2020. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12352. Acesso em: 16 out. 2025.

DILMAGHANI, Maryam *et al*. Workopolis or The Pirate Bay: what does Google Trends say about the unemployment rate? **Journal of Economic Studies**, Leeds, v. 46, n. 2, p. 422-445, 2019. Disponível em: http://doi.org/10.1108/JES-11-2017-0346. Acesso em: 15 out. 2025.

DINGDONG, Yi *et al*. Forecasting unemployment using internet search data via PRISM. **Journal of the American Statistical Association**, London, v. 116, n. 536, p. 1662-1673, 2021. Disponível em: https://doi.org/10.1080/01621459.2021.1883436. Acesso em: 16 out. 2025.

DOZ, Catherine; FULEKY, Peter. Dynamic factor models. *In*: FULEKY, Peter (ed.). **Macroeconomic forecasting in the era of Big Data**: theory and practice. New York: Springer, 2020. p. 27-64.

DOZ, Catherine; GIANNONE, Domenico; REICHLIN, Lucrezia. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. **Journal of Econometrics**, Amsterdam, v. 164, n. 1, p. 188-205, 2011. Disponível em: http://doi.org/10.1016/j.jeconom.2011.02.012. Acesso em: 15 out. 2025.

DURBIN, James; KOOPMAN, Siem Jan. **Time series analysis by state space methods**. 2nd ed. Oxford: Oxford University Press, 2012. (Oxford statistical science series, 38).

DURBIN, James; QUENNEVILLE, B. Benchmarking by state space methods. **International Statistical Review**, Hoboken, v. 65, n. 1, p. 23-48, 1997.

ELLIOTT, D. J.; ZONG, P. Improving timeliness and accuracy of estimates from the UK labour force survey. **Statistical Theory and Related Fields**, London, v. 3, n. 2, p. 186-198, 2019. Disponível em: https://doi.org/10.1080/24754269.2019.1676034. Acesso em: 16 out. 2025.

FONDEUR, Y.; KARAMÉ, Frederic. Can google data help predict French youth unemployment? **Economic Modelling**, Amsterdam, v. 30, p. 117-125, 2013. Disponível em: http://doi.org/10.1016/j.econmod.2012.07.017. Acesso em: 16 out. 2025.

FREITAS, Marcos Paulo Soares de *et al*. **Amostra Mestra para o Sistema Integrado de Pesquisas Domiciliares**. Rio de Janeiro: IBGE, 2007. (Texto para discussão, n. 23). Disponível em: https://www.ibge.gov.br/arquivo/projetos/sipd/texto_discussao_23.pdf. Acesso em: 16 out. 2025.

FRIEDMAN, Jerome H.; HASTIE, Trevor; TIBSHIRANI, Rob. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, Innsbruck, v. 33, n. 1, p. 1-22, 2010. Disponível em: https://www.jstatsoft.org/v33/i01/. Acesso em: 16 out. 2025.

GIANNONE, Domenico; REICHLIN, Lucrezia; SMALL, David. Nowcasting: the real-time informational content of macroeconomic data. **Journal of Monetary Economics**, Amsterdam, v. 55, n. 4, p. 665-676, 2008. Disponível em: http://doi.org/10.1016/j.jmoneco.2008.05.010. Acesso em: 16 out. 2025.

GINSBERG, Jeremy *et al*. Detecting influenza epidemics using search engine query data. **Nature**, New York, n. 457, p. 1012-1014, 2009. Disponível em: http://doi.org/10.1038/nature07634. Acesso em: 16 out. 2025.

GIORGINO, Toni. Computing and visualizing dynamic time warping alignments in R: the DTW package. **Journal of Statistical Software**, Innsbruck, v. 31, n. 7, p. 1-24, 2009. Disponível em: https://doi.org/10.18637/jss.v031.i07. Acesso em: 16 out. 2025.

GONÇALVES, Caio César Soares. **Produção de indicadores do mercado de trabalho com modelos de séries temporais de pesquisas repetidas**. 2023. 249 f. Tese (Doutorado em População, Território e Estatísticas Públicas) – Escola Nacional de Ciências Estatísticas, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 2023. Disponível em: https://ence.ibge.gov.br/images/Teses/Tese_CaioGoncalves.pdf. Acesso em: 16 out. 2025.

GONÇALVES, Caio *et al*. Single-month unemployment rate estimates for the Brazilian labour force survey using state-space models. **Journal of the Royal Statistical Society**: series A statistics society, Oxford, v. 185, n. 4, p. 1707-1732, 2022. Disponível em: http://doi.org/10.1111/rssa.12914. Acesso em: 16 out. 2025.

GOOGLE. **Google Trends lessons**. Califórnia, 2022. Disponível em: https://newsinitiative.withgoogle.com/pt-br/resources/products/google-trends/. Acesso em: 24 nov. 2025.

GOOGLE. **Google Trends**. Califórnia, 2025. Disponível em: https://trends.google.com.br/home. Acesso em: 24 nov. 2025.

GROVES, Robert M. Three eras of survey research. **Public Opinion Quarterly**, Oxford, v. 75, n. 5, p. 861-871, 2011. Disponível em: https://www.uvm.edu/~dguber/POLS234/articles/groves.pdf. Acesso em: 16 out. 2025.

GUZMAN, Giselle. Internet search behavior as an economic forecasting tool: the case of inflation expectations. **The Journal of Economic and Social Measurement**, Thousand Oaks, v. 36, n. 3, 2011. Disponível em: https://journals.sagepub.com/doi/abs/10.3233/JEM-2011-0342. Acesso em: 16 out. 2025.

HAND, David J. Statistical challenges of administrative and transaction data. **Journal of the Royal Statistical Society**: series A statistics society, Oxford, v. 81, n. 3, p. 555-605, 2018. Disponível em: http://doi.org/10.1111/rssa.12315. Acesso em: 16 out. 2025.

HARVEY, Andrew. **Forecasting, structural time series models and the Kalman Filter**. Cambridge: Cambridge University Press, 1989.

HARVEY, Andrew; CHIA-HUI, Chung. Estimating the underlying change in unemployment in the UK. **Journal of the Royal Statistical Society**: series A statistics society, Oxford, v. 163, p. 303-309, 2000. Disponível em: http://doi.org/10.1111/1467-985x.00171. Acesso em: 16 out. 2025.

HASSANI, Hossein; SILVA, Emmanuel Sirimal. Forecasting energy data with a time lag into the future and Google trends. **International Journal of Energy and Statistics**, Hackensack, v. 4, n. 4, e1650020, 2016. Disponível em: http://doi.org/10.1142/S2335680416500204. Acesso em: 16 out. 2025.

HIROAKI, Sakoe; SEIBI, Chiba. Dynamic programming algorithm optimization for spoken word recognition. **IEEE Transactions on Acoustics, Speech and Signal Processing**, [*s.l.*], v. 26, n. 1, 1978. Disponível em: https://jeffe.cs.illinois.edu/teaching/compgeom/refs/Sakoe-Chiba-DTW.pdf. Acesso em: 16 out. 2025.

HUI, Zou; HASTIE, Trevor. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society**: series B statistical methodology, Hoboken, v. 67, n. 2, p. 301-320, 2005. Disponível em: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467- 9868.2005.00503.x. Acesso em: 16 out. 2025.

HYNDMAN, Rob J.; KHANDAKAR, Yeasmin. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, Innsbruck, v. 26, n. 3, p. 1-22, 2008. Disponível em: http://doi.org/10.18637/jss.v027.i03. Acesso em: 16 out. 2025.

HYUNYOUNG, Choi; VARIAN, Hal. **Predicting initial claims for unemployment benefits**. New York: [*s.n.*], 2009a. Disponível em: https://static.googleusercontent.com/media/research.google.com/pt-BR//archive/papers/initialclaimsUS.pdf. Acesso em: 15 out. 2025.

HYUNYOUNG, Choi; VARIAN, Hal. **Predicting the present with Google Trends**. Califórnia: [*s.n.*], 2009b. Disponível em: https://static.googleusercontent.com/media/www.google.com/pt-BR//googleblogs/pdfs/google_predicting_the_present.pdf. Acesso em: 15 out. 2025.

HYUNYOUNG, Choi; VARIAN, Hal. Predicting the present with Google Trends. **Economic Record**, Hoboken, v. 88, n. 1, p. 2-9, 2012. Disponível em: http://doi.org/10.1111/j.1475-4932.2012.00809.x. Acesso em: 15 out. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa nacional por amostra de domicílios contínua**: notas técnicas: versão 1.12. Rio de Janeiro: IBGE, 2022. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca- catalogo?view=detalhes&id=2101992. Acesso em: 16 out. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Sobre o processo de ponderação da PNAD Contínua**. Rio de Janeiro: IBGE, 2021. Disponível em: https:// biblioteca.ibge.gov.br/visualizacao/livros/liv101803.pdf. Acesso em: 16 out. 2025.

JIANZHENG, Liu *et al*. Rethinking big data: a review on the data quality and usage issues. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, v. 115, p. 134-142, 2016. Disponível em: https://www.sciencedirect.com/science/article/pii/ S0924271615002567. Acesso em: 15 out. 2025.

KRIEG, Sabine; VAN DEN BRAKEL, Jan. Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. **Computational Statistics & Data Analysis**, Amsterdam, v. 56, n. 10, p. 2918-2933, 2012. Disponível em: https://ideas.repec.org/a/eee/csdana/v56y2012i10p2918-2933.html. Acesso em: 16 out. 2025.

KRISTOUFEK, Ladislav; MOAT, Helen Susannah; PREIS, Tobias. Estimating suicide occurrence statistics using Google trends. **EPJ Data Science**, New York, v. 5, n. 32, 2016. Disponível em: http://doi.org/10.1140/epjds%2Fs13688-016-0094-0. Acesso em: 16 out. 2025.

LAZER, David *et al*. The parable of Google flu: traps in big data analysis. **Science**, Washington, v. 343, n. 6176, p. 1203-1205, 2014. Disponível em: http://doi.org/10.1126/science.1248506. Acesso em: 16 out. 2025.

LUI, Catherine; METAXAS, Panagiotis T.; MUSTAFARAJ, Eni. On the predictability of the us elections through search volume activity. *In*: PROCEEDINGS OF THE IADIS INTERNATIONAL CONFERENCE ON E-SOCIETY, 2011, Avila. **Proceedings** [...]. Wellesley: [*s.n.*], 2011. Disponível em: https://repository.wellesley.edu/ object/ir153. Acesso em: 16 out. 2025.

MAAS, Benedikt. Short-term forecasting of the us unemployment rate. **Journal of Forecasting**, Hoboken, v. 39, p. 394-411, 2020. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2630. Acesso em: 16 out. 2025.

MASSICOTTE, Philippe. **gtrendsR**: perform and display Google Trends queries. [*S.l.: s.n.*], 2022. Disponível em: https://cran.r-project.org/web/packages/gtrendsR/index.html. Acesso em: 16 0ut. 2025.

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

MORETTIN, Pedro Alberto; SINGER, Julio da Motta. **Estatística e Ciência de dados**. Rio de Janeiro: LTC, 2022.

MORSY, S. *et al*. Prediction of zika- confirmed cases in Brazil and Colombia using Google Trends. **Epidemiology and Infection**, Cambridge, v. 146, n. 13, p. 1625-1627, 2018. Disponível em: https://doi.org/10.1017/S0950268818002078. Acesso em: 16 out. 2025.

NACCARATO, Alessia *et al*. Combining official and Google trends data to forecast the Italian youth unemployment rate. **Technological Forecasting and Social Change**, Amsterdam, v. 130, p. 114-122, 2018. Disponível em: https://doi.org/10.1016/j.techfore.2017.11.022. Acesso em: 16 out. 2025.

OFFICE FOR NATIONAL STATISTICS (London). **Experimental model-based single-month estimates for the labour force survey**: methods explained. London, 2019. Disponível em: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/experimentalmodelbasedsinglemonthestimatesforthelabourforcesurveymethodsexplained. Acesso em: 16 out. 2025.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **Quality framework and guidelines for OECD statistical activities**. [*S.l.*]: OECD, 2011. Disponível em: https://www.oecd.org/sdd/ qualityframeworkforoecdstatisticalactivities.htm. Acesso em: 16 out. 2025.

PALM, Franz C.; SMEEKES, Stephan; URBAIN, Jean-Pierre. Cross-sectional dependence robust block bootstrap panel unit root tests. **Journal of Econometrics**, Amsterdam, v. 163, n.1, p. 85-104, 2011. Disponível em: http://doi.org/10.1016/j.jeconom.2010.11.010. Acesso em: 16 out. 2025.

PETRIS, Giovanni. An R package for dynamic linear models. **Journal of Statistical Software**, Innsbruck, v. 36, n. 12, p. 1-16, 2010. Disponível em: http://www.jstatsoft. org/v36/i12/. Acesso em: 16 out. 2025.

PFAFF, Bernhard. **Analysis of Integrated and Cointegrated Time Series with R**. 2nd ed. New York: Springer, 2008.

PFEFFERMANN, Danny. Estimation and seasonal adjustment of population means using data from repeated surveys. **Journal of Business & Economic Statistics**, London, v. 9, n.2, p. 163-175, 1991. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/07350015.1991. 10509840. Acesso em: 16 out. 2025.

PFEFFERMANN, Danny. Methodological issues and challenges in the production of official statistics. **Journal of Survey Statistics and Methodology**, Oxford, v. 3, n. 4, p. 425-483, 2015. Disponível em: https://academic.oup.com/jssam/article/3/4/425/950591. Acesso em: 16 out. 2025.

PFEFFERMANN, Danny; FEDER, Moshe; SIGNORELLI, David. Estimation of autocorrelations of survey errors with application to trend estimation in small areas. **Journal of Business & Economic Statistics**, London, v. 16, n. 3, p. 339-348, 1998. Disponível em: https://doi.org/10.2307/1392510. Acesso em: 16 out. 2025.

PFEFFERMANN, Danny; TILLER, Richard. Small area estimation with state space models subject to benchmark constraints. **Journal of the American Statistical Association**, London, v. 101, p. 1387-1397, 2006. Disponível em: https://www.tandfonline.com/doi/abs/10.1198/016214506000000591. Acesso em: 16 out. 2025.

PREIS, Tobias; MOAT, Helen Susannah; STANLEY, H. Eugene. Quantifying trading behavior in financial markets using Google Trends. **Scientific Reports**, New York, v. 3, n. 1684, 2013. Disponível em: https://www.nature.com/search?q=Quantifying+trading+behavior+in+financial+markets+using+Google+Trends&journal=. Acesso em: 16 out. 2025.

RABINER, Lawrence; BIING, Hwang Juang. **Fundamentals of speech recognition**. Hoboken: Prentice Hall, 1993.

RAO, J. N. K.; MINGYU, Yu. Small-area estimation by combining time-series and cross-sectional data. **Canadian Journal of Statistics**, [*s.l.*], v. 22, p. 511-528, 1994. Disponível em: http://doi.org/10.2307/3315407. Acesso em: 16 out. 2025.

REISEN, V. A. *et al*. Robust dickey–fuller tests based on ranks for time series with additive outliers. **Metrika**, New York, v. 80, p. 115-131, 2017. Disponível em: http://doi.org/10.1007/s00184-016-0594-8. Acesso em: 16 out. 2025.

RIVERA, Roberto. A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. **Tourism Management**, [*s.l.*], v. 57, p. 12-20, 2016. Disponível em: http://doi.org/10.1016/j.tourman.2016.04.008. Acesso em: 16 out. 2025.

SALISU, Afees A.; OGBONNA, Ahamuefula E.; ADEDIRAN, Idris. Stock-induced Google Trends and the predictability of sectoral stock returns. **Journal of Forecasting**, Hoboken, v. 40, n. 2, p. 327-345, 2020. Disponível em: http://doi.org/10.1002/for.2722. Acesso em: 16 out. 2025.

SARDÁ-ESPINOSA, Alexis. **Time series clustering along with optimizations for the dynamic time warping distance**. Version 5.5.10. [*S.l.: s.n.*], 2022. Disponível em: https://cran.r-project.org/web/packages/dtwclust/index.html. Acesso em: 16 out. 2025.

SARDÁ-ESPINOSA, Alexis. Time-series clustering in r using the dtwclust package. **The R Journal**, [*s.l.*], v. 11, n. 1, 2019. Disponível em: http://doi.org/10.32614/RJ-2019-023. Acesso em: 16 out. 2025.

SÄRNDAL, Carl-Erik; SWENSSON, Bengt; WRETMAN, Jan. **Model assisted survey sampling**. New York: Springer, 1992.

SCHIAVONI, Caterina *et al*. A dynamic factor model approach to incorporate big data in state space models for official statistics. **Journal of the Royal Statistical Society**: series A statistic in society, Oxford, v. 184, n. 1, p. 324-353, 2020. Disponível em: http://doi.org/10.1111/rssa.12626. Acesso em: 16 out. 2025.

SCOTT, A. J.; SMITH, T. M. F. Analysis of repeated surveys using time series methods. **Journal of the American Statistical Association**, London, v. 69, n. 347, p. 674-678, 1974. Disponível em: https://www.jstor.org/stable/2286000. Acesso em: 16 out. 2025.

SCOTT, A. J.; SMITH, T. M. F.; JONES, R. G. The application of time series methods to the analysis of repeated surveys. **International Statistical Review**, Haia, v. 45, n. 1, p. 13-28, 1977. Disponível em: https://www.jstor.org/stable/1403000. Acesso em: 16 out. 2025.

SEUNG-PYO, Jun; HYOUNG, Sun Yoo; SAN, Choi. Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. **Technological Forecasting and Social Change**, Hoboken, v. 130, p. 69-87, 2017. Disponível em: https://doi.org/10.1016/j.techfore.2017.11.009. Acesso em: 16 out. 2025.

SHIKIDA, Cláudio D. *et al*. Informações e política econômica: um teste para aperfeiçoamento de erros de previsão a partir da utilização do Google Trends. **Revista Gestão & Políticas Públicas**, São Paulo, v. 2, n. p. 197-218, 2012. Disponível em: https://www.revistas.usp.br/rgpp/article/view/97858. Acesso em: 16 out. 2025.

SILVA, Denise Britz do Nascimento. **Modelling compositional time series from repeated surveys**. 1996. 235 f. Tese (Doutorado em Filosofia) – University of Southampton, Southampton, 1996. Disponível em: https://eprints.soton.ac.uk/462965/. Acesso em: 16 out. 2025.

SILVA, Denise Britz do Nascimento; SMITH, T. M. F. Modelling compositional time series from repeated surveys. **Survey Methodology**, [*s.l.*], v. 27, n. 2, p. 205-2015, 2001. Disponível em:

https://www150.statcan.gc.ca/n1/pub/12-001-x/2001002/article/6097-eng.pdf. Acesso em: 16 out. 2025.

SIMIONESCU, Mihaela. Improving unemployment rate forecasts at regional level in Romania using Google Trends. **Technological Forecasting and Social Change**, Amsterdam, v. 155, e120026, 2020. Disponível em: http://doi.org/10.1016/j.techfore.2020.120026. Acesso em: 16 out. 2025.

SIMIONESCU, Mihaela; CIFUENTES-FAURA, Javier. Can unemployment forecasts based on Google trends help government design better policies?: an investigation based on Spain and Portugal. **Journal of Policy Modeling**, Amsterdam, v. 44, n. 1, p. 1-21, 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S0161893821001101. Acesso em: 16 out. 2025.

SMEEKES, Stephan; WIJLER, Etienne. An automated approach towards sparse single-equation cointegration modelling. **Journal of Econometrics**, Amsterdam, v. 221, n. 1, p. 247-276, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0304407620302190. Acesso em: 16 out. 2025.

SMEEKES, Stephan; WILMS, Ines. BootUR: an R package for bootstrap unit root tests. **Journal of Statistical Software**, Innsbruck, v. 106, n. 12, 2023. Disponível em: https://www.jstatsoft.org/article/view/v106i12. Acesso em: 16 out. 2025.

SMITH, Paul. Google's MIDAS touch: predicting UK unemployment with internet search data. **Journal of Forecasting**, Hoboken, 2016. Disponível em: https://doi.org/10.1002/for.2391. Acesso em: 16 out. 2025.

SMITH, T. M. F. Principles and problems in the analysis of repeated surveys. *In*: NAMBOODIRI, N. Krishnan (ed.). **Survey sampling and measurement**. New York: Academic Press, 1978. p. 201-216.

STOCK, J. H.; WATSON, M. W. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *In*: TAYLOR, John; UHLIG, Harald. **Handbook of macroeconomics**. Amsterdam: Elsevier, 2017. v. 2A.

SUHOY, Tanya. **Query indices and a 2008 downturn**: Israeli data. Jerusalem: Bank of Israel, 2009. (Working paper, n. 6). Disponível em: https://www.boi.org.il/en/Research/Pages/papers_dp0906e.aspx. Acesso em: 16 out. 2025.

TAKUMI, Ito *et al*. Application of Google Trends-based sentiment index in exchange rate prediction. **Journal of Forecasting**, Hoboken, 2021. Disponível em: https://doi.org/10.1002/for.2762. Acesso em: 16 out. 2025.

TAO, Chen *et al*. The 2007-2008 U.S. recession: what did the real-time Google Trends data tell the United States? **Contemporary Economic Policy**, Hoboken, v. 33, n. 2, p. 395-403, 2015. Disponível em: http://doi.org/10.1111/coep.12074. Acesso em: 15 out. 2025.

TILLER, Richard B. Time series modelling of sample survey data from the U.S. current population survey. **Journal of Official Statistics**, Thousand Oaks, v. 8, n. 2, p. 149-166, 1992. Disponível em: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/time-series- modeling-of-sample-survey-data-from-the-u.s.-current-population-survey.pdf. Acesso em: 16 out. 2025.

TOEWS, Michael W.; WHITFIELD, Paul H.; ALLEN, Diana M. Seasonal statistics: the 'seas' package for R. **Computers & Geosciences**, Amsterdam, v. 33, n. 7, p. 944-951, 2007. Disponível em: https://doi.org/10.1016/j.cageo.2006.11.011. Acesso em: 16 out. 2025.

TSAY, Ruey S. **Multivariate time series analysis**: with R and financial applications. Hoboken: Wiley, 2013.

UNITED NATIONS. Statistics Division. **Fundamental principles of official statistic**. New York, 2014. Disponível em: http:// unstats.un.org/unsd/dnss/gp/FP-New-E.pdf. Acesso em: 16 out. 2025.

UNITED NATIONS. **United Nations fundamental principles of official statistics**: implementation guidelines. New York, 2015. Disponível em: https://unstats.un.org/unsd/dnss/gp/impguide.aspx. Acesso em: 16 out. 2025.

UNITED STATES. Bureau of Labor Statistics. **Handbook of methods**: local area unemployment statistics. Washington, 2018. Disponível em: https://www.bls. gov/opub/hom/lau/pdf/lau.pdf. Acesso em: 16 out. 2025.

VALK, Serge de; MATTOS, Daiane Marcolino; FERREIRA, Pedro Guilherme. Nowcasting: an R package for predicting economic variables using dynamic factor models. **The R Journal**, [*s.l.*], v. 11, n. 1, p. 230-244, 2019. Disponível em: https://doi.org/10.32614/RJ-2019-020. Acesso em: 16 out. 2025.

VAN DEN BRAKEL, Jan A.; BUELENS, Bart; BOONSTRA, Harm-Jan. Small area estimation to quantify discontinuities in repeated sample surveys. **Journal of the Royal Statistical Society**: series A statistic in society, Oxford, v. 179, n. 1, p. 229-250, 2016. Disponível em: http://doi.org/10.1111/rssa.12110. Acesso em: 15 out. 2025.

VAN DEN BRAKEL, Jan; KRIEG, Sabine. Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. **Survey Methodology**, Ottawa, v. 41, n. 2, p. 267-296, 2015. Disponível em: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14231-eng.pdf. Acesso em: 15 out. 2025.

VAN DEN BRAKEL, Jan; KRIEG, Sabine. Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. **Survey Methodology**, Ottawa, v. 35, n. 2, p. 177-190, 2009. Disponível em: https://www150.statcan.gc.ca/n1/pub/12-001-x/2009002/article/11040-eng.pdf. Acesso em: 15 out. 2025.

VAN DEN BRAKEL, Jan; KRIEG, Sabine. Small area estimation with state space common factor models for rotating panels. **Journal of the Royal Statistical Society**: series A statistic in society, Oxford, v. 179, n. 3, p. 763-791, 2016. Disponível em: http://doi.org/10.1111/rssa.12158. Acesso em: 15 out. 2025.

VAN DEN BRAKEL, Jan; SOUREN, Martijn; KRIEG, Sabine. Estimating monthly labour force figures during the Covid-19 pandemic in the Netherlands. **Journal of the Royal Statistical Society**: series A statistic in society, Oxford, v.185, n. 4, p. 1560-1583, 2021. Disponível em: https://academic.oup.com/jrsssa/article/185/4/1560/7069419?login=false. Acesso em: 15 out. 2025.

VICHI, Maurizio; HAND, David J. Trusted smart statistics: the challenge of extracting usable aggregate information from new data sources. **Statistical Journal of the IAOS**, Thousand Oaks, v. 35, n. 4, p. 605-613, 2019. Disponível em: http://doi.org/10.3233/SJI-190526. Acesso em: 16 out. 2025.

VOSEN, Simeon; SCHMIDT, Torsten. Forecasting private consumption: survey-based indicators vs. Google Trends. **Journal of Forecasting**, Hoboken, v. 30, p. 565-578, 2011. Disponível em: http://doi.org/10.1002/for.1213. Acesso em: 16 out. 2025.

WEI, William W. S. **Multivariate time series analysis and applications**. Hoboken: John Wiley & Sons, 2019. (Wiley series in probability and statistics).

WIJLER, Etiënne J. J. *High*-**dimensional time series analysis**: unit roots, cointegration and forecasting*.* 2021. Tese (Doutorado) – Maastricht University, Holanda, 2021. Disponível em: https://doi.org/10.26481/dis.20210114ew. Acesso em: 16 out. 2025.

WOO, Jaemin; OWEN, Ann L. Forecasting private consumption with Google Trends data. **Journal of Forecasting**, Hoboken, v. 38, n. 2, p. 81-91, 2018. Disponível em: http://doi.org/10.1002/for.2559. Acesso em: 16 out. 2025.

YANG, Yang; BING, Pan; HAIYAN, Song. Predicting hotel demand using destination marketing organization's web traffic data. **Journal of Travel Research**, Thousand Oaks, v. 53, n. 4, p. 433-447, 2014. Disponível em: http://doi.org/10.1177/0047287513500391. Acesso em: 16 out. 2025.

YUE, Teng *et al*. Dynamic forecasting of Zika epidemics using Google trends. **PLOS ONE**, v. 12, n. 1, e0165085, 2017. Disponível em: https://doi.org/10.1371/journal.pone.0165085. Acesso em: 16 out. 2025.

ZAMPROGNO, Bartolomeu *et al*. Principal component analysis with autocorrelated data. **Journal of Statistical Computation and Simulation**, London, v. 90, n. 12, p. 2117-2135, 2020. Disponível em: https://doi.org/10.1080/00949655.2020.1764556. Acesso em: 16 out. 2025.

ZHI, Su. Chinese online unemployment-related searches and macroeconomic indicators. **Frontiers of Economics in China**, Beijing, v. 9, n. 4, p. 573-605, 2014. Disponível em: https://journal.hep.com.cn/fec/EN/10.3868/s060-003-014-0027-3. Acesso em: 16 out. 2025.

APPENDIX A

A1: Initial words

Table 7: List of initial words (continue)

ID	Term in Portuguese	Term in English	Comments
1	agencia de emprego	employment agency	
2	agencias de emprego	employment agencies	
3	agendamento de carteira de trabalho	employment record book scheduling	
4	agendar seguro desemprego	schedule unemployment insurance	
5	balcão de empregos		Job advertiser's name
6	blog do emprego		Job advertiser's name
7	busca de trabalho	job search	
8	busca emprego	job search	
9	cadastro curriculo	resume registration	
10	carteira de trabalho	employment record book	Personal document with work records
11	catho		Job advertiser's name
12	catho vagas		Job advertiser's name
13	CLT		Acronym for Consolidação das Leis do Trabalho, set of formal labour regulations
14	CLT contrato de trabalho	CLT employment contract	
15	como agendar carteira de trabalho	how to schedule employment record book	
16	como fazer curriculo	how to make resume	
17	comunidade de emprego		Job advertiser's name
18	contrato de experiencia	experience contract	
19	contrato de trabalho	work contract	
20	contrato de trabalho determinado	fixed work contract	
21	contrato prazo determinado	fixed term contract	
22	curriculo	curriculum vitae	Written without the accent
23	currículo	curriculum vitae	
24	emprego	job	
25	empregos	jobs	
26	entrevista de emprego	job interview	

Table 7: List of initial words

(continued)

ID	Term in Portuguese	Term in English	Comments
27	fazer curriculo	make resume	
28	hora do emprego		Job advertiser's name
29	indeed		Job advertiser's name
30	indeed vagas	indeed vacancies	
31	infojobs		Job advertiser's name
32	infojobs vagas	infojobs vacancies	
33	informe vagas	report vacancies	
34	mercado de trabalho	labour market	
35	modelo contrato de trabalho	employment contract template	
36	modelo de contrato	contract template	
37	modelo de contrato de trabalho	employment contract template	
38	modelo de currículo	resume template	
39	OLX emprego	OLX job	Job advertiser's name
40	OLX empregos	OLX jobs	
41	oportunidade de emprego	employment opportunity	
42	oportunidade de trabalho	job opportunity	
43	oportunidades de emprego	employment opportunities	
44	oportunidades de trabalho	job opportunities	
45	primeiro emprego	first job	
46	processo do trabalho	work process	
47	procuro emprego	I am looking for a job	
48	remuneração	remuneration	
49	rescisão	termination	
50	rescisão contrato de trabalho	termination of employment contract	
51	rescisão de contrato	contract termination	
52	rescisão de contrato de trabalho	termination of work contract	
53	rescisão de trabalho	termination of employment	
54	salario	wage	Written without the accent

Table 7: List of initial words

(continued)

ID	Term in Portuguese	Term in English	Comments
55	salário	wage	
56	segunda via carteira de trabalho	second version employment record book	
57	seguro desemprego	unemployment insurance/ claimant count	
58	SINE		Acronym for Sistema Nacional de Emprego (SINE), public job search service.
59	SINE vagas	SINE vacancies	Acronym for Sistema Nacional de Emprego (SINE)
60	SINE vagas de emprego	SINE job openings	Acronym for Sistema Nacional de Emprego (SINE)
61	site de emprego	job site	
62	site de empregos	jobs site	
63	suspensão contrato de trabalho	work contract suspension	
64	suspensão de contrato de trabalho	suspension of employment contract	
65	termo de rescisão	term of termination	
66	termo de rescisão de contrato	contract termination term	
67	tirar carteira de trabalho	get a employment record book	
68	trabalho	work	
69	vaga	vacancy	
70	vaga de	vacancy of	
71	vaga de emprego	job opportunity	
72	vaga de empregos	job vacancy	
73	vaga de estagio	internship vacancy	
74	vaga de trabalho	job vacancy	
75	vaga emprego	employment vacancy	
76	vagas	vacancies	

Table 7: List of initial words

(conclusion)

ID	Term in Portuguese	Term in English	Comments
77	vagas de emprego	employment vacancies	
78	vagas de emprego SINE	employment vacancies SINE	
79	vagas de trabalho	job vacancies	
80	vagas emprego	vacancies employment	
81	vagas SINE	SINE vacancies	Acronym for Sistema Nacional de Emprego (SINE)
82	vagas.com		Job advertiser's name, web site

A 2: Final Selected Google Trends Words

Table 8: Selected final words – Brazil

(continue)

ID	Term in Portuguese	Term in English	Comments
1	Acorda cidade emprego	Acorda cidade employment	Job advertiser's name
2	Acorda cidade empregos	Acorda cidade employment	Job advertiser's name
3	Agencia de emprego	Employment agency	Written without the accent
4	Agência de emprego	Employment agency	
5	Agencia de emprego RJ	Employment agency RJ	DF acronym for Rio de Janeiro, Brazilian state/city
6	Agencias de emprego	Employment agencies	
7	Balcao		Written without the accent, job advertiser's name
8	Balcao de emprego		Written without the accent, job advertiser's name
9	Balcão de emprego		Job advertiser's name
10	Banda B empregos	Banda B employment	Job advertiser's name
11	Bauru empregos	Bauru jobs	Name of the city (Bauru)
12	Blog do emprego df		Job advertiser's name, DF acronym for Federal District
13	Blog emprego df		Job advertiser's name, DF acronym for Federal District
14	Bne		Job advertiser's name, BNE acronym for Banco Nacional de Empregros
15	Cadastrar curriculo	Register resume	
16	Carioca empregos		Job advertiser's name
17	Catho		Job advertiser's name
18	Catho entrar		Job advertiser's name
19	Catho online	Catho online	Job advertiser's name
20	Ciee		Acronym for Centro de Integração Empresa-Escola, association with internship programs
21	Ciee pe	Ciee pe	Acronym for Pernambuco (PE), Brazilian state
22	Classificados	Newspaper ads	

Table 8: Selected final words – Brazil

(continued)

ID	Term in Portuguese	Term in English	Comments
23	Classificados empregos	Newspaper ads jobs	
24	Como fazer currículo	How to make resume	
25	Como fazer um currículo	How to make a resume	
26	Comunidade de emprego		Job advertiser's name
27	Contratando	Hiring	
28	Correio braziliense emprego		Job advertiser's name
29	Correio emprego	Job courier	Job advertiser's name
30	Curriculo	Curriculum	Written without the accent
31	Currículo	Curriculum	
32	Curriculo jovem aprendiz	Young apprentice curriculum	
33	Currículo lattes	Curriculum lattes	Curriculum vitae specific to the academic area
34	Curriculo online	Curriculum online	
35	Currículo online pdf	Curriculum online pdf	
36	Currículo para primeiro emprego	Resume for first job	
37	Currículo pdf	Resume pdf	
38	Curriculo primeiro emprego	Resume first job	Written without the accent
39	Currículo primeiro emprego	Resume first job	
40	Currículo primeiro emprego objetivo	Resume first job objective	
41	Curriculo profissional	Professional resume	
42	Currículo profissional	Professional resume	
43	Curriculo pronto	Resume ready	Written without the accent
44	Currículo pronto	Resume ready	
45	Curriculo pronto word	Resume ready word	
46	Currículo simples	Simple resume	
47	Curriculo vitae	Curriculum vitae	Written without the accent
48	Currículo vitae	Curriculum vitae	
49	Curriculo word	Resume word	
50	Curriculum	Curriculum vitae	
51	Edital de emprego		Job advertiser's name
52	Emprega Brasil		Government service for publicizing vacancies and employment policies

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
53	Emprega Campinas		Job advertiser's name
54	Emprega Sao Jose		Job advertiser's name
55	Emprego	Job	
56	Emprego BH	Job BH	Acronym for Belo Horizonte (BH), Brazilian city
57	Emprego Brasilia	Job Brasilia	Brazilian capital (Brasília)
58	Emprego Curitiba	Job Curitiba	Brazilian city (Curitiba)
59	Emprego DF	DF Job	DF acronym for Federal District
60	Emprego ligado		Job advertiser's name
61	Emprego Manaus	Job Manaus	Brazilian city
62	Emprego na OLX	Job OLX	
63	Emprego OLX Salvador	Job OLX Salvador	Brazilian city
64	Emprego OLX SP	Job OLX São Paulo	Acronym for São Paulo (SP), Brazilian state/city
65	Emprego PE	Job PE	Acronym for Pernambuco (PE), Brazilian state
66	Emprego pra ontem		Job advertiser's name
67	Emprego RJ	Job RJ	Acronym for Rio de Janeiro (RJ), Brazilian state
68	Emprego Santista		Job advertiser's name
69	Empregos	Jobs	
70	Empregos BH	BH Jobs	Acronym for Belo Horizonte (BH), Brazilian city
71	Empregos Campo Grande MS	Jobs Campo Grande MS	Acronym for Mato Grosso do Sul (MS), Brazilian state and Campo Grande is a Brazilian city
72	Empregos Curitiba	Curitiba Jobs	Brazilian city (Curitiba)
73	Empregos DF	DF Jobs	DF acronym for Federal District
74	Empregos em Curitiba	Jobs in Curitiba	Brazilian city (Curitiba)
75	Empregos OLX	OLX Jobs	Job advertiser's name
76	Empregos pe	Jobs pe	Acronym for Pernambuco (PE), Brazilian state
77	Empregos SP	SP Jobs	Acronym for São Paulo (SP), Brazilian state

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
78	Enviar curriculo	Send resume	Written without the accent
79	Enviar currículo	Send resume	
80	Estagio	Internship	
81	Estagio CIEE	Internship CIEE	Acronym for Centro de Integração Empresa-Escola, association with internship programs
82	Estagios	Internships	
83	Fazer curriculo lattes	Make curriculum lattes	Curriculum vitae specific to the academic area
84	Fazer currículo online	Make resume online	
85	Fazer currículo pdf	Make resume pdf	
86	Fazer currículo pelo celular	Make resume by mobile	
87	Fazer um currículo	Make resume	
88	Gi group		Job advertiser's name
89	Global empregos		Job advertiser's name
90	Google empregos	Google jobs	
91	Habilidades para curriculo	Resume skills	
92	Hard franca empregos		Job advertiser's name
93	Havan currículo	Havan resume	Trading company's name
94	Hora do emprego		Job advertiser's name
95	Indeed		Job advertiser's name
96	Indeed emprego	Indeed employment	
97	Indeed empregos	Indeed jobs	
98	Indeed RJ	Indeed RJ	Acronym for Rio de Janeiro (RJ), Brazilian state
99	Infojobs		Job advertiser's name
100	Infojobs entrar	Infojobs login	
101	Infojobs rj		Acronym for Rio de Janeiro (RJ), Brazilian state
102	Infojobs sp		Acronym for São Paulo (SP), Brazilian state
103	Informe vagas PE		Job advertiser's name, scronym for Pernambuco (PE), Brazilian state

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
104	Informe vagas		Job advertiser's name
105	JF empregos	JF jobs	Acronym forJuiz de Fora (JF), Brazilian city
106	Jornal meu emprego	Jornal meu emprego	
107	Linkedin		Job advertiser's name
108	Luandre		Job advertiser's name
109	Mais emprego		Job advertiser's name
110	Mais empregos		Job advertiser's name
111	Maringa.com		Job advertiser's name. Brazilian city (Maringa)
112	Maringa.com empregos	Maringa.com jobs	Job advertiser's name. Brazilian city (Maringa)
113	Meu primeiro emprego	My first job	
114	Modelo currículo	Resume template	
115	Modelos currículo	Resume templates	
116	modelos de currículo	Resume Templates	
117	O que colocar no curriculo	What to put on resume	
118	O que é currículo	What is resume	
119	Objetivo	Objetive	
120	Objetivo curriculo	Resume objective	
121	Objetivo de curriculo	Resume objective	Written without the accent
122	Objetivo de currículo	Resume objective	
123	Objetivo no curriculo	Objective in resume	
124	Objetivo para currículo primeiro emprego	Objective in resume first job	
125	OLX Curitiba		Job advertiser's name, Brazilian city (Curitiba)
126	OLX DF		Job advertiser's name, DF acronym for Federal District
127	OLX emprego Curitiba	OLX job Curitiba	Job advertiser's name, Brazilian city (Curitiba)
128	OLX emprego DF	OLX jobs DF	Job advertiser's name, DF acronym for Federal District

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
129	OLX emprego RJ	OLX job RJ	Job advertiser's name. Acronym for Rio de Janeiro (RJ), Brazilian state
130	OLX empregos Curitiba	OLX jobs Curitiba	Job advertiser's name, Brazilian city (Curitiba)
131	OLX empregos RJ	OLX jobs RJ	Job advertiser's name. Acronym for Rio de Janeiro (RJ), Brazilian state
132	OLX empregos SP	OLX jobs SP	Job advertiser's name. Acronym for São Paulo (SP), Brazilian state
133	OLX MS	OLX MS	Job advertiser's name. Acronym for Mato Grosso do Sul (MS), Brazilian state
134	OLX RJ	OLX RJ	Job advertiser's name. Acronym for Rio de Janeiro (RJ), Brazilian state
135	OLX SP	OLX SP	Job advertiser's name. Acronym for São Paulo (SP), Brazilian state
136	OLX vaga de emprego	OLX job vacancy	Job advertiser's name.
137	OLX vagas de emprego Curitiba	OLX job vacancies Curitiba	Job advertiser's name, Brazilian city (Curitiba)
138	OLX vagas de empregos	OLX job vacancies	
139	Plataforma lattes	Platform lattes	Academic curriculum platform.
140	Portal mais emprego		Former government service for publicizing vacancies and employment policies
141	Primeiro curriculo	First resume	
142	Primeiro emprego RJ	First job RJ	Acronym for Rio de Janeiro (RJ), Brazilian state
143	Procura	Demand	
144	Procura emprego	Looking for a job	
145	Procuro emprego	Looking for job	
146	Quero bolsa		Scholarship advertiser's name.

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
147	Rio vagas comparecer		Brazilian state/city (Rio de Janeiro)
148	Riovagas	Riovagas	Job advertiser's name. Brazilian state/city (Rio de Janeiro)
149	Salário	Wage	
150	Seguro desemprego	Unemployment insurance/ claimant count	
151	SINE		Acronym for Sistema Nacional de Emprego (SINE), public job search service.
152	SINE BH		Acronym for Sistema Nacional de Emprego (SINE), public job search service. Acronym for Belo Horizonte (BH), Brazilian city
153	SINE Curitiba		Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Curitiba)
154	SINE Emprego	SINE job	Acronym for Sistema Nacional de Emprego (SINE)
155	SINE empregos	SINE jobs	Acronym for Sistema Nacional de Emprego (SINE)
156	SINE Fortaleza		Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Fortaleza)
157	SINE Goiania	SINE Goiania	Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Goiânia)
158	SINE idt		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Instituto de Desenvolvimento do Trabalho,
159	SINE RJ		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Rio de Janeiro (RJ), Brazilian state

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
160	SINE RS		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Rio Grande do Sul (RS), Brazilian state
161	SINE Sobral	SINE sobral	Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Sobral)
162	SINE vaga de emprego	SINE job vacancy	Acronym for Sistema Nacional de Emprego (SINE).
163	SINE vagas Curitiba	SINE vacancies Curitiba	Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Curitiba)
164	SINEBahia vagas de emprego	SINEBahia job vacancies	Acronym for Sistema Nacional de Emprego (SINE). Brazilian state (Bahia)
165	Sites de emprego	Job sites	
166	Sucessor rh		Job advertiser's name.
167	Trabalhe conosco	Work with us	
168	Trabalho	Work	
169	Vaga	Vacancy	
170	Vaga de	Vacancy of	
171	Vaga de emprego	Job opportunity	
172	Vaga de emprego DF	DF job vacancy	DF acronym for Federal District
173	Vaga de emprego RJ	RJ job vacancy	Acronym for Rio de Janeiro (RJ), Brazilian state
174	Vaga de emprego SP	SP job vacancy	Acronym for São Paulo (SP), Brazilian state
175	Vaga de empregos	Job vacancy	
176	Vaga de trabalho	Job vacancy	
177	Vaga emprego RJ	job vacancy RJ	Acronym for Rio de Janeiro (RJ), Brazilian state
178	Vaga emprego SP	SP job vacancy	Acronym for São Paulo (SP), Brazilian state
179	Vaga para emprego	job vacancy	

Table 8: Selected final words – Brazil (continued)

ID	Term in Portuguese	Term in English	Comments
180	Vagas	Vacancies	
181	Vagas BH	Vacancies BH	Acronym for Belo Horizonte (BH), Brazilian city
182	Vagas Curitiba	Curitiba vacancies	Brazilian city (Curitiba)
183	Vagas de emprego	Jobs	
184	Vagas de emprego BH	Job vacancies BH	Acronym for Belo Horizonte (BH), Brazilian city
185	Vagas de emprego Curitiba	Curitiba job vacancies	Brazilian city (Curitiba)
186	vagas de emprego df OLX	DF job vacancies	DF acronym for Federal District
187	Vagas de emprego DF OLX	DF OLX job vacancies	DF acronym for Federal District. Job advertiser's name.
188	Vagas de emprego do SINE	SINE job vacancies	Acronym for Sistema Nacional de Emprego (SINE)
189	Vagas de emprego em BH	Job vacancies in BH	Acronym for Belo Horizonte (BH), Brazilian city
190	Vagas de emprego em Fortaleza	Job vacancies in Fortaleza	Brazilian city (Fortaleza)
191	Vagas de emprego ES	ES job vacancies	Acronym for Espírito Santo (ES), Brazilian state
192	Vagas de emprego Fortaleza	Fortaleza job vacancies	Brazilian city (Fortaleza)
193	Vagas de emprego indeed	Indeed job vacancies	
194	Vagas de emprego na OLX	job vacancies at OLX	Job advertiser's name.
195	Vagas de emprego no SINE	Job vacancies on SINE	Acronym for Sistema Nacional de Emprego (SINE)
196	Vagas de emprego na OLX	OLX job vacancies	Job advertiser's name.
197	Vagas de emprego PE	Job vacancies PE	Acronym for Pernambuco (PE), Brazilian state
198	Vagas de emprego Recife	Recife job vacancies	Brazilian city (Recife)
199	Vagas de emprego RJ	RJ job vacancies	Acronym for Rio de Janeiro (RJ), Brazilian state
200	Vagas de emprego RS	RS job vacancies	Acronym for Rio Grande do Sul (RS), Brazilian state
201	Vagas de emprego Salvador	Salvador job vacancies	Brazilian city (Salvador)
202	Vagas de emprego SINE	Employment vacancies SINE	Acronym for Sistema Nacional de Emprego (SINE)

Table 8: Selected final words – Brazil (conclusion)

ID	Term in Portuguese	Term in English	Comments
203	Vagas de emprego SINE BH	SINE bh job vacancies	Acronym for Belo Horizonte (BH), Brazilian city
204	Vagas de emprego SP	Job vacancies SP	Acronym for São Paulo (SP), Brazilian state
205	Vagas de empregos	Job vacancies	
206	Vagas de trabalho	Job vacancies	
207	Vagas DF	Vacancies DF	DF acronym for Federal District
208	Vagas do SINE	SINE vacancies	Acronym for Sistema Nacional de Emprego (SINE)
209	Vagas emprego RJ	job vacancy RJ	Acronym for Rio de Janeiro (RJ), Brazilian state
210	Vagas home office	Home office vacancies	
211	Vagas indeed	Vacancies indeed	Job advertiser's name.
212	Vagas no SINE	Vacancies in the SINE	Acronym for Sistema Nacional de Emprego (SINE)
213	Vagas para emprego	Job vacancies	
214	Vagas para trabalho	Job vacancies	
215	Vagas rio	Vacancy rio	Brazilian city/state (Rio de Janeiro)
216	Vagas RJ	Vacancies RJ	Acronym for Rio de Janeiro (RJ), Brazilian state
217	Vagas SINE	SINE vacancies	Acronym for Sistema Nacional de Emprego (SINE)
218	Vagas SINE bh	Vacancies SINE bh	Acronym for Sistema Nacional de Emprego (SINE). Acronym for Belo Horizonte (BH), Brazilian city
219	Vagas SINE hoje	Vacancies SINE today	
220	Vagas SINE rj	Vacancies SINE rj	Acronym for Sistema Nacional de Emprego (SINE). Acronym for Rio de Janeiro (RJ), Brazilian state
221	vagas.com		job advertiser's name, web site
222	vagas.com.br		job advertiser's name, web site

Table 9: Selected final words – Minas Gerais (continue)

ID	words_selected	Term in English	Comments
1	Bhjobs	Bhjobs	Job advertiser's name
2	Catho	Catho	Job advertiser's name
3	Catho empregos	Catho jobs	Job advertiser's name
4	Como fazer currículo	How to make resume	
5	Comunidade de emprego		Job advertiser's name
6	Concurso publico	Public service contest	
7	Currículo	Curriculum vitae	
8	Currículo para primeiro emprego	Resume for first job	
9	Currículo pdf	Resume pdf	
10	Currículo primeiro emprego	Resume first job	
11	Curriculo pronto word	Word ready resume	
12	currículo vitae	Curriculum vitae	
13	Curriculum	Curriculum vitae	
14	Curriculum vitae	Curriculum vitae	
15	Educa mais brasil		Scholarship advertiser's name.
16	Emprego	Job	
17	Emprego BH	Job BH	Acronym for Belo Horizonte (BH), Brazilian city
18	Emprego Contagem	Job Contagem	Brazilian city (Contagem)
19	Emprego em BH	Job in BH	
20	Emprego Montes Claros	Montes Claros jobs	Brazilian city (Montes Claros)
21	Emprego Uberlandia	Job Uberlândia	Brazilian city (Uberlândia)
22	Empregos	Jobs	
23	Empregos BH	Jobs BH	Acronym for Belo Horizonte (BH), Brazilian city
24	Empregos MG	Jobs MG	Acronym for Minas Gerais (MG), Brazilian state
25	Empregos Uberlandia	Jobs Uberlândia	Brazilian city (Uberlândia)
26	Fato real empregos	Fato real jobs	Job advertiser's name
27	Fazer currículo	Make resume	
28	Fazer currículo online	Make resume online	
29	Indeed	Indeed	Job advertiser's name
30	indeed bh		Job advertiser's name. Acronym for Belo Horizonte (BH), Brazilian city

Table 9: Selected final words – Minas Gerais (continued)

ID	words_selected	Term in English	Comments
31	Indeed vagas	Indeed jobs	Job advertiser's name
32	Infojobs	Infojobs	Job advertiser's name
33	Ipecont vagas	Ipecont vacancies	Job advertiser's name
34	Jf emprego		Job advertiser's name. Acronym for Juiz de Fora (JF), Brazilian city
35	Jf empregos	Jf jobs	Job advertiser's name. Acronym for Juiz de Fora (JF), Brazilian city
36	Jfempregos	Jf Jobs	Job advertiser's name. Acronym for Juiz de Fora (JF), Brazilian city
37	Jornal balcao		Job advertiser's name.
38	Mais emprego		Government employment web-site
39	Minas curriculo	Minas curriculum	Reference to Minas Gerais, Brazilian state.
40	Objetivo para currículo	Objective for resume	
41	Objetivos para curriculo	Curriculum objectives	
42	OLX empregos	OLX jobs	Job advertiser's name.
43	OLX MG		Acronym for Minas Gerais (MG), Brazilian state
44	OLX Uberlandia	OLX uberlandia	Job advertiser's name. Brazilian city (Uberlândia)
45	Oportunidades BH	BH Opportunities	Acronym for Belo Horizonte (BH), Brazilian city
46	Procuro emprego	Looking for a job	
47	Rio vagas	River vacancies	Reference to Rio de Janeiro, Brazilian state/city
48	Salário	Wage	
49	Sest senat vagas	Sest senat vacancies	Organization for transport work-ers
50	SINE	SINE	Acronym for Sistema Nacional de Emprego (SINE), public job search service.
51	SINE bh		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Belo Horizonte (BH), Brazilian city

Table 9: Selected final words – Minas Gerais (continued)

ID	words_selected	Term in English	Comments
52	SINE bh empregos	SINE bh jobs	Acronym for Sistema Nacional de Emprego (SINE). Acronym for Belo Horizonte (BH), Brazilian city
53	SINE Contagem		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Contagem)
54	SINE de bh		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Belo Horizonte (BH), Brazilian city
55	SINE de Contagem		Acronym for Sistema Nacional de Emprego (SINE). Acronym for Sistema Nacional de Emprego (SINE). Brazilian city (Contagem)
56	SINE emprego	SINE job	Acronym for Sistema Nacional de Emprego (SINE), public job search service.
57	SINE empregos	SINE jobs	Acronym for Sistema Nacional de Emprego (SINE), public job search service.
58	Trabalho home office	Work from home office	
59	Vaga de	Vacancy of	
60	Vaga de emprego	Job vacancy	
61	Vaga emprego	Job vacancy	
62	Vagas	Vacancies	
63	Vagas bh	Vacancies BH	Acronym for Belo Horizonte (BH), Brazilian city
64	Vagas contagem	Vacancies Contagem	Brazilian city (Contagem)
65	Vagas de emprego	Jobs	
66	Vagas de emprego Contagem	Job vacancies Contagem	Brazilian city (Contagem)
67	Vagas de emprego em Contagem	Job vacancies Contagem	Brazilian city (Contagem)
68	Vagas de emprego em Uberlandia	Job vacancies in Uberlandia	Brazilian city (Uberlândia)
69	Vagas de emprego MG	MG job vacancies	Acronym for Minas Gerais (MG), Brazilian state

Table 9: Selected final words – Minas Gerais (continued)

ID	words_selected	Term in English	Comments
70	Vagas de emprego no SINE	Job vacancies in SINE	Acronym for Sistema Nacional de Emprego (SINE), public job search service.
71	Vagas de emprego Sete Lagoas	Job vacancies sete lagoas	Brazilian city (Sete Lagoas)
72	Vagas de emprego SINE	Employment vacancies SINE	Acronym for Sistema Nacional de Emprego (SINE), public job search service.
73	Vagas de emprego SINE BH	Job vacancies in BH	Acronym for Belo Horizonte (BH), Brazilian city
74	Vagas de emprego Uberlandia	Uberlandia job vacancies	Brazilian city (Uberlândia)
75	Vagas de estagio	Internship vacancies	Brazilian city (Uberlândia)
76	Vagas de emprego bh	BH job vacancies	Acronym for Belo Horizonte (BH), Brazilian city
77	Vagas em BH	Jobs bh	Acronym for Belo Horizonte (BH), Brazilian city
78	Vagas emprego	Job vacancies	
79	Vagas emprego BH	Job Vacancies BH	Acronym for Belo Horizonte (BH), Brazilian city
80	Vagas de emprego Uberlandia	Uberlandia job vacancies	Brazilian city (Uberlândia)
81	Vagas indeed BH	Vacancies indeed BH	Job advertiser's name. Acronym for Belo Horizonte (BH), Brazilian city
82	Vagas MG	Jobs mg	Acronym for Minas Gerais (MG), Brazilian state
83	Vagas no SINE	Vacancies in the SINE	Acronym for Sistema Nacional de Emprego (SINE), public job search service.
84	Vagas no SINE BH	Vacancies in SINE BH	Acronym for Sistema Nacional de Emprego (SINE), public job search service. Acronym for Belo Horizonte (BH), Brazilian city
85	Vagas PE	PE vacancies	Acronym for Pernambuco (PE), Brazilian state
86	Vagas pitagoras		Advertiser'name
87	Vagas SINE	SINE vacancies	Acronym for Sistema Nacional de Emprego (SINE), public job search service.

Table 9: Selected final words – Minas Gerais (conclusion)

ID	words_selected	Term in English	Comments
88	Vagas SINE BH	Vacancies in BH	Acronym for Sistema Nacional de Emprego (SINE), public job search service. Acronym for Belo Horizonte (BH), Brazilian city
89	Vagas SINE Contagem	Vacancies SINE Contagem	Acronym for Sistema Nacional de Emprego (SINE), public job search service. Brazilian city (Contagem)
90	Vagas Uberlandia	Jobs uberlandia	Brazilian city (Uberlândia)
91	Vagas urgentes BH	Urgent Vacancies BH	Acronym for Belo Horizonte (BH), Brazilian city
92	vagas.com.br		Job advertiser's name.
93	Varginha online		Job advertiser's name. Brazilian city (Varginha)
94	Vli		Job advertiser's name.

A3: State-space representation of the dynamic factor model

The state-space formulation of the dynamic factor model is given by:

$$\mathbf{z}_t = \mathbf{H}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \tag{20}$$

$$\boldsymbol{\alpha}_t = \mathbf{G}\boldsymbol{\alpha}_{t-1} + \mathbf{W}\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \qquad t = 1, \dots, T \tag{21}$$

Considering the case with r factors and n* series of Google Trends, the state vector α_t is defined as:

$$\boldsymbol{\alpha}_t' = [L_t \quad R_t \quad S_{1,t} \quad S_{1,t}^* \quad \dots \quad S_{6,t} \quad e_t \quad \dots \quad e_{t-2} \quad f_{1,t} \quad \dots \quad f_{r,t}] \tag{22}$$

$$\boldsymbol{\alpha}_{t-1}' = [L_{t-1} \quad R_{t-1} \quad S_{1,t-1} \quad S_{1,t-1}^* \quad \dots \quad S_{6,t-1} \quad e_{t-1} \quad \dots \quad e_{t-3} \quad f_{1,t-1} \quad \dots \quad f_{r,t-1}] \tag{23}$$

The system matrices are:

$$\mathbf{H}_t = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & \hat{c}_t & 0 & 0 & \mathbf{0}_{1\times r} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boldsymbol{\Lambda}_{n^*\times r} \end{bmatrix} \tag{24}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} e_t \\ \boldsymbol{\xi}_t \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \sigma_I^2 & \mathbf{0}_{1\times r} \\ 0 & \boldsymbol{\Psi}_{n^*\times r} \end{bmatrix} \tag{25}$$

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & \mathbf{0}_{1\times 11} & \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times r} \\ 0 & 1 & \mathbf{0}_{1\times 11} & \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times r} \\ \mathbf{0}_{11\times 1} & \mathbf{0}_{11\times 1} & \mathbf{G}^{(S)} & \mathbf{0}_{11x3} & \mathbf{0}_{11\times r} \\ \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 11} & \mathbf{G}^{(e)} & \mathbf{0}_{3\times r} \\ \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times 11} & \mathbf{0}_{r\times 3} & \mathbf{I}_r \end{bmatrix} \tag{26}$$

$$\mathbf{G}^{(S)} = \begin{bmatrix} 0.87 & 0.50 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.50 & 0.87 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.50 & 0.87 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.87 & 0.50 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.50 & 0.87 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.87 & -0.50 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.87 & 0.50 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.50 & -0.87 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{27}$$

$$\mathbf{G}^{(e)} = \begin{bmatrix} 0 & 0 & \hat{\phi} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{28}$$

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \mathbf{0}_{13\times r} \\ 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \\ & \mathbf{0}_{r\times 4} & & & \mathbf{I}_r \end{bmatrix}, \boldsymbol{\eta}_t = \begin{bmatrix} \eta_{R,t} \\ \eta_{S,t} \\ \eta^*_{S,t} \\ \eta_{e,t} \\ \mathbf{u}_t \end{bmatrix} \tag{29}$$

$$\mathbf{Q} = \begin{bmatrix} \sigma_R^2 & 0 & \mathbf{0} & \mathbf{Q}^{(\rho)} \\ 0 & \sigma_S^2 & 0 & \mathbf{0}_{1\times r} \\ 0 & 0 & \sigma_e^2 & \mathbf{0}_{1\times r} \\ \mathbf{Q}^{(\rho)} & \mathbf{0}_{r\times 1} & \mathbf{0}_{r\times 1} & \mathbf{I}_r \end{bmatrix} \tag{30}$$

$$\mathbf{Q}^{(\rho)} = \begin{cases} \mathbf{q}_{ii'}^{(\rho)} = \rho_{R,\mathbf{f}_r} \cdot \sigma_{R,t}, & i \neq i' \\ \mathbf{q}_{ii'}^{(\rho)} = 0, & i = i' \end{cases}, \quad i, i' = 1, \cdots, r \tag{31}$$

A 4 Results of targeting the predictor strategies – selected models

Table 10: Estimated coefficients for penalized elastic net regression – Brazil

Variables	Coef.
Intercept	-1.67
trabalhe conosco	1.38
vagas de emprego pe	7.51
classificados	2.52
vagas rj	1.36
modelo currículo	2.70
banda b empregos	3.46
olx curitiba	2.92
vagas indeed	0.42
meu primeiro emprego	-2.71
currículo para primeiro emprego	1.60
vagas sine hoje	-0.63
vagas sine rj	1.55
alpha	1
lambda	15
RMSE	87.28
Rsquared	0.04
MAE	69.08
RMSESD	16.89
RsquaredSD	0.04
MAESD	14.49

Note: The other coefficients were estimated equal to zero. The regression considers the first difference of the unemployment slope with the first difference of the Google Trends series.

Table 11: Classification of terms according to cluster – complete data – Minas Gerais (continue)

ID	Term	Cluster	ID	Term	Cluster
0	UNEMPLOYMENT_BLFS	1	49	vaga de	10
1	fazer currículo	1	50	vaga de emprego	11
2	procuro emprego	1	51	currículo pdf	12
3	vagas de emprego	1	52	vaga emprego	12
4	catho	2	53	trabalho home office	13
5	jf emprego	2	54	vagas	13
6	objetivos para curriculo	2	55	vagas de emprego sine	14
7	oportunidades bh	2	56	vagas pe	14
8	rio vagas	2	57	sine empregos	15
9	vagas de emprego em uberlandia	2	58	vagas emprego	15
10	vagas emprego uberlandia	2	59	emprego em bh	16
11	vagas sine contagem	2	60	empregos bh	17
12	comunidade de emprego	3	61	empregos mg	18
13	sine emprego	3	62	sine contagem	19
14	concursos	4	63	vagas pitagoras	19
15	currículo	4	64	vagas sine bh	19
16	emprego bh	4	65	empregos uberlandia	20
17	infojobs	4	66	vli	20
18	jfempregos	4	67	emprego contagem	21
19	vagas de emprego em contagem	4	68	emprego uberlandia	21
20	vagas em bh	4	69	jornal balcao	21
21	emprego	5	70	ipecont vagas	22
22	sest senat vagas	5	71	vagas de emprego sine bh	22
23	vagas bh	5	72	varginha online	23
24	vagas de estagio	5	73	indeed bh	24
25	vagas designação	5	74	mais emprego	24
26	vagas indeed bh	5	75	como fazer currículo	25
27	bhjobs	6	76	curriculum	25
28	catho empregos	6	77	minas curriculo	25
29	currículo primeiro emprego	6	78	vagas uberlandia	25
30	emprego montes claros	6	79	fato real empregos	26
31	empregos	6	80	fazer currículo online	26
32	indeed vagas	6	81	olx uberlandia	26
33	vagas contagem	6	82	vagas de emprego contagem	26
34	vagas mg	6	83	vagas de emprego mg	26

Table 11: Classification of terms according to cluster - complete data - Minas Gerais

(conclusion)

ID	Term	Cluster	ID	Term	Cluster
35	indeed	7	84	curriculo pronto word	27
36	sine	7	85	currículo vitae	27
37	sine bh	7	86	objetivo para currículo	27
38	vagas no sine	7	87	vagas de emprego uberlandia	28
39	curriculum vitae	8	88	vagas emprego bh	28
40	olx empregos	8	89	jf empregos	29
41	currículo para primeiro emprego	9	90	sine de bh	30
42	olx mg	9	91	sine de contagem	31
43	salário	9	92	educa mais brasil	32
44	sine bh empregos	9	93	vagas urgentes bh	32
45	vagas de emprego no sine	9	94	vagas de emprego sete lagoas	33
46	vagas no sine bh	9			
47	vagas sine	9			
48	vagas.com.br	9			

Table 12: Estimated correlations and likelihood ratio tests in the clustering combined with bivariate structural models - youth unemployment – Brazil

Terms	ρ^R_{xy}	LR test - ρ^R_{xy} = 0 (p-value)
curriculo word	0.7585	0.02
mais empregos	0.3178	0.28
sine idt	0.7063	0.01
vagas de emprego fortaleza	0.6264	0.05