Socio-Economic Planning Sciences 95 (2024) 101963

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps

Real-time nowcasting the monthly unemployment rates with daily Google Trends data

Eduardo André Costa [a,*], Maria Eduarda Silva [a,b], Ana Beatriz Galvão [c]

[a] *University of Porto, School of Economics and Management, Rua Dr. Roberto Frias s/n, Porto, 4200-464, Portugal*
[b] *INESC-TEC, LIAAD, Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal*
[c] *University of Warwick, Department of Economics, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom*

ARTICLE INFO

Keywords:
Google Trends
Predictors
Nowcasting
Unemployment rate
Mixed Data Sampling
Portugal

ABSTRACT

Policymakers often have to make decisions based on incomplete economic data because of the usual delay in publishing official statistics. To circumvent this issue, researchers use data from Google Trends (GT) as an early indicator of economic performance. Such data have emerged in the literature as alternative and complementary predictors of macroeconomic outcomes, such as the unemployment rate, featuring readiness, public availability and no costs. This study deals with extensive daily GT data to develop a framework to nowcast monthly unemployment rates tailored to work with real-time data availability, resorting to Mixed Data Sampling (MIDAS) regressions. Portugal is chosen as a use case for the methodology since extracting GT data requires the selection of culturally dependent keywords. The nowcasting period spans 2019 to 2021, encompassing the time frame in which the coronavirus pandemic initiated. The findings indicate that using daily GT data with MIDAS provides timely and accurate insights into the unemployment rate, especially during the COVID-19 pandemic, showing accuracy gains even when compared to nowcasts obtained from typical monthly GT data via traditional ARMAX models.

1. Introduction

The strength of labour markets is conventionally assessed through macroeconomic indicators, with the unemployment rate playing a pivotal role in this evaluation. However, this and other official statistics, which rely on surveys, are announced with delays, hampering the clear sight of the current state of the economy.

In response to the pressing need for timely and accurate macroeconomic indicators, a variety of alternative data sources, from administrative records to web scraping, are currently being used to generate nowcasts. Nowcasts provide immediate estimates of the current economic state informed by leading indicators. Among these indicators, online activity data is gaining traction due to its rapid availability and the absence of associated costs. The integration of online activity data into research methodologies has been on the rise, serving to refine preliminary estimates, anticipate announcement dates [1], supplement conventional data sources [2] and offer a more comprehensive view of macroeconomic indicators [3].

Among online activity data, search engine data, reflecting online search behaviour, offers a glimpse into users' needs, concerns, and interests [4] and has been used for augmented forecasting approaches. The exploration of Google Trends (GT) data is of particular interest in the literature. GT provides an updated measure of Google search engine users' interest in a subject or a specific keyword over time, given a geographical location. Data from this source are featured in *quasi*-real-time and available at various frequencies, from hourly to monthly. The literature has been focusing on its potential to predict unemployment figures, such as unemployment rates, the number of unemployed individuals and the quantity of initial claims for unemployment benefits. Over the past decade, numerous studies have scrutinised the predictive capabilities of GT data for unemployment metrics across various countries. The typical finding is that predictions based on GT data tend to outperform traditional benchmarks in forecasting unemployment metrics. This conclusion is supported by research conducted for the UK [5–7], Spain [e.g., 8–10], the USA [e.g., 7,11,12], Portugal [13,14], France [13,15], Italy [13,16], Türkiye [17], Romania [18], the Czech Republic, Hungary and Poland [19], and Canada and Japan [7]. However, it is worth noting that GT data has been found to underperform against a benchmark for Slovakia [19] and Germany [7] data.

In addition, GT data as unemployment predictors of the US unemployment data have served to forecast major economic events, such as the 2008 financial crisis and the COVID-19 pandemic, with studies revealing the ability of GT data to enhance forecast accuracy. Studies

[*] Corresponding author.
E-mail addresses: up201800115@edu.fep.up.pt (E.A. Costa), mesilva@fep.up.pt (M.E. Silva), ana.b.galvao@proton.me (A.B. Galvão).

https://doi.org/10.1016/j.seps.2024.101963
Received 19 July 2023; Received in revised form 7 February 2024; Accepted 28 May 2024
Available online 31 May 2024
0038-0121/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

report accurate improvements over benchmarks, showcasing the resilience and efficacy of GT data in capturing the dynamics of labour markets in disturbed economic scenarios [e.g., 20–22].

GT data have also been employed in assessing youth unemployment, a global concern. These data have been shown to perform well in predicting the youth unemployment rate in countries such as France [15], Italy [16] and Spain [23]. Other studies have also investigated the digital divide among different demographic groups, such as those defined by sex, age or race [see, 24,25]. These studies have further examined the bias in the selection of Internet audiences. They associate the accuracy of unemployment predictions with the disparities in access to and usage of digital technologies and the Internet. This in-depth understanding provides valuable insights into the challenges of using GT data to forecast a wide range of demographic segments.

The integration of the publishing timeline and high-frequency sampling predictors enables the production of nowcasts. The varied sampling frequencies provided by GT allow researchers to use a range of forecasting techniques, adapting to the different publication frequencies of various unemployment indicators. Most studies use GT data at monthly or weekly sampling frequencies as predictors of unemployment. A possible explanation for these frequency choices is that Google Trends limits the availability of comprehensive historical data at high frequencies (daily and hourly). The main forecasting approaches so far in the literature involve classic time series models, incorporating (V)ARMAX modelling with or without seasonal and error correction terms [e.g., 5,8,11,13,14,16,19,20,22,24–26]. Mixed frequency models [e.g., 6,12,27] and Bayesian approaches [e.g., 7,23] are also considered, among others [e.g., 15,21]. Consequently, the publication timeline for unemployment metrics can influence the choice of prediction approach in terms of sampling frequencies, such as opting for a mixed-frequency model or a more traditional method. Specifically, the delay between the end of a period and the publication schedule can favour one prediction approach over another, depending on the unemployment indicator and the frequency at which a GT predictor is made public. For instance, a single sampling frequency is preferred when modelling the monthly number of individuals registering for employment at public employment offices in Spain due to the slight one-week delay in releasing the official statistics after the reference period [25]. In this case, the short delay (7/30 days) favours a more traditional approach, forecasting the monthly unemployment metric with monthly GT predictors. Conversely, the weekly US unemployment insurance claims count, which becomes publicly available after five days (5/7 days), suggests that using daily GT predictors [27] makes the mixed-frequency model suitable.

This paper contributes to the growing field of economic forecasting that incorporates alternative data sources to enhance the timeliness and accuracy of predictions. In light of this, it focuses on real-time nowcasting of the Portuguese unemployment rate between January 2019 and November 2021. The official unemployment rate in Portugal is published with a two-month delay from the end of a reference period (60/30 days). Consequently, this paper constructs a framework to nowcast monthly unemployment rates using GT series as predictors, which considers the timing between the official unemployment rate publication and the availability of GT data. To this end, two approaches for producing nowcasts are envisaged. The first involves using daily GT predictors via MIxed DAta Sampling (MIDAS) regressions. The second approach adopts GT monthly series as predictors in AutoRegressive Moving Average with eXogeneous input (ARMAX) models, which aligns with the traditional approach in the Portuguese literature related to this subject. The objectives of this paper are a valuable addition to the toolkit of policymakers and economists who need to make informed decisions in a rapidly changing economic landscape.

Complementing the current body of literature, this paper makes distinctive contributions in nowcasting the unemployment rate using GT data: (i) it generates real-time nowcasts that account for the official unemployment rate's publication delay, not considered in previous literature using Portugal's unemployment data and GT series; (ii) it employs MIDAS regressions with an extensive GT-daily series as predictors, a framework not yet utilised in the Portuguese unemployment-GT-related literature; (iii) the paper unveils the prediction accuracy of daily GT predictors compared to monthly GT predictors, the predominant sampling frequency of previous studies, which is unprecedented to the authors' knowledge; (iv) the study analyses the predictive power of nowcasts within the context of the COVID-19 pandemic, a novelty among Portuguese unemployment-GT-related literature; (v) it compares the predictions developed to the provisional unemployment statistics provided by the Portuguese National Statistical Office, an innovative aspect with Portuguese data.

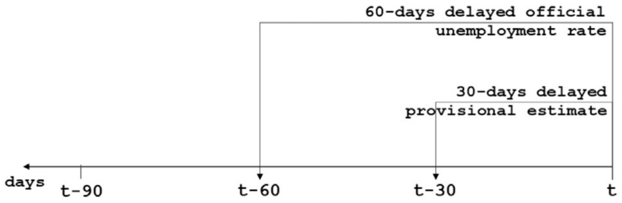


Fig. 1. The Portuguese unemployment rate publishing timeline.

The remainder of this paper is organised as follows. Section 2 presents the data and details the construction of the Google Trends predictors. Section 3 specifies the modelling and predictive approaches adopted to generate the predictions, and Section 4 designs the methodology utilised. The empirical findings are described and discussed in Section 5, and Section 6 presents the concluding remarks.

2. Data

The data sets considered in this work are the monthly Portuguese unemployment rate and the daily and monthly Google Trends time series related to proxies concerning the labour situation. The following sections describe the datasets, establish the procedure to collect Google Trends data and construct predictors relevant to the purpose of this study.

2.1. Unemployment data

On average, Portuguese official unemployment rates[1] are issued monthly with a delay of 60 days from the end of a reference month, while provisional estimates are available with a delay of 30 days. As illustrated in Fig. 1, at the end of a calendar month, designated by time t, the Portuguese National Statistical Office (INE) issues the unemployment rate for $t-60$ and a provisional estimate for $t-30$.[2]

The official monthly seasonally-adjusted Portuguese unemployment rate analysed in this work is obtained from the European Union Statistical Office [28], comprising data from January 2004 to November 2021. Provisional unemployment rates are from INE Monthly Reports [29]. These estimates concern data from January 2019 to February 2020 and April 2020 to November 2021 since INE could not supply a provisional unemployment rate for March 2020 due to the COVID-19 pandemic.

The official Portuguese unemployment rates between 2004 and 2021[3] are illustrated in Fig. 2. From the third quarter of 2008 to early

[1] Appendix A contains details on the Portuguese unemployment rate definition.

[2] The Portuguese unemployment rate publishing schedule for 2019 is detailed in Appendix B.

[3] The Portuguese unemployment figures have been impacted by government measures related to COVID-19, including lockdowns and layoffs, which have influenced the classification of individuals in terms of their conditions related to work [30].

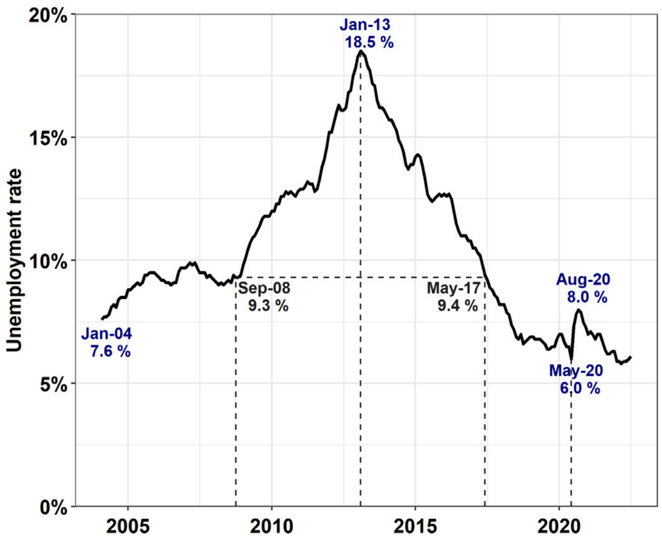


Fig. 2. Monthly seasonally adjusted official Portuguese unemployment rate from January 2004 to November 2021 — own figure based on Eurostat [28].

2013; the country experienced an upward trend in unemployment due to the 2008 global financial crisis. After a peak in January 2013, the unemployment rate decreased, reaching the level of the pre-2008 turmoil in May 2017. Afterwards, the data showed some mild oscillations until the impact of the coronavirus disease outbreak started emerging in May 2020.

2.2. Google Trends data

Google search engine data have been publicly available via GT since 2006, comprising anonymous data from 2004 onwards. GT provides readily available time series describing the relative demand for keywords or categories[4] through the Google Index (GI), a normalised measure given a geographical area and period [11]. Users can target keywords as search criteria from the whole Google search universe or Google's categories. Also, the searches can target the demand for a specific Google category without relying on keywords.

The GI is normalised relative to the highest demand in a given area during the period queried. The index ranges from 0, indicating no relevant quantity of requests for a keyword, to 100, the maximum popularity [6]. It may also assume the label "<1" for insignificant volumes. Thus, given a maximum value for a period, GI represents the demand variation relative to the maximum in the searched period for the geographical domain.

The main apprehension regarding GT data relates to the selection bias triggered by internet user characteristics regarding sex, income and age [e.g., 5,12,16]. Additionally, two issues related to GT data extraction deserve attention. The first concerns the noise that GT sampling methods induce [16]. Indexes collected on different days are slightly diverse [20], even when all collection rules are constant. This issue is overcome by averaging the data collected over multiple days [3,5,13,31]. The other constraint is the limited historical data in high-frequency samplings [21]. GT limits daily series to approximately nine months of data and weekly series from nine months to five years, whereas monthly sampling grants more than five years of data. Because of the normalised characteristic of the index, stacking data frames to obtain extended series in high-frequency data of weekly and daily series would conceal the eventual presence of trends.

To address the restricted historical data in high-frequency samplings, Eichenauer et al. [31] developed a procedure for obtaining extended daily GT series consistent with weekly and monthly sampling frequencies, thus reflecting long-run trends. This method collects GT data given a keyword and a geographical area over multiple days and sampling frequencies (daily, weekly and monthly) using shifted overlapping rolling windows. It resorts to averaging data on intersecting periods across the overlapping windows and over the multiple collections, which addresses the sampling noise issue and results in a daily, weekly and monthly series. Finally, over these resulting series, the Chow and Lin [32] disaggregation procedure is then applied to low-frequency (LF) series, monthly and weekly, to obtain a high-frequency (HF) series, daily, consistent with both LF series.

The typical search terms to explore unemployment issues include *job(s)* [3,5,11,18,19,33], derivative terms like *job offer(s)* [8,10,16,33], the keyword *unemployment* [5,9,12–14,34], as well as names of local job search engines or job placement agencies [10,17,35,36]. The keyword selection strategy adopted in this work considers the most frequent search terms in the literature as well as Portuguese cultural aspects, leading to four keywords[5]:

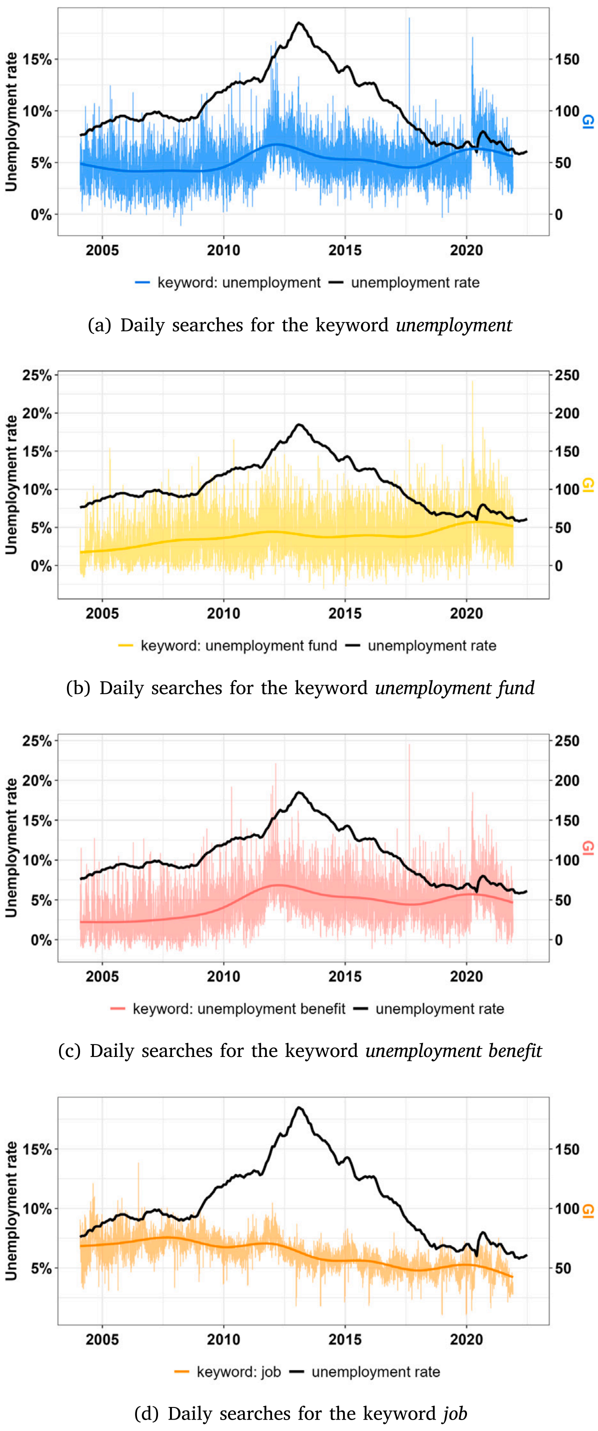
- *desemprego* (unemployment)
- *fundo desemprego* (unemployment fund)
- *subsídio desemprego* (unemployment benefit)
- *emprego* (job)

GT time series for each keyword are collected in two steps circumscribed to searches performed in Portugal. The first step aims to obtain monthly GT series following the regular literature on unemployment-GT. It comprises the collection of monthly GT series for each search term covering data from January 2004 to November 2021. The collection is repeated over 18 days and then averaged over the multiple extractions to overcome the sampling variability intrinsic to GT. The second step seeks to build an extensive daily GT series for each search term following the Eichenauer et al. [31] procedure. It comprises the collection of daily, weekly, and monthly GT time series intended to build a daily series that spans the period from January 2004 to November 2021. To this end, for each keyword, daily, weekly and monthly GT data are collected using overlapping rolling windows as follows[6]: the daily data collection uses six-month time windows moved by 15 days (e.g., from 01-Jan-04 to 01-Jul-04, from 16-Jan-04 to 16-Jul-04, and so on), resulting in 425-time series with intersecting dates; the weekly data extraction considers five-year time windows shifted by 11 weeks (e.g., from 01-Jan-04 to 01-Jan-09, from 18-Mar-04 to 18-Mar-09), resulting in 63-time series with intersecting weeks; the monthly data collection employs 15-year time windows moved by a month (e.g., from 01-Jan-04 to 01-Jan-19, from 01-Feb-04 to 01-Feb-19), thus generating 35-time series with coincident months. The intersecting periods in each sampling frequency are summarised into averages. Repeating the procedure on 18 different days results in 54 series for each keyword: 18 at daily, 18 at weekly and 18 at monthly frequencies. Subsequently, for each frequency and each keyword, the 18 series are averaged, resulting in three series — one daily, one weekly and one monthly. Finally, for each keyword, the Chow and Lin [32] temporal disaggregation procedure is performed twice: first, disaggregating monthly to weekly and, then, disaggregating the resulting weekly to daily. This process

[4] Google assigns categories to the searches performed in the search engine using natural language classification [20]. Examples of categories include *Jobs* and *Welfare & Unemployment*.

[5] Preliminary experiences also considered the keyword *sapo empregos* (the name of a relevant job search portal in Portugal) and Google's categories *Jobs* and *Welfare & Unemployment*. These were excluded from this paper to present relevant results.

[6] Rules for overlapping rolling windows are summarised in tabular form in Appendix C.



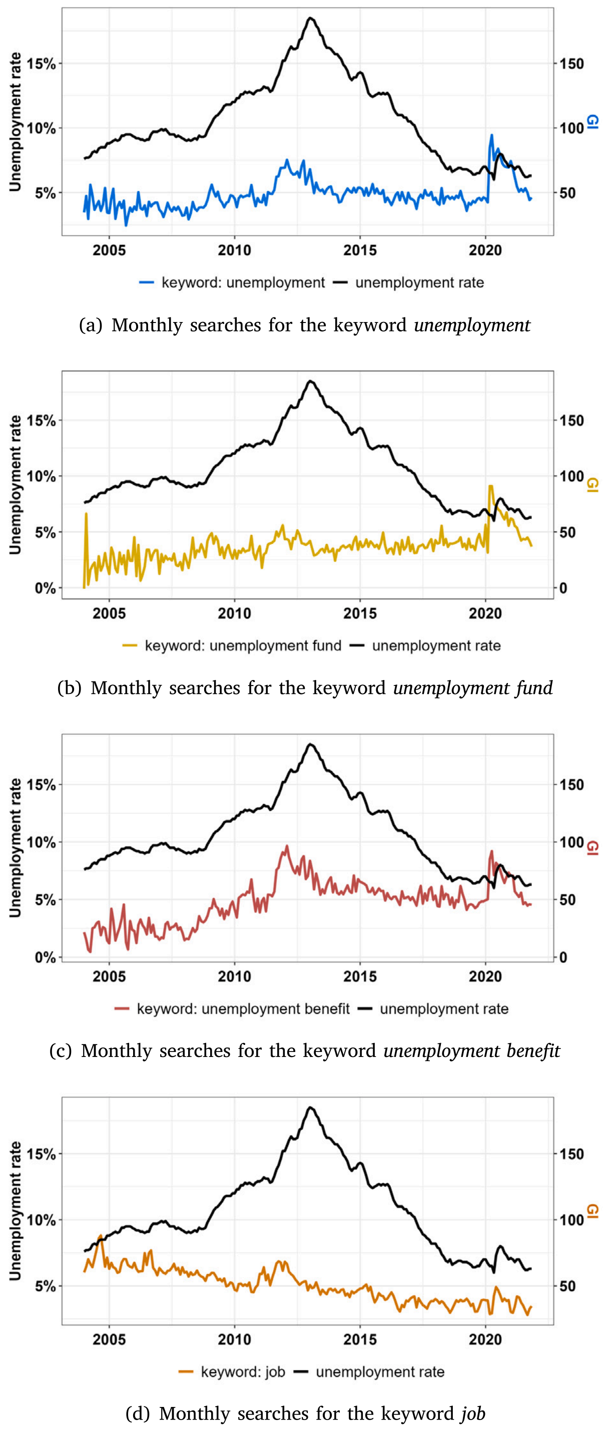
(a) Daily searches for the keyword *unemployment*

(b) Daily searches for the keyword *unemployment fund*

(c) Daily searches for the keyword *unemployment benefit*

(d) Daily searches for the keyword *job*

Fig. 3. GT daily series, its smoothed trend lines and the Portuguese official unemployment rate from January 2004 to November 2021.



(a) Monthly searches for the keyword *unemployment*

(b) Monthly searches for the keyword *unemployment fund*

(c) Monthly searches for the keyword *unemployment benefit*

(d) Monthly searches for the keyword *job*

Fig. 4. GT monthly series and the Portuguese official unemployment rate from January 2004 to November 2021.

results in an extensive daily GT series for each keyword consistent with weekly and monthly sampling frequencies.

It is essential to highlight that the second step relies on the methodology proposed by Eichenauer et al. [31], specifically designed to construct a comprehensive daily series by collecting multi-frequency data via overlapping windows and employing temporal disaggregation. While this approach also gathers weekly and monthly GT series, their prompt utilisation is not advisable, thus requiring pre-processing before their use. Further investigation into alternative methods is necessary to obtain an extensive weekly series, especially considering that long monthly series can be collected directly, as executed in the first step.

Figs. 3 and 4 display the daily and monthly GT time series seasonally adjusted by Prophet procedures [37]. Some keywords present clear leading signals related to either the unemployment peak in January 2013 or the shock caused by the COVID-19 pandemic from May 2020, such as Figs. 3(a), 3(c), 4(a) and 4(c).

3. The econometric models

Considering the data collection details outlined in Section 2, we utilise the multiple GT series in monthly and daily frequencies to predict the monthly official unemployment rate. For the sake of simplicity, the equations presented in this section employ single explanatory variables, consistent with the empirical prediction application of this study. Expanding these equations to include multiple predictors is straightforward.

Two modelling approaches are assumed to nowcast unemployment rates using GT data issued at a higher pace than the delayed emission of the official unemployment rate. The first approach relies on the most popular method in GT-unemployment literature, the ARMAX model, using solely monthly-frequency series, i.e., considering Google monthly series as predictors of the monthly unemployment rate. To set the notation, let y_t denote a stationary time series satisfying an ARMA(p, q) model given by

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \tag{1}$$

where ε_t is a white noise process, i.e., a sequence of independent and identically distributed random variables with zero mean and variance σ_ε^2. The ARMA(p, q) model in Eq. (1) can be extended to consider an exogenous predictor x_t sampled at the same frequency as y_t, leading to the following ARMAX model

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t + \sum_{k=0}^{K} \beta_k x_{t-k}. \tag{2}$$

The second approach resorts to MIDAS models to accommodate the sampling imbalance between the daily-frequency GT time series and the monthly unemployment rate. An overview of these models is presented in the following subsection.

3.1. MIDAS regression

MIDAS models [38] have emerged in econometrics literature over the past decades and aim to estimate the relationship between low-frequency (LF) time series and high-frequency (HF) predictors, preserving the relationship across different frequencies.

Consider that y_t^L denotes a stationary LF-dependent time series, and x_t^H represents a stationary HF-predictor time series. In addition, consider that the sampling frequency of x_t^H is m times that of y_t^L. The basic MIDAS equation involving a single low-and-high-frequency variable to forecast h-step-ahead, as first introduced by Ghysels et al. [38], can be written as[7]

$$y_t^L = \beta_0 + \beta_1 B(L^{1/m}; \boldsymbol{\theta}) x_{t-h}^H + \varepsilon_t^L, \tag{3}$$

where the error term, ε_t^L, is a sequence of independent and identically distributed random variables with mean zero and variance σ_ε^2, and independent of x_t^H. The term $B(L^{1/m}; \boldsymbol{\theta})$ denotes the weighting scheme applied to x_t^H, and determined by the function, $b(k; \boldsymbol{\theta})$, as follows

$$B(L^{1/m}; \boldsymbol{\theta}) = \sum_{k=0}^{K} b(k; \boldsymbol{\theta}) L^{k/m} \tag{4a}$$

with

$$\sum_{k=0}^{K} b(k; \boldsymbol{\theta}) L^{k/m} = 1. \tag{4b}$$

The term $B(L^{1/m}; \boldsymbol{\theta})$ in Eq. (4a) assigns weights, restricted to sum to one, to the K lags of x_t^H using the lag operator $L^{k/m}$. This operator shifts x_t^H back in time by k/m periods such that $L^{k/m} x_t^H = x_{t-k/m}^H$, shaping a distributed lag polynomial. The functional form of $b(k; \boldsymbol{\theta})$ assumes the following representation

$$b(k; \boldsymbol{\theta}) = \frac{\psi(k; \boldsymbol{\theta})}{\sum_{k=0}^{K} \psi(k; \boldsymbol{\theta})}, \tag{5}$$

satisfying the restriction in Eq. (4b).

Several specifications for the function $b(k; \boldsymbol{\theta})$ are available [39]. Two popular specifications [40] are the Beta lag function, defined as

$$b(k; \theta_1, \theta_2) = \frac{f(\frac{k}{K}; \theta_1, \theta_2)}{\sum_{k=0}^{K} f(\frac{k}{K}; \theta_1, \theta_2)}, \tag{6a}$$

where $f(\cdot)$ is the Beta probability density, according to

$$f(z, a, b) = \frac{z^{a-1}(1-a)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{6b}$$

$$\Gamma(a) = \int_0^{\infty} \exp^{-z} z^{a-1} dz, \tag{6c}$$

and the exponential Almon lag function, given by

$$b(k; \boldsymbol{\theta}) = \frac{\exp(k\theta_1 + \cdots + k^Q \theta_Q)}{\sum_{k=0}^{K} \exp(k\theta_1 + \cdots + k^Q \theta_Q)}, \tag{7}$$

where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_Q\}$, with Q denoting the number of shape parameters. This function with only two shape parameters, $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, allows lagged coefficients to assume increasing, decreasing or hump-shaped designs[8] [39].

Involving a weighting function with a few parameters can parsimoniously capture the dynamics between the mixed sampling series, avoiding the need to estimate multiple coefficients that would result from the high-frequency-lagged explanatory series otherwise. As such, the use of Eq. (3) with the Beta lag or the exponential Almon of order two into the weighting scheme requires the estimation of only four parameters, i.e., β_0, β_1, θ_1 and θ_2, regardless of K, the number of HF lags, thus favouring large m ratios, as in the case of y_t^L sampled at a monthly frequency and x_t^H at daily — accordingly, the parameters in Eq. (3) are estimated via non-linear least squares.

The concept of leads in MIDAS regressions [41] refers to using additional i/m observations available at a nowcasting time, hence, observing the HF predictor up to $x_{t+i/m}^H$. The accommodation of leads in Eq. (3) is given by Eq. (8), maintaining the original MIDAS regression features.

$$y_t^L = \beta_0 + \beta_1 B(L^{1/m}; \theta_h) x_{t+i/m-h}^H + \varepsilon_t^L \tag{8}$$

Another variant of MIDAS regression does not enforce constraints on the weighting scheme; thus, it is termed Unrestricted MIDAS (U-MIDAS). The linear regression function provided by U-MIDAS [42] is defined as

$$y_t^L = \beta_0 + \sum_{k=0}^{K} \beta_k x_{t-k-h}^H + \varepsilon_t^L. \tag{9}$$

In this variant, the ordinary least squares (OLS) method consistently estimates the unknown parameters as long as the error term follows the usual OLS specifications and the lag length size of x_t^H is adequate to obtain uncorrelated errors. Foroni et al. [42] compared the forecasting accuracy of U-MIDAS to the standard MIDAS in a simulated experiment, with the former outperforming the latter, especially when there is no significant difference in sampling frequencies. Accordingly, in U-MIDAS regression, the number of estimated parameters grows with the number

[7] Although suppressed in the notation for the sake of simplification, the parameters on MIDAS regressions are h-specific, implying that given the choice of h and available data, the model estimation relies on distinct information sets, producing direct forecasts.

[8] Appendix D illustrates, from the exponential Almon polynomial with a bivariate parameter and the Beta lag, examples of the weighting shapes on 20 high-frequency lags.

of HF-lagged terms, favouring modest sampling differences between the series, such as y^L_t observed quarterly and x^H_t monthly.

A recently developed variant, TF-MIDAS [43], employs transfer functions to establish the relationship between x^H_t and y^L_t target. This method offers a more comprehensive representation than previous models based on distributed lags by eliminating the truncation error and incorporating a moving average component into the model. A TF-MIDAS model is defined by two equations given by

$$y^L_t = \beta_0 + \sum_{k=0}^{K} \frac{a_k(Z)}{b_k(Z)} x^H_{t-k-h} + \eta^L_t \tag{10a}$$

$$\phi(Z)\eta^L_t = \theta(Z)\varepsilon^L_t. \tag{10b}$$

In these expressions, $a_k(Z)$ and $b_k(Z)$ denote finite lag polynomials, where $Z \equiv L^K$. Additionally, $\phi(Z)$ and $\theta(Z)$ are autoregressive and moving average polynomials of orders p and q, respectively, aligning with the standard processes in the ARMA framework. The TF-MIDAS model is estimated using the exact maximum likelihood method and may involve many estimated parameters contingent on the lag length. Like U-MIDAS, TF-MIDAS regression is recommended for modelling modest sampling differences between series.

Other variations of MIDAS regressions can be found in Foroni et al. [42] and Ghysels et al. [44].

The mixed-frequency approach is well-suited to build nowcasts combining the monthly Portuguese unemployment rate with publishing delay and the prompt available daily series from GT. Considering this and the substantial difference in sampling frequencies denoted by the wide m ratio between series, the standard MIDAS model with leads from Eq. (8) is employed. Once such a model is estimated, nowcasts for y^L_{t+h} conditional on information available at period $T + i/m$ is calculated as

$$\hat{y}^L_{t+h|T+i/m} = \hat{\beta}_0 + \hat{\beta}_1 B(L^{1/m}; \hat{\theta}) x^H_{t+i/m}. \tag{11}$$

Note that the parameters of Eq. (11) are h-specific, implying that given the choice of h and available data, the model estimation relies on distinct information sets, producing direct forecasts.

4. Methodology

This study develops two sets of nowcasts for the monthly Portuguese official unemployment rate from January 2019 to November 2021, assuming the monthly and daily GT series collected and constructed according to Section 2 as predictors. Each keyword in the set $\mathbb{W} = \{unemployment, unemployment\ fund, unemployment\ benefit, job\}$ gives rise to a GT predictor at monthly and daily frequencies. Each GT predictor at monthly frequency produces a nowcast using the ARMAX models, Eq. (2). Each GT predictor at daily frequency produces a nowcast using the MIDAS regression with leads, Eq. (8), given the *quasi*-real-time feature of GT data and the large sampling ratio between the official unemployment rate and daily GT series. In summary, at time t, each keyword from the set $\mathbb{W}$ generates a nowcast for a GT sampling frequency. Then, the four nowcasts of each GT frequency are averaged, resulting in a combined search term nowcast. Henceforward, nowcasts at time t based on keyword $\mathbb{W}$ daily GT predictor are represented by $\hat{y}^L_{\mathbb{W},t}$ and the combined nowcast is denoted as $\hat{y}^L_t = \frac{1}{4}\sum_{\mathbb{W}} \hat{y}^L_{\mathbb{W},t}$. Nowcasts based on monthly series are denoted as $\hat{y}_{\mathbb{W},t}$ and $\hat{y}_t = \frac{1}{4}\sum_{\mathbb{W}} \hat{y}_{\mathbb{W},t}$.

All the time series are tested for unit roots using the Augmented Dickey–Fuller [45] and undergo first differencing. Accordingly, in Eq. (8), y^L_t represents the monthly changes in unemployment rates, and x^H_t represents the daily changes in Google searches of a single keyword from $\mathbb{W}$. Similarly, in Eq. (2), y_t represents the monthly changes in unemployment rates, and x_t corresponds to monthly changes in Google searches for a single keyword from $\mathbb{W}$.

The weighting function, $b(k;\boldsymbol{\theta})$, employed within Eq. (8) is the exponential Almon lag formulation governed by two parameters to shape the distributed lag polynomial, Eq. (7) with $Q = 2$. To this end, null starting values initialise the optimisation process for $\boldsymbol{\theta}$ parameters.[9] In addition, a grid search is performed with a minimum of 5 and a maximum of 250 days to determine the length K of the weighting scheme in Eq. (4a), with the selection of K determined by the Bayesian Information Criterion (BIC). Furthermore, for robustness purposes, the employment of the Beta lag formulation, Eq. (6), is also used in an analogous selection of length K and with ones as initial values for optimisation, i.e., values equivalent to those used in exponential Almon lag formulation.

The orders p and q of the ARMAX models, Eq. (2), are determined by the Hyndman–Khandakar algorithm [46]. Additionally, the quantity of K lags from the predictor is selected according to the BIC minimisation within a grid search ranging between 0 and 12 months.

The general framework to build monthly predictions based on a single series of monthly or daily GT data is as follows. The information sets $\mathbb{I}_{\mathbb{W},t}$ available at a prediction time t, the end of the month, for a keyword from $\mathbb{W}$ contain changes in the official unemployment rate with a delay of 60 days and changes in predictors up to time t, $\mathbb{I}_{\mathbb{W},t} = \{\boldsymbol{Y}_{t-60}, \boldsymbol{X}_{\mathbb{W},t}\}$, which ensures real-time predictions and models specifications relying only on *ex-ante* information at the nowcasting time. Based on this, the nowcasts for the unemployment rates from January 2019 to November 2021, 35 in total, are built on expanding samples starting on 01-Jan-2004. Thus, the first nowcast is based on data available on 31-Jan-2019, namely historical changes in official unemployment rates from January 2004 to November 2018 and predictors comprising GT time series ranging from 01-Jan-04 to 31-Jan-19 for daily series and from January 2004 to January 2019 in the case of monthly data. The sample is then expanded with data corresponding to one month. As such, on 28-Feb-2019, the changes in official unemployment rates nowcast for February 2019 is issued based on the unemployment data from January 2004 to December 2018, changes in the GT daily series from 01-Jan-04 to 28-Feb-19 and changes in the GT monthly series from January 2004 to February 2019. Accordingly, the 35th nowcast is issued for the difference in unemployment of November 2021 on the 30th of November 2021. This nowcast is built based on unemployment information from January 2004 to September 2021 and GT daily predictors ranging from 01-Jan-04 to 30-Nov-21 or GT monthly predictors ranging from January 2004 to November 2021.[10] Finally, as explained previously, at each time t, the final nowcasts are obtained by averaging over the nowcasts from the set $\mathbb{W}$ at daily and monthly frequencies.

In addition to the GT-based nowcasts, predictions from two benchmark models are obtained, namely, Random Walk (RW) and the ARMA(p,q), from Eq. (1) with orders selected via the Hyndman-Khandakar algorithm [46], and denoted as $\hat{y}^{RW}_t$ and $\hat{y}^{ARMA}_t$, respectively. Note that both benchmarks rely exclusively on the monthly changes in unemployment data available at the time of the issue of the nowcasts, $\mathbb{I}_t = \{\boldsymbol{Y}_{t-60}\}$. Consequently, the RW forecasts represent the last observed change in unemployment, e.g., the January 2019 prediction is the observed change in unemployment of November 2018.

The forecasting performance is assessed with the Root-Mean-Squared Forecast Error (RMSFE), defined as follows

$$RMSFE^{*} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}^{*}_t - y_t)^2}{T}}, \tag{12}$$

where y_t is the observed change in unemployment in time t, T represents the number of forecasts under evaluation[11] and * denotes $\hat{y}^L_t$,

[9] Combinations among the pairs of starting values $\theta_1 = \{-0.5; -0.1; 0.1; 0.3; 0.4; 0.5; 0\}$ and $\theta_2 = \{-0.0025; -0.015; -0.01; 0\}$ were also evaluated. However, the combination $\boldsymbol{\theta} = (0, 0)$ performed best.

[10] Appendix E summarises the samples and information sets of the 35 nowcasts produced in this study.

[11] Over the complete set of forecasts, $T = 35$; over 2019 or 2020, $T = 12$; and over 2021, $T = 11$.

$\hat{y}_t$, $\hat{y}_t^{RW}$ or $\hat{y}_t^{ARMA}$. Also, the superscript considers predictions of each keyword from set $\mathbb{W}$ according to the sampling frequency.

The predictive accuracy of any two sets of predictions is compared via the modified version of the Diebold–Mariano (mDM) test [47], accounting for small samples [48]. This test uses a squared error loss-differential function, $d_t = e_{1t}^2 - e_{2t}^2$, where e_1 and e_2 represent the forecast errors. The test evaluates the equal predictive accuracy between the sets of predictions, $E[d_t] = 0$, as the null hypothesis against the alternative of a forecast set delivering statistically lower (higher) errors. Moreover, more than two sets of predictions for equal predictive capability are assessed via the test developed by Mariano and Preve [49], MultiDM, with finite-sample correction. This test acts as a comparative metric for predictions linked to multiple keywords, with the null hypothesis evaluating whether the forecasts demonstrate identical accuracy.

5. Results and discussion

This section analyses and discusses the results of the nowcasting procedures described above with a view to (i) evaluate whether Google Trends data are effective predictors to nowcast the Portuguese unemployment rate; (ii) compare daily and monthly Google Trends data as predictors of unemployment rate; (iii) assess the predictive performance of each keyword; (iv) compare Google Trends-based nowcasts and provisional estimates issued by the INE; (v) assess the impact of COVID-19 in the nowcasting procedures proposed in this work.

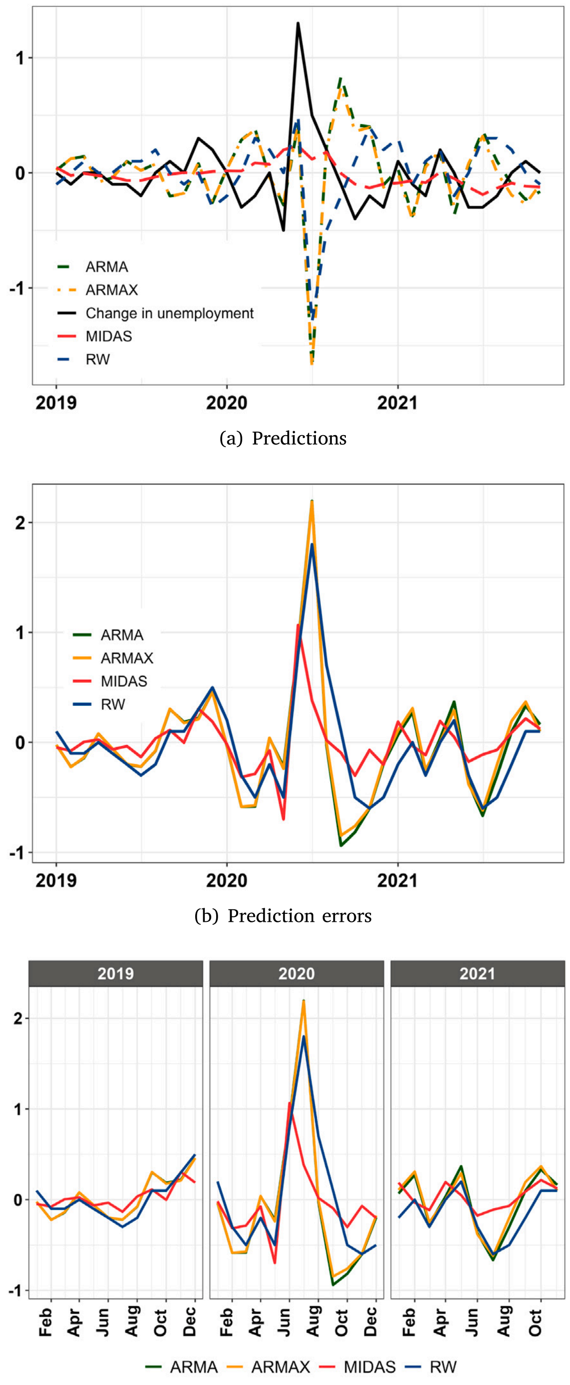
5.1. Evaluating daily and monthly GT predictors

Table 1 reports the out-of-sample RMSFEs for the nowcasts based on GT data at daily and monthly frequencies and benchmark predictions. In addition, Fig. 5 illustrates the forecasts and the corresponding errors.

The results indicate that nowcasts using daily GT data present significantly lower RMSFEs than benchmark predictions for the whole forecasting period (42% over RW and 50% over ARMA) and for each of the years 2019 (45% over RW and 44% over ARMA) and 2020 (40% over RW and 50% over ARMA). For 2021, nowcasts based on daily data outperform ARMA benchmark predictions by 57% RMSFE and are as accurate as RW predictions.[12] In addition, nowcasts built on daily predictors demonstrate 49% superior forecasting performance than those generated on monthly predictors for the entire nowcasting timeframe and consistently outperform across the years, with gains varying from 44%–55%. To further validate MIDAS predictions and ensure robustness in the use of the exponential Almon lag for designing the distributed lag polynomial, Appendix F indicates that the nowcasts derived from this specification are statistically equivalent to the Beta lag parametrisation.[13]

The nowcasts built on monthly GT data outperform the ARMA benchmark predictions, showing statistically significant error reductions over the entire forecasting period (2%), including 2019 (1%) and 2020 (2%). In contrast, monthly GT data nowcasts generally show no statistical difference compared to RW benchmark predictions, except for 2020, the first year of the COVID-19 pandemic, when GT monthly predictors significantly underperformed by 16%. Remarkably, while monthly GT data did not consistently outperform the RW, they offer a more reliable alternative to ARMA when predicting Portuguese unemployment.

Despite 2020 exhibiting higher RMSFEs than the other two years for all models, as evident in Table 1, MIDAS predictions consistently outperform benchmarks and ARMAX in this period. Statistically significant error improvements range from 40%–50%, underscoring the robust predictive capability of daily GT data in the chaotic context of the COVID-19 event.



(a) Predictions

(b) Prediction errors

(c) Prediction errors by year

Fig. 5. Out-of-sample predictions and prediction errors resulting from MIDAS, ARMAX and the RW.

5.2. Assessing the predictive performance of each keyword

An important issue when using GT data is the choice of keywords. Table 2 reports out-of-sample RMSFEs for the nowcasts based on GT data for each keyword by sampling frequency considered in this study. Although the difference in performance is minimal across keywords, the results indicate that when predicting with daily data, the keyword *unemployment* produces the overall lower nowcast errors, while when forecasting with monthly data, the keyword *job* stands out with reduced errors. However, the MultiDM test does not permit the assertion that

[12] This comparison has reached 53% in favour of daily predictors without statistical significance.

[13] A similar conclusion is drawn from Maas [12].

Table 1
Nowcasting performance.

Prediction model	RMSFE by prediction period			
	Total[a]	2019	2020	2021[b]
(A) MIDAS (based on daily predictors)	**0.2665**	**0.1205**	**0.4186**	**0.1374**
(B) ARMAX (based on monthly predictors)	0.5207	0.2144	0.8129	0.3028
(C) RW (benchmark)	0.4601	0.2179	0.7018	0.2908
(D) ARMA (benchmark)	0.5328	0.2157	0.8304	0.3168
Comparison (A) MIDAS vs (B) ARMAX	−49%*	−44%**	−48%*	−55%**
mDM *p*-value	(0.0607)	(0.0000)	(0.0707)	(0.0369)
Comparison (A) MIDAS vs (C) RW	−42%**	−45%*	−40%*	−53%
mDM *p*-value	(0.0346)	(0.0517)	(0.0512)	(0.1259)
Comparison (A) MIDAS vs (D) ARMA	−50%*	−44%**	−50%*	−57%*
mDM *p*-value	(0.0606)	(0.0000)	(0.0717)	(0.0656)
Comparison (B) ARMAX vs (C) RW	13%	−2%	16%*	4%
mDM *p*-value	(0.1023)	(0.3646)	(0.0642)	(0.3864)
Comparison (B) ARMAX vs (D) ARMA	−2%*	−1%**	−2%*	−4%
mDM *p*-value	(0.0579)	(0.0074)	(0.07474)	(0.1951)
Comparison (C) RW vs (D) ARMA	−14%*	1%	−15%**	−8%**
mDM *p*-value	(0.0912)	(0.5869)	(0.0611)	(0.0034)

mDM, one-sided modified Diebold–Mariano test: $^{**}p < 0.05$, $^{*}p < 0.10$.

[a] January 2019 to November 2021.

[b] January 2021 to November 2021.

Table 2
GT keywords prediction performance by sampling frequency.

Series frequency	Search terms	RMSFE by prediction period			
		Total[a]	2019	2020	2021[b]
Daily	*unemployment*	**0.2532**	0.1171	**0.3974**	**0.1294**
	unemployment fund	0.2864	**0.1081**	0.4532	0.1556
	unemployment benefit	0.2815	0.1452	0.4305	0.1644
	job	0.2998	0.1397	0.4675	0.1622
	MultiDM *p*-value	(0.6916)	(0.1734)	(0.9773)	(0.0005)
Monthly	*unemployment*	0.5295	0.2151	0.8242	0.3172
	unemployment fund	0.5309	0.2161	0.8282	0.3122
	unemployment benefit	0.5283	0.2165	0.8229	0.3132
	job	**0.5051**	**0.2104**	**0.7856**	**0.3002**
	MultiDM *p*-value	(0.9126)	(0.7510)	(0.8503)	(0.9266)

MultiDM, Multivariate DM test: $^{**}p < 0.05$, $^{*}p < 0.10$.

[a] January 2019 to November 2021.

[b] January 2021 to November 2021.

the tiny differences among the sets of keywords at the distinct frequencies are statistically different considering the whole period. Figs. 6 and 7 illustrate the nowcasts produced by daily and monthly GT data by search term and their combination.

5.3. Comparing GT-based nowcasts and INE provisional estimates

As outlined in Section 2, INE provides a provisional unemployment rate with a 30-day delay. Accordingly, the RMFEs in Table 3 compare the errors of this provisional estimative relative to the official unemployment rate with the nowcast errors obtained from monthly and daily GT predictors. From this, it is clear that for 2019, before the pandemic, the accuracy of nowcasts from both GT sampling frequencies is statistically equal to the accuracy of the provisional estimates. For 2020, 2021 and the complete forecasting period, nowcasts based on monthly predictors consistently underperform the provisional estimates. On the other hand, daily GT-based nowcasts outperform INE projections by 21% for 2021 and provide statistically equal performance for 2020 and the whole prediction timeframe. It is worth noting that nowcasts produced by daily GT predictors are issued concomitantly with month ends, indicating timelier predictions than INE provisional estimates with similar statistical accuracy.

Remarkably, the pronounced impact of the COVID-19 pandemic in 2020 drives the total prediction errors shown in the second column of Table 3. The reassessment of this column, depicted in the sixth column, relies on trimming the two highest absolute prediction errors of 2020 and provides robustness to the findings reported in the previous paragraph.

5.4. Discussion

Studies conducted over different countries' data and timeframes indicate the potential of GT data to enhance the accuracy of unemployment figures predictions, as also demonstrated in this study. By the use of mixed-frequency modelling approaches, weekly GT data has proven effective in predicting monthly unemployment counts in the UK, yielding via MIDAS regressions substantial accuracy gains ranging between 6%–11% in RMSFE compared to naive forecasts [6]. Similarly, in the USA, in predicting weekly unemployment claims, a combination of U-MIDAS and machine learning for selecting the main features of a high-dimensional set of keywords on daily GT data enhanced nowcasts, resulting in a performance boost of 25%–45% in RMSFE compared to an AR benchmark model, with the extent of improvement depending on the specific nowcast week [27]. The predictive performance observed in our current study based on the standard MIDAS approach utilising daily GT predictors outperformed benchmarks significantly by at least 42% in RMSFE. This result aligns with findings from the American context [27], which has a distinct approach to obtaining extensive daily GT series that overlook the coherency with monthly and weekly GT

Table 3
Accuracy of nowcasts produced by daily and monthly GT predictors and of INE provisional estimates.

Prediction model	RMSFE by prediction period				Trimmed total RMSFE[d]
	Total[a]	2019	2020[b]	2021[c]	
(A) MIDAS (based on daily predictors)	0.2659	**0.1205**	0.4286	**0.1374**	**0.1405**
(B) ARMAX (based on monthly predictors)	0.5190	0.2144	0.8311	0.3028	0.3442
(E) INE provisional estimates	**0.2249**	0.2160	**0.2747**	0.1732	0.1862
Comparison (A) MIDAS vs (E) INE provisional	18%	−44%	56%	−21%*	−25%
mDM *p*-value	(0.3231)	(0.2329)	(0.1805)	(0.0857)	(0.1920)
Comparison (B) ARMAX vs (E) INE provisional	131%*	−1%	203%*	75%**	85%*
mDM *p*-value	(0.0965)	(0.4946)	(0.0809)	(0.0246)	(0.0610)

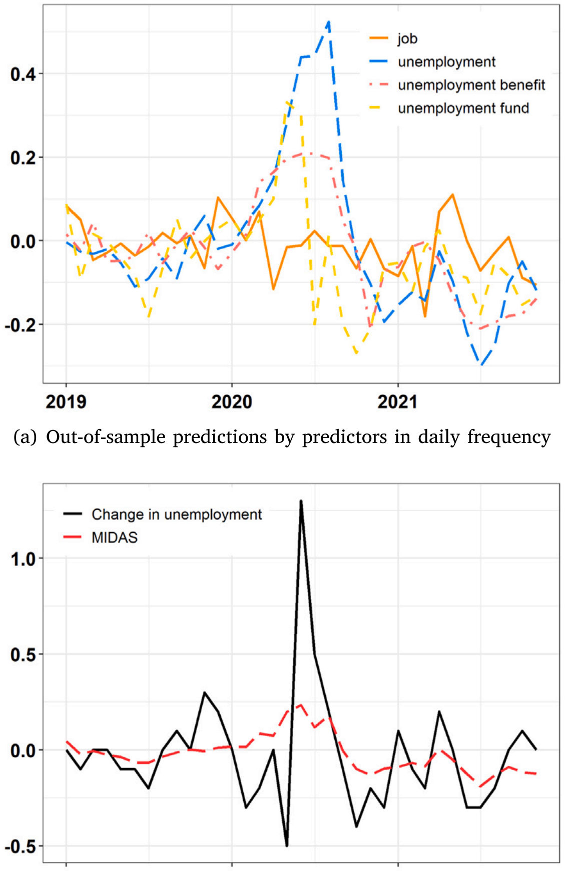
mDM, one-sided modified Diebold–Mariano test: $^{**}p < 0.05$, $^{*}p < 0.10$.

[a] January 2019 to November 2021, except March 2020, the reference month INE did not issue a provisional estimate due to the pandemic.

[b] Except March 2020.

[c] January 2021 to November 2021.

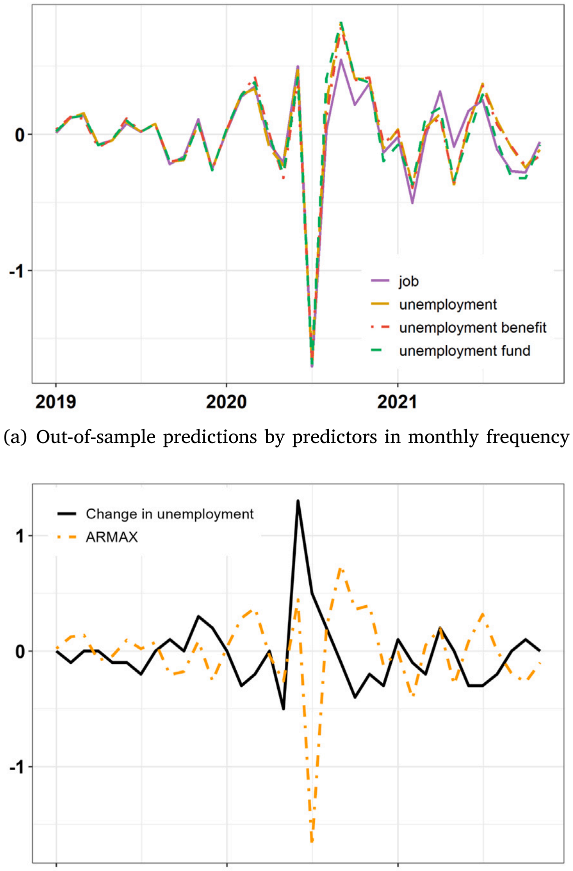
[d] Recalculated total by trimming the two highest absolute errors in 2020.



(a) Out-of-sample predictions by predictors in daily frequency

(b) Comparison of out-of-sample nowcasts to the official unemployment rate

Fig. 6. Predictions based on GT daily series by keywords and nowcasts compared to the official unemployment rate.



(a) Out-of-sample predictions by predictors in monthly frequency

(b) Comparison of out-of-sample nowcasts to the official unemployment rate

Fig. 7. Predictions based on GT monthly series by keywords and nowcasts compared to the official unemployment rate.

series considered in this study. In-depth studies are required to comprehensively compare methods and findings related to constructing an extensive daily GT series as employed in this study based on Eichenauer et al. [31] and Borup et al. [27].

On the other hand, traditional same-frequency modelling with monthly GT data has demonstrated effectiveness, showcasing improved accuracy in predicting job demands in Spain and outperforming ARIMA benchmarks by 13% in RMSFE [8]. Additionally, it has successfully forecasted the US national unemployment rate, consistently achieving 18%–32% improvement in RMSFE across various forecasting horizons [11]. For the COVID-19 pandemic period, previous studies using same-frequency modelling highlighted the efficacy of GT data in enhancing forecasts, especially for weekly US initial unemployment benefit claims. These studies concluded that GT-augmented predictions outperformed 37% in RMSFE (between March and July of 2020) over random walk (RW) forecasts [21] and at least 40% in mean-absolute forecast error from February to October of 2020 over RW or AR forecasts [22]. The results of our study with same-frequency modelling during the first full year of the COVID-19 pandemic indicate a modest but significant 2% superior accuracy in RMSFE compared to the ARMA benchmark and a meaningful 16% underperformance compared to naive forecasts from the RW. Conversely, the mixed-frequency approach in our current study demonstrates daily GT data consistently outperforming benchmarks by at least 40%. It is noteworthy that mixed-frequency models may produce comparable prediction performance over an entire forecasting timeframe but can diverge across distinct subperiods based on the economic context, with such variations being particularly more pronounced in extraordinary situations like the COVID-19 pandemic [50,51].

Among unemployment-GT forecasting studies for Portugal, our study distinguishes by showcasing GT's responsiveness to the impact of COVID-19, utilising daily predictors and integrating series publishing timelines. Unlike previous studies with Portuguese unemployment data that rely exclusively on GT monthly predictors building 1-step-ahead predictions, our approach introduces predictions with a 60-day delay in the publication schedule, leading to 2-step-ahead predictions. Previous studies consider solely AR-based predictions as benchmarks, excluding the consideration of an RW, report predictions augmented by monthly GT data as more accurate than benchmarks by 10%–16% in RMSFE between September 2010 and August 2013 [13] and by 1% in RMSFE from January 2021 to June 2021 [14]. Meanwhile, ours observes a 2% enhancement against the ARMA benchmark from January 2019 to November 2021. Moreover, complementing previous studies, this research extends the benchmark to RW predictions and argues that monthly GT-augmented predictions are on par with such a benchmark for the entire forecasting period, 2019 and 2021, and fall short in 2020. Furthermore, our findings reveal daily GT data standing out in unemployment predictions, surpassing monthly GT predictors and benchmarks in accuracy, even during the challenging initial COVID-19 period.

6. Conclusion

The conventional approach in macroeconomic forecasting relies heavily on data from National Statistics Offices. However, worldwide events like the COVID-19 outbreak and the 2007–2008 financial crisis have hindered the issuance of timely estimates [52,53], fostering a growing interest in non-traditional data sources as predictive tools. Google Trends (GT) stands out among such sources due to its promptness, cost-effectiveness and diverse sampling frequencies. A thorough examination of the role of this data source employing high-frequency sampling predictors reveals its potential to yield reliable economic indicator predictions, empowering policymakers to assess current economic states and make informed decisions.

This study corroborates the evidence of GT's predictive power for unemployment within the Portuguese labour market. However, selecting GT data involves navigating a myriad of potential keywords or categories, making it a complex endeavour. According to this, exploring other job-related terms beyond the ones used in this study could improve the predictive performance of GT-based nowcasts. The conceptualisation of this study took into account the trade-off between relying on a single keyword [e.g., 13–15] or multiple keywords [e.g., 6,7,25] to generate predictions. The decision to adopt a combination of four culturally relevant job-related keywords in Portugal was driven by the recognition that depending on a single keyword may risk losing the relationship with the dependent variable if that term becomes less relevant over time [6]. Conversely, employing multiple keywords may address the concern associated with a single keyword and facilitate the convergence toward desired outcomes by capturing hidden trends [10].

Comparing the accuracy of daily GT predictors against monthly counterparts, this study offers a distinguished perspective from previous unemployment-GT literature. Moreover, it introduces innovative approaches concerning Portugal's data, including considerations for the exact timing of data availability, the utilisation of daily GT series as predictors, comparisons with official provisional estimates and an analysis of the COVID-19 pandemic's impact. Other perspectives could involve the prediction of a country's disaggregated unemployment using GT data as predictors, such as regionalised predictions [e.g., 11,23] and demographic predictions, providing valuable insights into the digital divide and delving deeper into the selection bias from search engine users discussed in Dilmaghani [24], and Mulero and Garcia-Hiernaux [25]. In this context, future work could address GT data as predictors for developing economies.

This study predicts Portuguese monthly unemployment by employing mixed-frequency and same-frequency approaches, i.e., considering daily GT predictors via a MIDAS model and monthly GT predictors via an ARMAX model. This prediction strategy relies on econometric methods, disregarding alternatives like machine learning-based approaches. The choice between MIDAS models and machine learning methods depends on the data's specific characteristics, the problem's nature and the analysis's goals. This paper's prediction aims to guide policymakers in assessing the current state of an economy based exclusively on *ex-ante* information. The option for the standard MIDAS regression relies on the large frequency ratio between the monthly predicted series and daily GT series, given that this modelling parsimoniously reduces the number of estimated parameters. In addition, extending the framework employed in this study could facilitate econometric inference, including hypothesis testing, estimating confidence intervals and imposing economic constraints or assumptions into the model. Compared to most machine learning approaches, these tasks are typically more amenable to accommodated by econometric methods. Notwithstanding, the authors of this paper recognise that machine learning methods are a valuable tool for producing unemployment predictions based on GT predictors, as seen in Xu et al. [34], Li et al. [54], Borup and Schütte [3], Jeremias [55].

Among the results found in this paper, nowcasts based on high-frequency daily GT predictors demonstrate exceptional performance, outperforming benchmarks and predictions augmented by monthly GT predictors. This result underscores the effectiveness of daily GT predictors in addressing the challenges posed by the global health issue of COVID-19. In contrast, nowcasts based on monthly GT predictors exhibit better accuracy performance than the ARMA as a benchmark but fall short in accuracy compared to predictions from the RW benchmark and daily-based data, highlighting the superior performance of daily GT predictors.

This study emphasises that resorting to daily GT predictors in the Portuguese context yields accurate and timely information on the unemployment rate, delivering proper estimates 60 days and 30 days in advance than official and provisional estimates issued by the INE, respectively.

The framework illustrated in this study, tailored to the Portuguese case due to the necessity of selecting culture-dependent keywords for extracting GT data, offers a blueprint for collecting relevant GT data when extending similar analyses to other countries or regions.

CRediT authorship contribution statement

Eduardo André Costa: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Maria Eduarda Silva:** Investigation, Methodology, Supervision, Writing – review & editing. **Ana Beatriz Galvão:** Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors gratefully acknowledge support from CEF.UP. Portuguese public funds have financed this research via FCT (Fundação para a Ciência e a Tecnologia) and ESF (European Social Fund) under the reference 2021.07583.BD, and in the framework of the FCT project with reference LA/P/0063/2020, DOI 10.54499/LA/P/0063/2020.

Appendix A. Portuguese unemployment rate definition

The International Labor Organization guidelines conduct the Portuguese labour market survey to gauge the local unemployment rate. The survey focuses on people aged between 15 and 74. It classifies the surveyees into *employed*, *unemployed*, and *economically inactive*. The *employed* classification refers to employed people, self-employed or family workers who either worked for profit during the interview reference period or were temporarily absent from their jobs or businesses during the interview period of reference. The *unemployed* categorisation includes individuals not classified as employed who are available for work, sought a job within the interview reference period or three weeks before, or found a position to start up to three months after the unemployment measurement reference period. The *economically inactive* group comprises individuals who are neither classified as employed nor unemployed. The employed and unemployed populations constitute a country's labour force, also known as the active population, for producing goods and services. Accordingly, the unemployment rate is the ratio of people classified as unemployed to the active population.

Appendix B. Timetable of provisional and official 2019 Portuguese unemployment rate

Calendar date	Month of reference	
	Provisional	Official
28-Feb-19	January 2019	December 2018
29-Mar-19	February 2019	January 2019
29-Apr-19	March 2019	February 2019
03-Jun-19	April 2019	March 2019
28-Jun-19	May 2019	April 2019
30-Jul-19	June 2019	May 2019
29-Aug-19	July 2019	June 2019
27-Sep-19	August 2019	July 2019
30-Oct-19	September 2019	August 2019
28-Nov-19	October 2019	September 2019
08-Jan-20	November 2019	October 2019
29-Jan-20	December 2019	November 2019
28-Feb-20	January 2020	December 2019

Appendix C. Summary of overlapping windows to build an extensive keyword GT daily series

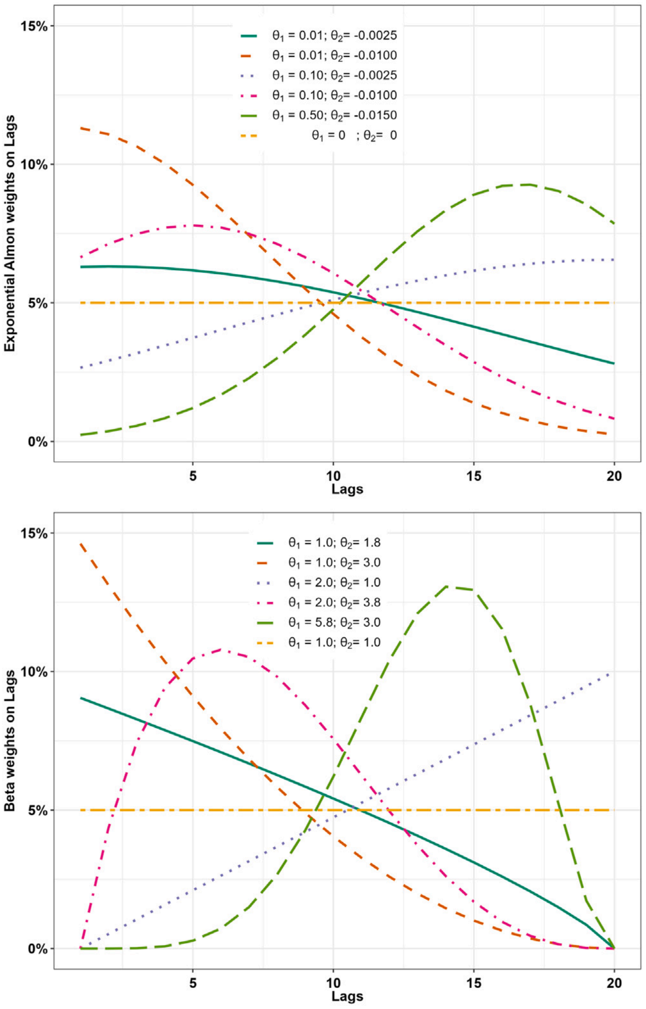
See Table C.4.

Table C.4
Summary of overlapping windows to build an extensive keyword GT daily series.

Frequency	Window[a]	Shift[a]
Daily	6 months	15 days
Weekly	5 years	11 weeks
Monthly	15 years	1 month

[a] Windows and shifts proposed in Eichenauer et al. [31].

Appendix D. Experimental designs of the exponential Almon and Beta functions as distributed polynomials



Appendix E. Data available for forecasting purposes integrating the timing of the unemployment rate publication at prediction points

See Table E.5.

Appendix F. RMSFE comparison between MIDAS nowcasting performance using exponential Almon polynomial and Beta lag into weighting functions

See Table F.6.

Table E.5
Data available for forecasting purposes integrating the timing of the unemployment rate publication at prediction points.

Prediction point	Date	Prediction target	Last available data reference			
			Google series		Unemployment rate	
			Daily	Monthly	Official	Provisional
1	31-Jan-19	January 2019	31-Jan-19	January 2019	November 2018	December 2018
2	28-Feb-19	February 2019	28-Feb-19	February 2019	December 2018	January 2019
3	31-Mar-19	March 2019	31-Mar-19	March 2019	January 2019	February 2019
⋮	⋮	⋮	⋮	⋮	⋮	⋮
33	30-Sep-21	September 2021	30-Sep-21	September 2021	July 2021	August 2021
34	31-Oct-21	October 2021	31-Oct-21	October 2021	August 2021	September 2021
35	30-Nov-21	November 2021	30-Nov-21	November 2021	September 2021	October 2021

Table F.6
RMSFE comparison between MIDAS nowcasting performance using exponential Almon polynomial and Beta lag into weighting functions.

Search terms	Weighting function	RMSFE by prediction period			
		Total[a]	2019	2020	2021[b]
unemployment	Exponential Almon polynomial	0.2532	0.1171	0.3974	0.1294
	Beta	0.2656	0.1234	0.4087	0.1602
	Comparison Almon vs Beta	−5%	−5%	−3%	−19%**
	mDM p-value	(0.1704)	(0.3195)	(0.3324)	(0.0327)
unemployment fund	Exponential Almon polynomial	0.2864	0.1081	0.4532	0.1556
	Beta	0.2932	0.1355	0.4580	0.1571
	Comparison Almon vs Beta	−2%	−20%*	−1%	−1%
	mDM p-value	(0.4220)	(0.0781)	(0.4711)	(0.4823)
unemployment benefit	Exponential Almon polynomial	0.2815	0.1452	0.4305	0.1644
	Beta	0.2947	0.1472	0.4573	0.1565
	Comparison Almon vs Beta	−5%	−1%	−6%	5%
	mDM p-value	(0.1668)	(0.3894)	(0.1464)	(0.1906)
job	Exponential Almon polynomial	0.2998	0.1397	0.4675	0.1622
	Beta	0.2908	0.1312	0.4514	0.1671
	Comparison Almon vs Beta	3%	7%	4%	−3%
	mDM p-value	(0.1370)	(0.1594)	(0.1460)	(0.2026)
Daily combined search terms	Exponential Almon polynomial	0.2665	0.1205	0.4186	0.1374
	Beta	0.2746	0.1277	0.4315	0.1376
	Comparison Almon vs Beta	−3%	−6%	−3%	0%
	mDM p-value	(0.1452)	(0.1222)	(0.1840)	(0.4922)

mDM, one-sided modified Diebold–Mariano test: $^{**}p < 0.05$, $^{*}p < 0.10$.

[a] January 2019 to November 2021.

[b] January 2021 to November 2021.

References

[1] Buono D, Mazzi GL, Kapetanios G, Marcellino M, Papailias F. Big data types for macroeconomic nowcasting. Eurostat Rev Natl Acc Macroecon Indic 2017;1(2017):93–145, URL: https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2017-art4.pdf.

[2] Barcellan R, Nielsen PB, Calsamiglia C, Camerer C, Cantillon E, Crépon B, et al. Developments in data for economic research. In: Matyas L, Blundell R, Cantillon E, Chizzolini B, Ivaldi M, Leininger W, Marimon R, Steen F, editors. Economics without borders: economic research for European policy challenges. Cambridge: Cambridge University Press; 2017, p. 568–611. http://dx.doi.org/10.1017/9781316636404.015.

[3] Borup D, Schütte ECM. In search of a job: Forecasting employment growth using Google Trends. J Bus Econom Statist 2020;40(1):186–200. http://dx.doi.org/10.1080/07350015.2020.1791133.

[4] Ettredge M, Gerdes J, Karuga G. Using web-based search data to predict macroeconomic statistics. Commun ACM 2005;48(11):87–92. http://dx.doi.org/10.1145/1096000.1096010.

[5] McLaren N, Shanbhogue R. Using internet search data as economic indicators. Bank Engl Q Bull 2011;51(2):134–40. http://dx.doi.org/10.2139/ssrn.1865276.

[6] Smith P. Google's MIDAS touch: Predicting UK unemployment with internet search data. J Forecast 2016;35(3):263–84. http://dx.doi.org/10.1002/for.2391.

[7] Niesert RF, Oorschot JA, Veldhuisen CP, Brons K, Lange RJ. Can Google search data help predict macroeconomic series? Int J Forecast 2020;36(3):1163–72. http://dx.doi.org/10.1016/j.ijforecast.2018.12.006.

[8] Vicente MR, López-Menéndez AJ, Pérez R. Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? Technol Forecast Soc Change 2015;92:132–9. http://dx.doi.org/10.1016/j.techfore.2014.12.005.

[9] González-Fernández M, González-Velasco C. Can Google econometrics predict unemployment? Evidence from Spain. Econom Lett 2018;170:42–5. http://dx.doi.org/10.1016/j.econlet.2018.05.031.

[10] Mulero R, García-Hiernaux A. Forecasting Spanish unemployment with Google Trends and dimension reduction technique. SERIEs 2021;12:329–49. http://dx.doi.org/10.1007/s13209-021-00231-x.

[11] D'Amuri F, Marcucci J. The predictive power of Google searches in forecasting US unemployment. Int J Forecast 2017;33(4):801–16. http://dx.doi.org/10.1016/j.ijforecast.2017.03.004.

[12] Maas B. Short-term forecasting of the US unemployment rate. J Forecast 2020;39(3):394–411. http://dx.doi.org/10.1002/for.2630.

[13] Barreira N, Godinho P, Melo P. Nowcasting unemployment rate and new car sales in South-Western Europe with Google Trends. NETNOMICS: Econ Res Electron Netw 2013;14(3):129–65. http://dx.doi.org/10.1007/s11066-013-9082-8.

[14] Simionescu M, Cifuentes-Faura J. Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. J Policy Model 2022;44(1):1–21. http://dx.doi.org/10.1016/j.jpolmod.2021.09.011.

[15] Fondeur Y, Karamé F. Can Google data help predict French youth unemployment? Econ Model 2013;30(1):117–25. http://dx.doi.org/10.1016/j.econmod.2012.07.017.

[16] Naccarato A, Falorsi S, Loriga S, Pierini A. Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technol Forecast Soc Change 2018;130:114–22. http://dx.doi.org/10.1016/j.techfore.2017.11.022.

[17] Chadwick MG, Sengül G. Nowcasting the unemployment rate in Turkey: Let's ask Google. Working Papers, Research and Monetary Policy Department, Central Bank of the Republic of Turkey; 2012, URL: https://EconPapers.repec.org/RePEc:tcb:wpaper:1218.

[18] Simionescu M. Improving unemployment rate forecasts at regional level in Romania using Google Trends. Technol Forecast Soc Change 2020;155:120026. http://dx.doi.org/10.1016/j.techfore.2020.120026.

[19] Pavlicek J, Kristoufek L. Nowcasting unemployment rates with Google searches: Evidence from the visegrad group countries. PLoS One 2015;10(5):1–11. http://dx.doi.org/10.1371/journal.pone.0127084.

[20] Choi H, Varian H. Predicting the present with Google Trends. Econ Rec 2012;88(s1):2–9. http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x.

[21] Yi D, Ning S, Chang C-J, Kou SC. Forecasting unemployment using internet search data via PRISM. J Amer Statist Assoc 2021;116(536):1662–73. http://dx.doi.org/10.1080/01621459.2021.1883436.

[22] Aaronson D, Brave SA, Butters RA, Fogarty M, Sacks DW, Seo B. Forecasting unemployment insurance claims in realtime with Google Trends. Int J Forecast 2022;38(2):567–81. http://dx.doi.org/10.1016/j.ijforecast.2021.04.001.

[23] Simionescu M, Cifuentes-Faura J. Forecasting national and regional youth unemployment in Spain using Google Trends. Soc Indic Res 2022;164(3):1187–216. http://dx.doi.org/10.1007/s11205-022-02984-9.

[24] Dilmaghani M. The racial 'digital divide' in the predictive power of Google Trends data for forecasting the unemployment rate. J Econ Soc Meas 2018;43(3–4):119–42. http://dx.doi.org/10.3233/JEM-180458.

[25] Mulero R, Garcia-Hiernaux A. Forecasting unemployment with Google Trends: Age, gender and digital divide. Empir Econ 2023;65:587–605. http://dx.doi.org/10.1007/s00181-022-02347-w.

[26] Larson WD, Sinclair TM. Nowcasting unemployment insurance claims in the time of COVID-19. Int J Forecast 2022;38(2):635–47. http://dx.doi.org/10.1016/j.ijforecast.2021.01.001.

[27] Borup D, Rapach DE, Schütte ECM. Mixed-frequency machine learning: Nowcasting and backcasting weekly initial claims with daily internet search volume data. Int J Forecast 2023;39(3):1122–44. http://dx.doi.org/10.1016/j.ijforecast.2022.05.005.

[28] Eurostat. 2022, Online URL: https://ec.europa.eu/eurostat/data/database (Accessed 31 January 2022).

[29] Statistics Portugal. Monthly estimates of employment and unemployment. 2022, Online URL: https://www.ine.pt/xportal/xmain?PORTLET_ID=JSP&xpgid=ine_destaques&xpid=INE&PORTLET_NAME=ine_cont_header_dest_en&PORTLET_UID=%23JSP%3Aine_cont_header_dest_en%23&DESTAQUESdata_inicial=&DESTAQUESdata_final=&x=15&y=12&DESTAQUESfreeText=Monthly+Employment+and+Unemployment+Estimates (Accessed 31 January 2022).

[30] Statistics Portugal. Monthly estimates of employment and unemployment - august 2020. 2020, Online URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=415271578&DESTAQUESmodo=2 (Accessed 25 January 2024).

[31] Eichenauer VZ, Indergand R, Martínez IZ, Sax C. Obtaining consistent time series from Google Trends. Econ Inq 2022;60(2):694–705. http://dx.doi.org/10.1111/ecin.13049.

[32] Chow GC, Lin A-l. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. Rev Econ Stat 1971;53(4):372–5, URL: http://www.jstor.org/stable/1928739.

[33] Nagao S, Takeda F, Tanaka R. Nowcasting of the U.S. unemployment rate using Google Trends. Finance Res Lett 2019;30:103–9. http://dx.doi.org/10.1016/j.frl.2019.04.005.

[34] Xu W, Li Z, Cheng C, Zheng T. Data mining for unemployment rate prediction using search engine query data. Serv Orient Comput Appl 2013;7(1):33–42. http://dx.doi.org/10.1007/s11761-012-0122-2.

[35] Askitas N, Zimmermann KF. Google econometrics and unemployment forecasting. Appl Econ Q 2009;55(2):107–20. http://dx.doi.org/10.3790/aeq.55.2.107.

[36] Dilmaghani M. Workopolis or The Pirate Bay: What does Google Trends say about the unemployment rate? J Econ Stud 2019;46(2):422–45. http://dx.doi.org/10.1108/JES-11-2017-0346.

[37] Taylor S, Letham B. Prophet: Automatic forecasting procedure. 2021, URL: https://CRAN.R-project.org/package=prophet R package version 1.0.

[38] Ghysels E, Santa-Clara P, Valkanov R. The MIDAS Touch: Mixed Data Sampling Regression Models. Working paper, UNC and UCLA; 2004.

[39] Ghysels E, Sinko A, Valkanov R. MIDAS regressions: Further results and new directions. Econometric Rev 2007;26(1):53–90. http://dx.doi.org/10.1080/07474930600972467.

[40] Ghysels E, Marcellino M. Applied Economic Forecasting Using Time Series Methods. Oxford University Press; 2018.

[41] Andreou E, Ghysels E, Kourtellos A. Should macroeconomic forecasters use daily financial data and how? J Bus Econom Statist 2013;31(2):240–51. http://dx.doi.org/10.1080/07350015.2013.767199.

[42] Foroni C, Marcellino M, Schumacher C. Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. J Roy Statist Soc Ser A: Statist Soc 2013;178(1):57–82. http://dx.doi.org/10.1111/rssa.12043.

[43] Bonino-Gayoso N, Garcia-Hiernaux A. TF-MIDAS: A transfer function based mixed-frequency model. J Stat Comput Simul 2021;91(10):1980–2017. http://dx.doi.org/10.1080/00949655.2021.1879082.

[44] Ghysels E, Kvedaras V, Zemlys-Balevičius V. Chapter 4 - mixed data sampling (MIDAS) regression models. In: Vinod HD, Rao C, editors. Financial, macro and micro econometrics using r. Handbook of statistics, vol. 42, Elsevier; 2020, p. 117–53. http://dx.doi.org/10.1016/bs.host.2019.01.005.

[45] Said SE, Dickey DA. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 1984;71(3):599–607. http://dx.doi.org/10.1093/biomet/71.3.599.

[46] Hyndman RJ, Khandakar Y. Automatic time series forecasting: The forecast package for R. J Stat Softw 2008;27(3):1–22. http://dx.doi.org/10.18637/jss.v027.i03.

[47] Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econom Statist 1995;13(3):253–63. http://dx.doi.org/10.1080/07350015.1995.10524599.

[48] Harvey D, Leybourne S, Newbold P. Testing the equality of prediction mean squared errors. Int J Forecast 1997;13(2):281–91. http://dx.doi.org/10.1016/S0169-2070(96)00719-4.

[49] Mariano RS, Preve D. Statistical tests for multiple forecast comparison. J Econometrics 2012;169(1):123–30. http://dx.doi.org/10.1016/j.jeconom.2012.01.014.

[50] Bonino-Gayoso N. Mixed-frequency models: Alternative comparison and macroeconomic variables nowcasting (Ph.D. thesis), Universidad Complutense de Madrid; 2022, Available at https://docta.ucm.es/entities/publication/4379324f-9ddb-4578-99b3-a971b8fd47ee.

[51] Bonino-Gayoso N, Garcia-Hiernaux A. Macroeconomic forecasting evaluation of MIDAS models. In: Valenzuela O, Rojas F, Herrera LJ, Pomares H, Rojas I, editors. Theory and applications of time series analysis. Cham: Springer Nature Switzerland; 2023, p. 135–53.

[52] Simionescu M, Zimmermann KF. Big Data and unemployment analysis. GLO Discussion Paper 81, Maastricht: Global Labor Organization (GLO); 2017, URL: http://hdl.handle.net/10419/162198.

[53] United Nations and World Bank. Monitoring the state of statistical operations under the COVID-19 pandemic: Highlights from the second round of a global COVID-19 survey of national statistical offices. Washington, D.C.: World Bank Group; 2020, URL: http://documents.worldbank.org/curated/en/297221597442670485/Monitoring-the-State-of-Statistical-Operations-under-the-COVID-19-Pandemic-Highlights-from-the-Second-Round-of-a-Global-COVID-19-Survey-of-National-Statistical-Offices.

[54] Li Z, Xu W, Zhang L, Lau RY. An ontology-based web mining method for unemployment rate prediction. Decis Support Syst 2014;66:114–22. http://dx.doi.org/10.1016/j.dss.2014.06.007.

[55] Jeremias G. Forecasting with Machine Learning methods: A case study with unemployment. Universidade do Porto; 2023, Available at https://hdl.handle.net/10216/156518.

Eduardo André Costa is a Ph.D. Candidate in Economics at the School of Economics and Management of the University of Porto. He holds an undergraduate degree in Statistics and a master's in Industrial Engineering. His research interests include linear and non-linear time series, forecasting, quantitative methods and non-traditional data sources.

Maria Eduarda Silva is an Associate Professor with habilitation at the University of Porto School of Economics and Management, where she is also the Director of the Master in Data Analytics, and former Secretary General of the Federation of the National Statistical Societies. Her main research interests are non-linear time series, time series of counts, long-memory, time series classification/clustering and economic applications.

Ana Beatriz Galvão is an Economic Modelling and Forecasting Professor at Warwick Business School, University of Warwick. She is a Research Fellow of the Centre for Economic Policy Research and a fellow of the International Association for Applied Econometrics. Ana Galvão's research interests are in empirical macroeconomics, forecasting and non-linear time series models.