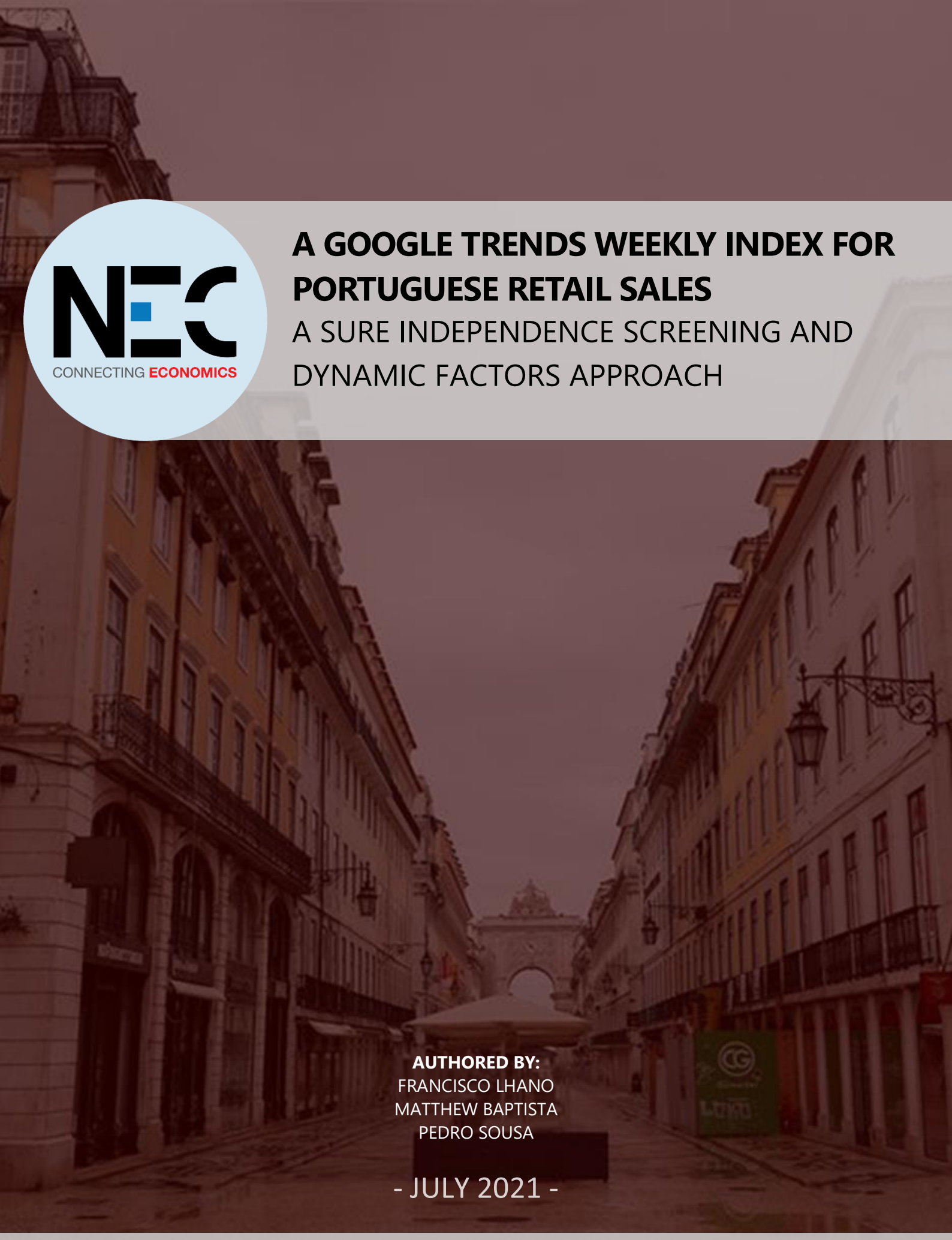


A GOOGLE TRENDS WEEKLY INDEX FOR PORTUGUESE RETAIL SALES

A SURE INDEPENDENCE SCREENING AND DYNAMIC FACTORS APPROACH

AUTHORED BY:
FRANCISCO LHANO
MATTHEW BAPTISTA
PEDRO SOUSA

- JULY 2021 -

In partnership with:

Gabinete de Estratégia e Estudos
Ministério da Economia

Nova Economics Club is the Economics club at Nova School of Business and Economics. Since 2012, its focus is to create a bridge between the theory learnt in classes with the work performed by economists on a daily basis.

e-mail: novaeconomicsclub@novasbe.pt
website: www.novaeconomicsclub.pt

Title: *A Google Trends Weekly Index for Portuguese Retail Sales: the Sure Independence Screening and Dynamic Factors Approach*
Authors: *Francisco Lhano, Matthew Baptista, and Pedro Sousa*
Partner: *GEE – Gabinete de Estratégia e Estudos do Ministério da Economia*
Advisor: *Gabriel Osório de Barros, Nuno Tavares, and Rita Bessone Basto*
Date: *July 2021*

Cover photo by Rafael Marchante/Reuters from Rádio Renascença

Acknowledgements

All the opinions shared in this report are solely those of the authors and do not necessarily coincide with those of the Office for Strategy and Studies of the Portuguese Ministry of Economy (GEE) or the Nova Economics Club. Any errors and omissions are the responsibility of the authors. The authors would like to thank all individuals who helped build the dataset used in this report. Furthermore, we would like to thank GEE in general and Gabriel Osório de Barros, Nuno Tavares, and Rita Bessone Basto in particular for their guidance and academic stimulation to get a grip of this broad and challenging scientific field.

A GOOGLE TRENDS WEEKLY INDEX FOR PORTUGUESE RETAIL SALES

THE SURE INDEPENDENCE SCREENING AND DYNAMIC FACTORS APPROACH

Francisco Lhano

Nova Economics Club

Nova School of Business and Economics

lhano.francisco@gmail.com

Matthew Baptista

Nova Economics Club

Nova School of Business and Economics

baptista.matthew@yahoo.fr

Pedro Sousa

Nova Economics Club

Nova School of Business and Economics

pnunosousa@gmail.com

Abstract

We develop a composite indicator based on high-frequency data to answer the need for timely information in a period of increased short-term volatility and uncertainty in the economy. We target retail sales. The scarcity of publicly available high-frequency data is solved by using Portuguese Google searches. They are treated as per Nicolas Woloszko's comprehensive methodology. In this big-data context, we implement a Sure Independence Screening procedure to parsimoniously select the most relevant keywords. The selected subsample is then included in a dynamic factor model, to extract the signal that will be our index, in a fashion similar to the FED's Weekly Economic Index and the BoP's Daily Economic Index. The indicator behaves decently and allows us to re-tell recent history. Quarterly and MIDAS regressions show good in-sample performance. As often found in the Google Trends nowcasting literature, our indicator is particularly well-suited for detecting big economic downfalls but has a mixed performance in normal times.

Table of Contents

1) Non-Technical Summary

In recent years there has been much research focusing on finding ways to balance the trustworthiness of institutional quarterly economic reports (e.g., resorting to the Central Banks' reports and respective standard indicators), and the updating speed of information on the news, which naturally translates in faster reactions from the general public to new economic conjunctures or public policies.

We propose to build a composite indicator that monitors economic activity. The need to build this indicator arises from the fact that official macroeconomic data is reported with low periodicity compared to its urgent need and with significant time-lags. Given these conditions, it makes it extremely hard for policymakers and economists to assess not only exogenous and unconventional shocks (such as the Covid-19 pandemic), but also the numerous economic policies put through by the relevant institutions. Hence, we develop an alternative indicator that relies on weekly data to act as a proxy of said macroeconomic variables. We intend to proxy them by finding significant correlations between the data and those macroeconomic variables, such that we manage to timely monitor the evolution of the Portuguese economy, either in the pandemic context or many others. After such inspection, factor analysis will help us aggregate the information contained in those many series, summing it up in one single variable, our indicator.

Bearing in mind the highly demanding task of addressing one of the oldest problems in the field of Economics (accurate predictions in a timely fashion), a new field of forecasting is explored, one that resorts to information extracted from a feature of Google – "Google Trends". We gather Portuguese search results whose frequency reflects either the retail sales variation or economic growth during a given period. To this effect, a keywords dataset is created. The trends are extracted and downloaded using a Google Trends' *Python* library called "*Pytrends*" and then the relevant variables are picked by machine-learning (SIS – Sure Independence Screening procedure) using the *R* software. Subsequently, the data is included in a dynamic factor model to capture the relationships among these variables' time series.

2) Introduction

In this project, we try to answer some of the usual problems of timeliness and noisiness in the information available to policymakers. We focus on the retail sales series, a measure of the economic conjuncture that was particularly affected by the turmoil of the Covid-19 pandemic. The recent economic events were characterized by drastic and overnight policy decisions, making it necessary to follow economic developments at a horizon shorter than the month.

Throughout the last couple of years, the nowcasting literature has veered towards the inclusion of higher frequency data in its models. When mobilising this data for our problem, we faced a supplementary obstacle: the limited public availability of classical series (e.g., daily credit card payments' data), and the rather short time span of the publicly available series in Portugal.

Although the literature on Google Trends for nowcasting is as old as the tool itself (2004), and even though some of it has proven to be inconclusive, it has known a certain resurgence in the recent years: advances in specific data treatments and selection methods has made it possible to broaden perspectives to those eager to use such data source.

The solution we chose was thus to build an index like the FED's[1], but on the basis of Google trends information, whose data are publicly available, exist since 2004, and are updated weekly. However, the solution presented by Google Trends data opens new problems to the forecaster. These data were not meant for nowcasting, and their coverage is immense: a danger would be indiscriminate inclusion, leading to a noisy index.

Before building the indicator *per se*, we follow three notable steps, aimed at targeting the objective-series and reducing the noisiness of our own series. First, we include trends at the keyword level, that is, we personalize the set with our knowledge of the Portuguese economy's specificities. For this specific task, we shared a survey with Portuguese residents in order to further diversify the dataset. Our dataset contains keywords that are Google search terms people would essentially type to satisfy their needs. Second, we treat those trends with Woloszko's methods (the most advanced to our knowledge). Ferrara (2019) had successfully used the SIS variable selection for nowcasting with Google Trends. The third step is to use it.

[1] The Weekly Economic Index (WEI).

Our application of SIS is slightly different, as we aim at constituting an index, before using it for nowcasting. In short, we solve the problem of nowcasting data needs by resorting to Google data. We then solve the problem of Google Trends being Big Data, and thus noisy, by applying feature selection (SIS) and then feature extraction (DFM) processes on the retained features to build our index.

Our results are encouraging and show that Google trends do contain relevant information, which can solve the peculiar problems of nowcasting. In particular we find that our index follows downfalls properly, both in timing and amplitude. It does however produce false high growth signals, and still presents several improvement perspectives.

In section 3, we present the review of the literature we used. In section 4, we show how the dataset was built and treated. Section 5 describes the methods used to select the relevant features of our dataset and extract the signal from them. In section 6, we present estimation results, and in section 7, we evaluate the in-sample forecasting performance, before concluding.

3) Literature Review

Our work draws on two strands of the nowcasting literature. First, we look at the recent work on high-frequency indicators of economic activity. Second, we see how Google data have been used for that purpose and how it performed.

The History of tracking economic activity in real-time dates to the last century, when some statisticians and econometricians resorted to weekly data to support policy analysis. Some of them focused on commodities' price quotations, whereas others focused on bank debts. Nowadays, there is a lot more availability [and accessibility] to get data. There were substantial breakthroughs in the high-frequency data field in the year 2020, due to the urgency of tracking down economic activity amid the Covid-19 pandemic. The Federal Reserve Bank of New York, Bundesbank and Bank of Portugal were three of the institutions which understood the need to put forth high-frequency economic indicators like the ones we have been talking about (WEI, WAI and DEI, respectively).

(Lewis, Mertens, Stock, & Trivedi, 2020), FED authors, proposed a Weekly Economic Index (WEI). This indicator of real economic activity offers its users two services. Its primary purpose is to provide a weekly index of real economic activity, whose output is the overall change in macroeconomic activity during the reference week, relative to the corresponding week one year earlier. The second purpose of the WEI is to nowcast monthly and quarterly economic activity series. Our indicator shares the same purposes. The authors include series of consumption, labour input and production in the index, excluding series that underestimate the economic activity, such as air traffic, which are hit particularly hard by pandemic-related policies. These series are chosen thanks to their theoretical relevance to economic activity and based on the correlation of their quarterly variant with the GDP growth rate. They find that the series have a narrative value and track events properly. It is also important to point out that, in MIDAS[2] nowcasting exercises, the WEI has a great forecasting power for GDP growth.

The Bundesbank's WAI (Eraslan & Götz, 2020) is also a principal component over high-frequency indicator. Its originality is that it also includes the latest GDP growth rate and traditional macroeconomic indicators, thus combining traditional and high-frequency series in the principal component analysis and considering the low-frequency variables have missing observations. The authors use twelve time series, capturing transport, electricity, pollution, consumer sentiment and labour market[3].

More relevant to us, the Bank of Portugal (BoP), in (Rua & Lourenço, 2020), created its own Daily Economic Index (DEI). The process for selecting the data and building the index is similar to the FED's WEI. The originality of their paper is the treatment of the specific problems of daily data. Using a LOESS de-seasonalization, they correct for the weekly seasonality effects, as well as the public holidays, even when those do not have a fixed date. MIDAS tests show that the index has predictive power. It also tracks lockdowns properly. They surveyed and collected all the available high-frequency data, which has started to be collected later for Portugal. As explained before, we aim at developing an index accessible to the public and that can be easily adjustable to one's needs, therefore we do not resort to the same unconventional data sources the BoP.

[2] Mixed-Data Sampling

[3] For this last economic proxy they use keywords from Google Trends (also known as Google search terms), namely "state support", "unemployment" and "short-time work".

Both the FED and BoP chose unconventional economic series that present strong co-movements with GDP growth. We follow a different variable selection procedure, as stated below. To estimate the indicator after the data pre-treatment, both authors follow a dynamic factor model with a single latent factor and so do we. Finally, the series are normalized to a common scale so that the authors give the index interpretable units.

The DEI performs well under the unusual circumstances of the SARS-CoV-2 outbreak. We share their motivation and take advantage of some of their methodology to improve our own indicator, such as the warning to some issues regarding data treatment, calendar effects, as well as the usage of a MIDAS regression to corroborate the results.

(Bortoli & Combes, 2015) assess whether Google Trends can forecast the short-term economic outlook in France or not. For that purpose, and since household consumption represents more than half of the French GDP, the authors would rather use as target variable the household expenditure, instead of GDP. They conclude that, as the consumption of goods and services is very heterogeneous, "using Google trending searches improves the forecasting of household expenditure in only a limited way", even though they observe/acknowledge benefits in using it for specific products, rather than for macroeconomic aggregates. We use their log-difference data treatment and include a Portuguese translation of their list of Google keywords in our own dataset. We consider the heterogeneity argument a very strong one, and since Portugal's share of consumption in GDP is even greater than that of France, the use of monthly retail sales as our target variable lets us think that our results could apply to GDP growth.

(INSEE, 2020) notes that Google search data bring information overlooked by traditional sources (e.g., Purchasing Managers' Index) in times of crisis. The forecasts that use high-frequency data[4] perform better at seizing the magnitude of the Covid-19 economic impact of early 2020. It finds that, although high-frequency data do seize significant portions of the variability of macroeconomic aggregates, they are not a big addition to forecasts that already include the usual variables. Another advantage is the timeliness of this data. Google Trends can indeed be obtained five days after the reference date, which would allow the user to receive timely information on the state of the economy.

[4] Notably, Google Trends.

Finally, (Woloszko, 2020) came up with an OECD Weekly Tracker for forty-six OECD and G20 countries resorting to Google Trends search data. Using a machine-learning procedure in panels (neural network model), and purging the data from their common long-term bias[5], the paper provides evidence that Google Trends data are very useful to track GDP growth rates (it outperforms the AR(4) benchmark) and shifts in consumption patterns.

The data includes "variables based on both Google Search categories and topics" (the latter being a collection of related keywords). Our approach differs from that of the author in aspects such as the machine-learning procedure employed and the indicator constitution process. Most importantly, we focus on extra-personalization of the keywords to a country-specific/the national economic context, rather than a broad translation of more general keywords which can be applied cross-country.

(Ferrara & Simoni, 2019) use Google Trends information to forecast GDP growth. They use the Sure Independence Screening procedure to pre-select the variables from their big dataset. They find its results superior to the other machine-learning methods they tested. To verify the power of the data, they add the series to a model including industrial production and opinion surveys, in a MIDAS regression, for the six largest Euro Area economies. They test 1776 variables per country. They conclude that Google data have two main advantages: they improve forecasts in periods of important swings, and they can be substitutes for other more official, but less timely, data, in the beginning of the quarter (i.e., in the first five weeks at least). They also show that pre-selection is useful, and that the data are relevant in real time conditions.

Within this literature, our index will try to answer several questions. We will wonder whether the previous conclusions are valid for the case of Portugal too. Most importantly, we'll ask whether Google Trends data are a sufficient source of information to build a DEI-like economic index (even though one should bear in mind that our target variable is monthly retail sales rather than quarterly GDP growth rate).

[5] This bias is associated with an ever-increasing usage of the internet in general and the Google search engine in particular. The ratio "interest over time" against "total search volume" of a certain keyword is downward-sloping in many cases, since it is not adjusted by the increasing total search volume over time.

4) Data

4.1. Motivations for using Google Trends

The omnipresence of Google in the daily life of economic agents makes it a privileged source of information on their economic decisions, particularly those relating to consumption. We believe Google searches are a correlate of purchasing plans or desires and can thus help us track consumer behaviour. Other economically relevant behaviours, anticipations, emotions (related to saving, uncertainty, fears…) may also be reflected in the data. Furthermore, around 66% of Portuguese GDP is driven by consumption: we have reason to believe that Google Trends may contain relevant information for this variable.

Due to those characteristics, we chose to study macroeconomic aggregates linked to private consumption. GDP was the natural candidate for the target variable, due to its centrality in economics. Another candidate was retail sales. We chose the latter in part because its composition is well defined and easier to map to Google searches that are representative of its content.

The problem of the number of observations and frequencies then arose. We decided to exploit the longest possible sample. Our dataset covers 207 months from January 2004 until February 2021 and has 900 weekly observations for each series. Weekly data was preferred, as daily data raised several specific problems (such as bank holidays or intra-weekly seasonality) while not providing an obvious timeliness gain – the Google Trends are indeed released every Tuesday, so they cannot be updated daily anyway.

Our compromise was to target a variable of monthly frequency to estimate the index and judge it graphically. As we still want to verify its performance in a multiple frequency framework, we will estimate a MIDAS model of quarterly retail sales. We will also run simple regressions of monthly and quarterly frequencies and compare them to an AR benchmark.

4.2. Target Series

For this project, we have decided to study retail sales in Portugal. The target variable is the index of deflated turnover and measures the retail trade (except of motor vehicles and motorcycles)[6] from Eurostat. The series used in variable selection is monthly, calendar-adjusted, but not seasonally adjusted. It is specified as a percentage change (i.e., relative to the same period in the previous year).

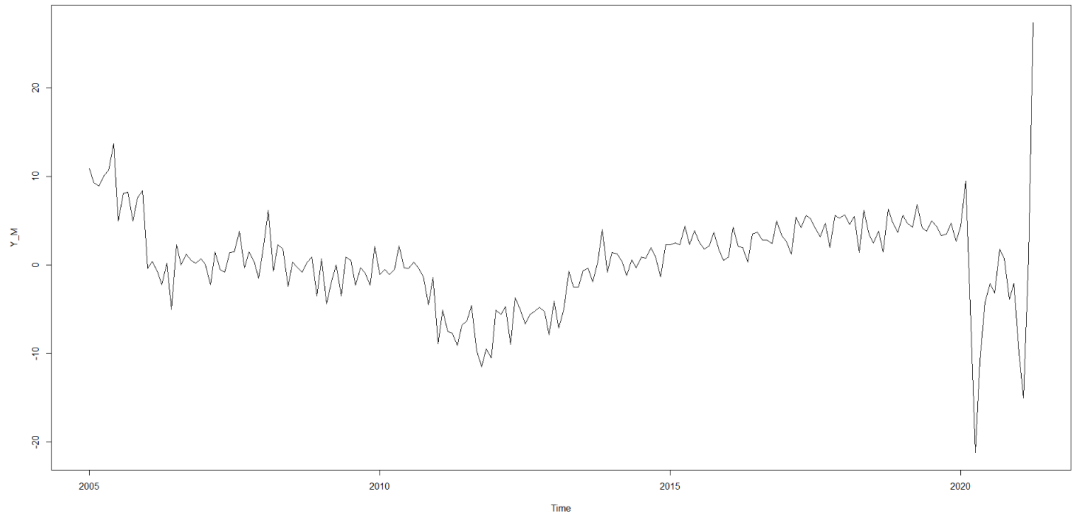


Figure 1 - Retail trade, except of motor vehicles and motorcycles, in Portugal (January 2005 - April 2021).

Even though the household consumption as percent of the Portuguese GDP has been decreasing since 1975, the country still belongs to the top 15 of European countries whose household consumption as a percentage of the respective GDP is the highest (63.89% in 2019, The World Bank). So it is reasonable to say that the Figure 1 reflects some of the economic fluctuations we have experienced in the last 16 years. This argument leads us to believe that, if our index performs well for retail, it may be informative for GDP as well, in later works.

One can observe the significant downward trend experienced in the second half of the first decade of the 2000's. This trend can arguably be decomposed in two major macroeconomic phenomena. First, the Dot-Com Bubble's aftermath of the early 2000's.

[6] We drew this time series from the "DBnomics" database. The importation code is the following: [Q.TOVV.G47.CA.PCH_SM.PT] for the quarterly series, [M.TOVV.G47.CA.PCH_SM.PT] for the monthly series. For the period under analysis, it is stationary according to an ADF test.

Second, and more important, the second major downward economic trend was due to the Great Recession. Such wider downturn finally reached its historical low between the years 2010 and 2014, during the Portuguese Sovereign Debt Crisis. Portugal experienced a rather shy recovery between 2013 and until the Covid-19 pandemic hit in the early 2020, followed by some significant booms and busts according to the lockdown measures of each period.

As has been described, the behaviour of retail trade can indeed track down contemporaneous major economic fluctuations. Hence we believe the retail trade series to be an economically relevant target variable. Success here will mean that Gtrends could have further uses.

4.3. Trend Choices – Theoretical Justification

Since the target variable we chose was retail sales, many of the keywords are related to retail trade. The remaining ones (concerning, for example, credit requests, job seeking, tourism, personal finance, …) are grouped into categories. Of course such an approach has a drawback. Even though it offers a privileged perspective on consumer behaviour, it is not expected to perform that well when it comes to business, supply-side decisions.

The choice of keywords was first aimed at representing the content of the retail trade index. Retail sales are built using the "Statistical classification of economic activities in the European Community", NACE Rev. 2. We referred to division 47, "Retail trade", of NACE Rev. 2. However, we did not try to target each category individually. We created thematic groups based on those. Not all keywords that might be relevant for retail forecasting can be included in division 47. Thus, out-of-NACE groups were added, based on the categories we found in the literature. The six quasi-NACE categories similar to division 47 are: "Apparel", "Foods, beverages, tobacco", "Leisure goods", "Other goods", "Other household items", "Tech products". The remaining four quasi-NACE categories different from division 47 are: "Conjuncture and labour market", "External factors", "Financial situation of households", "Mix of services".

Such categories were filled with keywords by three approaches. First, they were filled with general keywords that can be found in (Woloszko, 2020), (Bortoli & Combes, 2015) and

(INSEE, 2020). Second, with those directly inspired from the NACE subcategories. Finally, and as previously exposed in the introduction part of this report, the ten categories were filled with keywords we have obtained first from the literature, and then from our common knowledge and from that of colleagues of ours, interrogating them about their economic behaviours. What distinguishes our work from that of the current literature is precisely the combination of a big, diverse dataset and of the disaggregation level, that is, using specific keywords rather than general and pre-made categories. The gain will be a higher fit to our object of study. We can include Portuguese brands, specific cultural activities, and tourist-frequented Portuguese locations. All these dimensions can differ from one country to another and more intensively than the broad categories they are in.

Google Trends allows its users to access information about the relative frequency with which a keyword is searched during the requested period. An observation for a trend series is interpreted as the "number of hits" for a given keyword over time. We extracted 700 trends using the *Pytrends* API (for *Python*).

We allowed those who contributed to the dataset to write down car brands if they wished to, since even though the target series is retail trade without motor vehicles and motorcycles, the variable selection procedure we follow in this report is centred on correlation, not causality. Therefore, one could write a car brand as well as a keyword of any other dimension related to household consumption, since what would be relevant at the end of the day would always be correlation. The purpose of the Google Trends Weekly Index for Portuguese Retail Sales is to *nowcast* rather than to *explain per se*. One could argue that Portuguese people prefer to travel low-cost so it would make sense for the machine-learning procedure to present the keyword "Ryanair" as a relevant airline. In case such thing happens, it does not mean there is causality involved. There might be, and an inquisitive economist might even dig further on that, but the indicator *per se* does not tell us that. The same goes to car brands. Hence, the irrelevance of including them as keywords in the dataset to nowcast said target series.

A fundamental drawback to our method is that even though it offers a privileged perspective on consumer behaviour, there are not many reasons to believe it allows us to track business decisions (other than the signals that consumer behaviour provides us with). We lack supply-side elements in our analysis, then.

4.4. Importing the Trends

In the importation code, the user must choose the keywords, the country of origin of the searches, the period that fits the analysis the researcher intends to pursue, regional segmentation of the results, among others. The API interprets a keyword as a character vector with the Google Trends query keywords. The program can pick the information from several sources, such as the web (default), news, images, Froogle (shopping tool from Google, which ceased to exist back in 2007) and YouTube. Because the nature of this research is to use internet searches to nowcast retail sales, we use the web as the source of information.

Furthermore, the API lets the user filter the search results by the respective country of origin. We chose to limit ourselves to the Portuguese trends. An interesting extension of our work would be to include keywords from the leading countries visiting Portugal for tourism purposes (e.g., France, the UK, Spain,…).

The API delivers daily, weekly or monthly information depending on the requested timespan. Because the frequency returned depends on the timespan requested, we could not download the whole dataset in one go. The trends series are search volume indexes. They are rescaled so as to give the value 100 at the date (within the requested timespan) when the search intensity for a keyword was at its maximum. This problem of a relative index, combined with the frequency request constraint forced us to download several time intervals, to obtain weekly series. To harmonize the concatenated series, we multiplied their values by the monthly search volume index, which in turn was comparable for periods between 2004 and 2021. We then divided the series by 100, to retain the initial size of the series.

4.5. Data Treatments

To our knowledge, the study that treats Google Trends in the deepest fashion for nowcasting is that of OECD's economist Nicolas Woloszko. As such, we will follow the same data treatments as this author. Nonetheless, the first two treatments presented are exclusive to our dataset, since we faced some issues that the OECD author did not face, due to some of our keywords being less broad than those of Woloszko. First, we present the concept of search

volume indices, essential to this section. The operations that follow are presented in the order of their application.

4.5.1. Rationale

4.5.1.1. The Search Volume Index

The series we use are relative search volume indices. Their form is described by Equation 1.

$$SVI_{ct} = \frac{SV_{ct}}{SVT_t} * C_c$$

Equation 1 - Search Volume Index (SVI).

The search volume index is the ratio of the actual searches for the term, at time *t*, over the total searches on Google, also at time *t*. Because those searches trend upwards, due to Google's [and overall internet] increasing popularity, all the series mechanically suffer from a downward bias, which does not reflect actual searches for the terms (more on this issue below).

4.5.1.2. Dropping Low-Frequency Keywords

As a preliminary treatment, given the amount of zeros some series present due to the nature of the variable "interest over time", we need to drop some keywords in order not to bias the common time trend that is extracted from the search volume index afterwards (more on this topic further below). Hence, and since we did not manage to find any literature regarding the issue at hand, we arbitrarily decided to impose a 35% threshold. That is, were a series to present more than 35% of zeros in its observations and the respective keyword would be dropped.

4.5.1.3. Treating the Log of Zeros

The values retrieved by Google Trends can take a value of 0, which poses a mathematical problem for log-transformation. (Bellego & Pape, 2019) show there is no prevailing consensus in the literature. They review all AER papers from 2016 to 2020, and found that the most frequent solution adopted is to add a positive value to the variables. We followed their solution (adding the value 1 so that those logs still map to the value 0), for several reasons. Our zeros are "real zeros", that is, they are not measurement errors. The elimination of the common time trend observed among search volumes requires giving an additive form

to the series' constitutive components. The growth rates implied by the transformation make the data interpretable.

4.5.1.4. Common Time Trend

To eliminate the common time trend effect of the denominator on the series, (Woloszko, 2020) proposes a method that we reproduce and describe here and in the application part of this section.

We first take the log of the search volume indices, giving an additive form to (equation 1). We use an HP filter to eliminate the short-term cycles. We then apply PCA over the whole dataset of the long-term SVI trends. The first component obtained is rescaled to have the same mean and variance as the average log-SVI. This rescaled first component will be the common time trend (SVT element of equation 1), and we will subtract it from the initial log SVI, for all series. This transformation allows us to work with the search volume of each keyword and eliminates the downward bias in the series.

4.5.1.5. Seasonality

As is common practice in the Google Trends literature, we then apply differences to our series, previously log-transformed. A further benefit is that it allows us to treat potential non-stationarities. It also allows us to avoid the meaninglessness of the Google Trends' normalized values. We are now able to interpret our series as the increase or decrease of interest for the trend, relative to what it was on the reference week of the year before. (Woloszko, 2020) uses this transformation because it eliminates seasonality, and we reap this benefit as well.

4.5.1.6. Stationarity Testing

Due to some specificities of the machine-learning procedure we use, the log-SVI series must be stationary. For simplicity, and due to the size of the dataset, we decided to use the ADF test only, at the 10% threshold. If a series is non-stationary, it is discarded. Indeed, Google Trends data are normalized and passed by a log-difference already, so modifying them would reduce their interpretability, comparability, and informational content. The target series was tested for stationarity at the 5% threshold and posed no problem.

4.5.2. Application

The data treatments described for both the weekly and the monthly series were performed independently. We use a sample that goes from the first week of 2004 until the end of the first quarter of 2021.

From the initial dataset of 700 keywords, 290 keywords were eliminated by the process of dropping low-frequency keywords given the 35% threshold. We then pass the $log(SVI_{ct} + 1)$ transformation. We run the HP filters and PCA on those series and eliminate the common time trend. As expected, the common time trend we extract is positive (Figure 2)[7]. The common trends are roughly similar despite the two datasets' different frequencies. From 2019 onwards, however, a certain divergence appears and grows. Since the machine-learning feature selection selects the keywords with the monthly dataset, but computes the index with the weekly one, this may prove problematic, and it is an area of improvement.

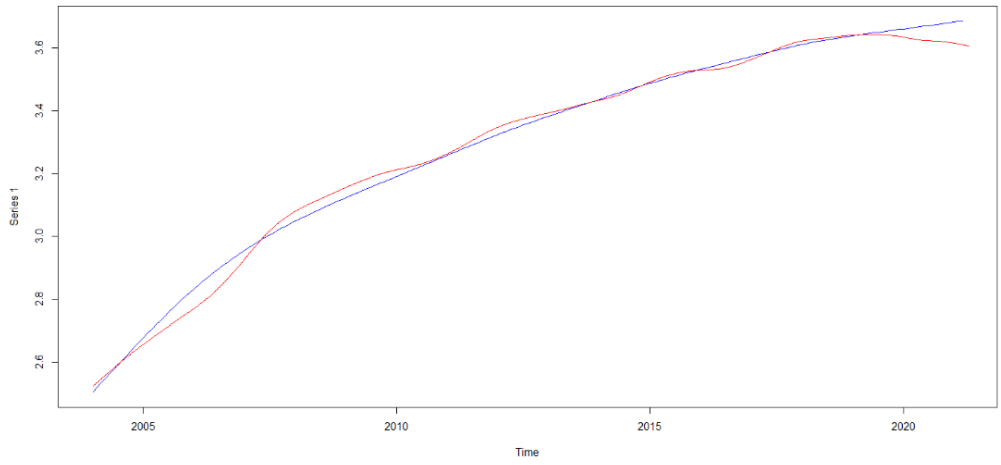


Figure 2 - The estimated common time trend. Note: Monthly frequency (blue), Weekly frequency (red).

The correlation screening (SIS) we apply further below requires all series to be stationary. Therefore, the ADF test eliminates another 72 series, leaving us with 338 series for the Sure Independence Screening to analyse.

[7] If Woloszko's interpretation holds and our keywords are sufficiently representative of the Portuguese Google Trends, the graph shows the upwards path of Google searches' usage from 2004 until 2021.

5) Methodology

5.1. Sure Independence Screening

5.1.1. Motivation

Given the size and content of our dataset, we face several problems. Our dataset's number of variables is higher than the time dimension (ultrahigh dimensionality case), which prevents direct application of the traditional techniques[8] (e.g., LASSO). Furthermore, using all the variables in the DFM would add much noise and detract us from our purpose. Although the average of the trends we selected does seem to reflect the economic activity to some extent, our "expert judgement" is insufficient for the task. Analysing and judging each of the series is impractical and arbitrary. Furthermore, Google searches are ambiguous, and their meaning for users may not correspond to the interpretation we extrapolate from them. For all the reasons above, we prefer to use an agnostic and rigorous variable selection approach.

Following the nowcasting successes of (Ferrara & Simoni, 2019) with Google Trends, we preselect the data with the SIS method (Fan & Lv, 2008), a feature-selection algorithm. The motivation for using the technique is first the volume of our sample[9], as well as the selection properties of SIS: sure independence, sparsity, the treatment of false negatives in correlation screening thanks to iterations, low computational requirements and the ability to confront the results over split-samples.

5.1.2. Basic Procedure

SIS is a framework for variable screening via independent correlation learning, in a setting of ultrahigh dimensionality (i.e., $p > n$, as in our sample which has only 207 monthly observations (n) for 338 variables (p)).

It consists in a two-stage procedure. Firstly, the variables that present the weakest absolute correlation with the target variable are eliminated, reducing the number of variables to a number below n. The Sure Screening property refers to the fact that we retain all the relevant covariates of the target variable. It is ensured by a few statistical conditions, among

[8] See "curse of dimensionality".
[9] I.e., *dimensions x observations*.

which normality of the residuals. Under these conditions, a threshold number of variables to keep can be defined: $d = \frac{n}{\log(n)}$. "Sure screening" means that these *d* highest correlates will include all the relevant variables.

After the first step, the number of remaining variables can still be too high: notably, we can have variables that the "sure screening" regarded as relevant before applying a regularization procedure but that, are not relevant for understanding the response. A regularization technique (a penalized regression method, e.g., LASSO or SCAD) is then applied to the set of *d* variables retained. The advantage of SIS-SCAD (which we chose to implement, over the LASSO), is reducing the false discovery rate, that is, the inclusion of variables that are not useful to predict the target series.

The "sure screening property" can be obtained under certain regularity conditions. It means that the procedure will be able to retain all the important features of the model. In contrast, the LASSO procedure would tend to outright eliminate some important variables, if they were highly correlated with other important explanatory variables. Numerical simulations under SIS show the procedure is able to achieve sparsity (i.e., to find which traits are relevant and sufficient to explain the response under study).

5.1.3. Iterative SIS

This procedure tackles the problem of the finite sample, and the potential failure of the aforementioned regularity conditions. Some important traits may be missed by the "naïve" correlation screenings. A feature that presents a low correlation, at first glance, would be eliminated, although it may have a high marginal correlation (i.e., it may be part of the true model). On the contrary, a feature that has a high correlation, but a low marginal one (i.e., it becomes useless in a model, once we put the other high correlation variables), may pass the screening.

To overcome this problem, the SIS procedure is simply re-iterated. After selecting the *d* first predictors, and eliminating those that were not relevant via regularization, we keep the residual of the regularized regression, and treat it as the target variable for the next iteration; hence, we will end up choosing the variables that have the highest "marginal utility", in a sense. Importantly, the iterative process also allows feature deletion at each step, thanks to the penalized likelihood estimations conducted.

5.1.4. Sample-Splitting ISIS

The goal here is to reduce the false selection rate. The dataset is randomly split in two. Over each half, we run the correlation screening. We obtain a subset of predictors for each subsample. The important predictors are supposed to appear in each subset, with a probability tending to one (sure screening). If we take the intersection of the two subsets, the probability limit of having all the relevant predictors must hold for the intersection too. However, the falsely selected variables will be eliminated. The process then goes on to iterations of the previously described kind, but with sample-splitting at each correlation screening step. The sparse penalized likelihood estimations are then conducted on the intersection of the subsets of variables selected in each subsample, for each step.

5.2. Dynamic Factor Model

Once we determined, at a monthly frequency, which trends are relevant for forecasting retail sales YoY growth, we return to the weekly series. We move from feature selection with the SIS to feature extraction with the DFM.

This way, we create an index in the spirit of the FED and the Bank of Portugal's. This framework supposes that the series in the vector of trends X_t are driven by a set of common unobserved variables, F_t, plus a vector of idiosyncratic shock ε_t. The DFM will try to pin down the unobserved variable F_t guiding retail sales.

$$X_t = \lambda(L) * F_t + \varepsilon_t$$

Equation 2 – Dynamic factor model

For determining the number of factors, we follow the (Lewis, Mertens, Stock, & Trivedi, 2020) practice: we suppose that our data is explained by only one dynamic factor. We get our indicator after a rescaling to give it the same mean and standard deviation as the retail sales growth series. For practicality, we have chosen to use the "*dynfactoR*" package. This imposes a constraint on the model: only the state equation is dynamic, not the measurement equation (i.e., F_t can depend on F_{t-p} but the X_t cannot).

To test for the adequacy of our DFM specification, we verify whether the residuals (idiosyncratic components of the Google trends) are white noise. The lag order can be chosen

by application of the AIC or BIC criteria. In general, the literature suggests that using one or two lags is sufficient for whitening the residuals.

6) Results

6.1. Sure Independence Screening

We ran the three variants of the procedure we described. Although the iterative SIS yielded the regularized regression with the lowest residual sum of squares, we preferred the sample-splitting iterative SIS, which selected less variables (24 vs 36). The process with a splitting of the sample is more "aggressive" in its elimination of keywords, which yields a more parsimonious index, and is desirable. Most importantly, the sample-splitting reduces the likelihood that we choose exclusively series that fit well only in crises or only in peaceful times. The sample-splitting ISIS ran 6 iterations and selected 24 series (Table 1).

Table 1 - Selected keywords from the sample-splitting ISIS.

Audi	Chicco	Citroën	Emprego	Ford	Honda
Ikea	Nissan	Pingo doce	Ryanair	Area	Alentejo
Arroz	Ebay	GPS	Oeiras	REMAX	Peugeot
Roupa	seguros	Seat	smart	Toyota	Volkswagen

6.2. Dynamic Factor Model

We ran the DFM using 6 lags and the three estimators (Figure 3). The results are consistent among the different estimators. We retained the Quasi-Maximum Likelihood (QML) estimator for our index. Our index is the QML estimated factor, rescaled to have the same mean and variance as monthly retail sales. The index obtained is stationary.

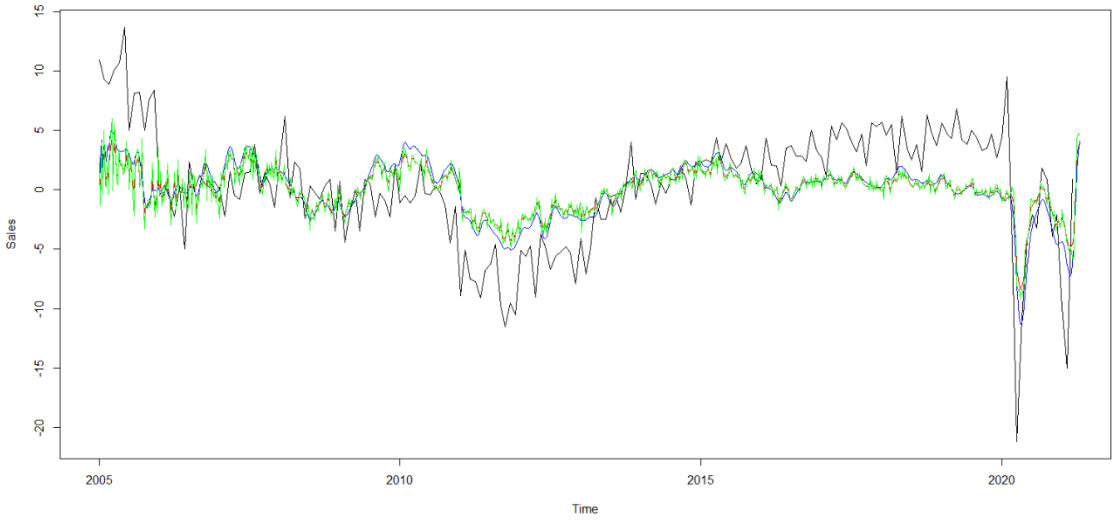


Figure 3 - First Principal Components from the DFM, 6 lags. Note: Quasi-maximum likelihood estimate (blue), Two-step estimate (red), Principal component analysis estimate (green).

6.3. Narrative Timeline of our Indicator

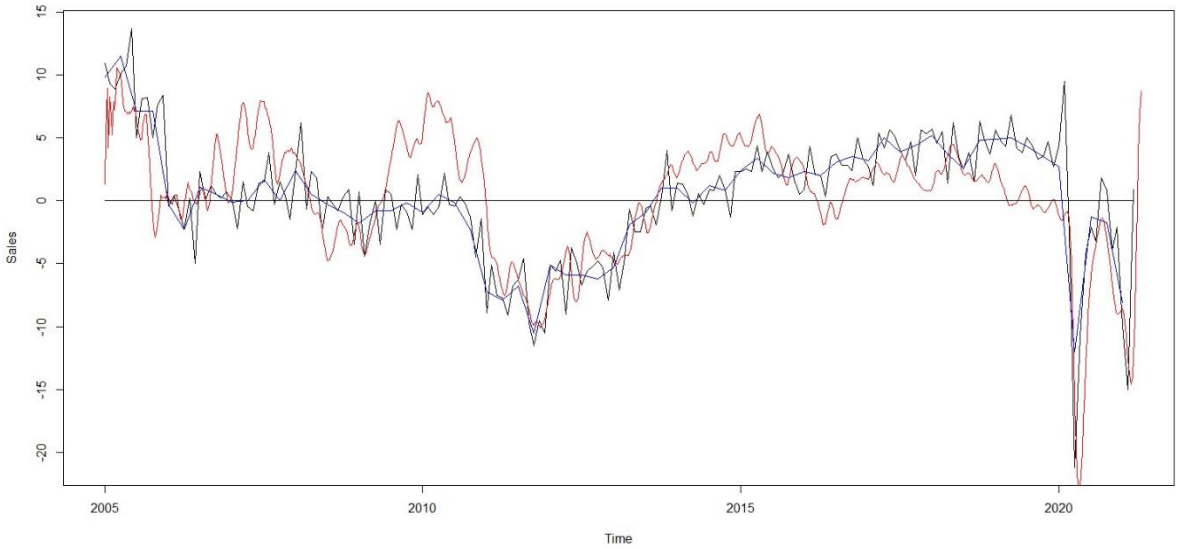


Figure 4 - The Google Trends Weekly Index for Portuguese Retail Sales from the QML estimator (red), plotted against the target series (weekly YoY retail sales growth (black)) and the quarterly YoY retail sales growth (blue).

The most eye-catching difference of our weekly index (Figure 4, red) from that of the target variable (Figure 4, black) concerns the deviation in the period 2008-2011. While the

indicator presented a significant growth rate during those years, retail sales have in fact nearly stagnated. However, the indicator captures remarkably well the deepest recessions (namely, the Portuguese sovereign debt crisis and the Covid-19 crisis). The second problem relates to the failure of the index to properly capture the accelerating trend of retail trade that is observed during the period of 2015-2020.

One can observe a very substantial and unexpected recession in early 2020, coinciding with Portugal's first lockdown (March 2020), along with the declaration of the State of Emergency in the whole country. A couple months in 2020, the trough is observed around April, followed by a significant recovery, which the weekly index captures as well. In mid-2020 there was a modest YoY increase (most likely associated with fewer Covid-19 new cases and the season in which tourism in Portugal peaks), bringing to an end the unprecedented economic upheaval.

Regarding the second half of 2020, there are two major fluctuations to highlight. There is a relatively smaller YoY decrease after the summer and a modest progression during the Christmas season, as predictable. Since the interpretation of the index is done looking at the relative variation related to the year before, one has to pay closer attention to the behaviour of the index in the end of 2020 and beginning of 2021. The last trough in the weekly index corresponds to the post-Christmas strict lockdown. As one approaches February and March of 2021, there is an observable upward trend, because indeed retail sales were not that much below the retail sales' levels of February and March of 2020, years that were relatively bad nonetheless.

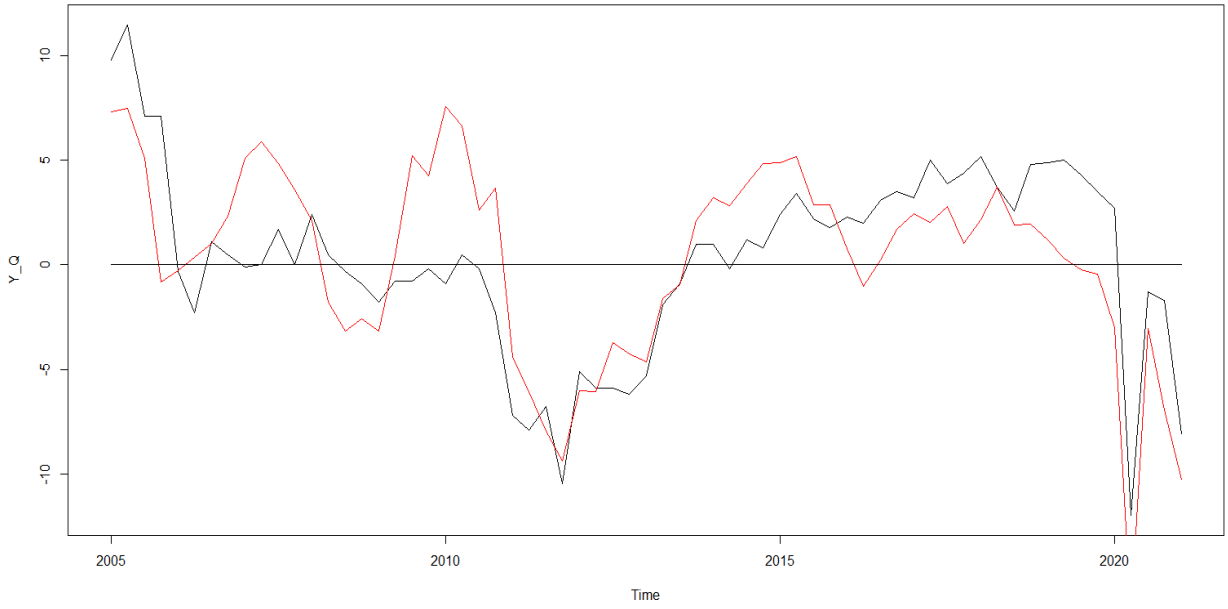


Figure 5 - The Google Trends Quarterly Index for Portuguese Retail Sales from the QML estimator (red), plotted against the quarterly YoY retail sales growth (black).

As per the quarterly version of our indicator, even though it does not achieve a performance as convincing as that our weekly index, it somewhat follows the dynamics of the quarterly retail sales. It should be noted that the indicator follows both the Portuguese sovereign debt crisis and Covid-19 crisis remarkably well. The challenge on this specific regression is that the SIS resorts to monthly frequency, while we plot the quarterly index. In spite of this mismatch, the indicator seems to have some informational content. However, it is even clearer here that our indicator is biased. It detects decreases in a timely fashion and appears to seize their amplitude. However, it fails on the upside, often providing false signals. This indicator thus seems more apt to detect negative events.

7) Performance Evaluation

To evaluate our indicator in a high frequency framework, we need to make sure that the target variable's frequency is a multiple of the indicator's frequency. Because weeks do not fit

neatly into months, we will evaluate the index with regard to the quarterly target[10]. We consider 13 weeks per quarter.

As in our reference papers for the WEI and DEI, we only conduct in-sample exercises and leave more detailed out-of-sample examinations for further research. We limit ourselves to commenting the MAE, RMSE and R^2 of our three models. The publication delay of monthly retail sales is about one month, which determines the forecaster's information set. Both the outputs of OLS and MIDAS regressions including and excluding lags of retail sales are presented in the Annex section of the report.

7.1. Autoregressive Benchmark

Before moving forward, we need to establish a benchmark for Portuguese retail sales, at the quarterly frequency. We define the adequate AR(p) model, using the auto-ARIMA function from the R software. The chosen model was an ARIMA(1,0,1).

7.2. Natural Nowcast

Like (Lewis, Mertens, Stock, & Trivedi, 2020), we rescaled our series to have the same mean and standard deviation as our target series. This means our quarterly averages will provide natural nowcasts, and we can thus immediately compute the performance criteria, and compare them to our AR(1) benchmark.

7.3. Quarterly Same Period Linear Regressions

Following (Lewis, Mertens, Stock, & Trivedi, 2020) and (Rua & Lourenço, 2020), we regress our quarterly target on our quarterly index, in an OLS framework. We do not include lags of the target variable, though.

[10] Bear in mind that, due to the variable selection being conducted for monthly frequencies (and not quarterly) we may not achieve the best performance evaluation we could get.

7.4. MIDAS Regressions

Finally, we take advantage of the high frequency fashion of our index in three different Mixed-Data Sampling exercises to assess whether our index shows any forecasting power (Tables 2 and 3). (Ghysels, Santa-Clara, & Valkanov, 2004) first proposed this type of model, which allows to regress a low frequency variable (e.g., quarterly retail) over a higher frequency variable (e.g., a weekly indicator). We use the "*midasr*" package to implement the technique, with the "*midas_r*" function.

The following exercises will help test whether the indicator has a leading content or whether it should stick with the original trait of it (a coincident indicator, rather than a leading one). First, we consider that our only information is on the 4 first weeks of Google trends for the quarter to forecast. Second, we consider the case in which we have all the Google trends observations for the quarter, but no retail sales data. In the third test, we suppose the same as in the previous one, but knowing the past quarter of YoY quarterly retail sales growth rate.

A second battery of tests considers one lag of quarterly retail sales (Table 3). Unlike GDP, retail sales are released monthly and with a monthly lag. Nowcasting quarterly retail sales puts us in a special case in which we already know part of the answer (e.g., in the last day of the quarter, we already know two thirds of the information on quarterly retail sales growth). Because our aim is to understand our indicator and assess its capabilities, rather than to produce a good forecast, we disregard intra-quarter new retail sales information. These specifications are realistic since the quarterly retail sales would be obtained with a one-month delay.

7.5. A Brief Comment on the Performance Metrics

The exercises without lags present systematically worse MAEs than the ARMA benchmark. The ARMA's RMSE is beaten when 8 weeks of information become available. The R^2 of our index is relatively satisfying when used on its own. Once all weeks are available, we capture 64% of the variations in quarterly retail sales.

Once we include lags, our R^2 approximates 75% in all specifications. All specifications beat the ARMA benchmark on MAE and RMSE. As expected, the further we move into the

quarter, the better the nowcast gets. The improvement is not drastic, however, showing that early quarter information is the most valuable. Even when controlling for the series' past behaviour, our index brings information.

Table 2 - In-sample performance for the regressions without retail sales growth rates' lags.

Performance metrics	ARMA (1,1)	Natural nowcast	OLS (quarterly)	MIDAS (4 weeks)	MIDAS (8 weeks)	MIDAS (all weeks)
R^2	--	--	0.5422	0.54700	0.62106	0.64024
Adj. R^2	--	--	0.535	--	--	--
MAE	1.75499	2.69520	2.49799	2.42994	2.29826	2.17955
RMSE	2.86913	3.30900	3.01251	2.99675	2.74085	2.67058

Table 3 - In-sample performance for the regressions with retail sales growth rates' one quarter lag.

Performance metrics	ARMA (1,1)	OLS (quarterly)	MIDAS (4 weeks)	MIDAS (8 weeks)	MIDAS (all weeks)
R^2	--	0.7442	0.75813	0.77144	0.78225
Adj. R^2	--	0.7358	--	--	--
MAE	1.754991	1.73575	1.67650	1.63759	1.58736
RMSE	2.869138	2.18816	2.12757	2.06821	2.01873

7.6. "Naïve" Correlation Benchmarks

Because one should not use a sledgehammer to crack a nut, we compare the performance of our index to simpler correlation screenings. We build datasets for the feature extraction by dynamic factor model using Pearson correlation coefficients. We create one index based on a factor of the 5 most correlated Google trends and one of the 25 most correlated Figure 6).

The SIS index captures deep falls the best, as exemplified in the first Covid-19 economic dip and the Portuguese sovereign debt crisis. The "naïve" indices are not very reactive to falls

in the growth rate. However, they share the same false signals as our index. They are unambiguously worse at capturing retail sales' fluctuations.

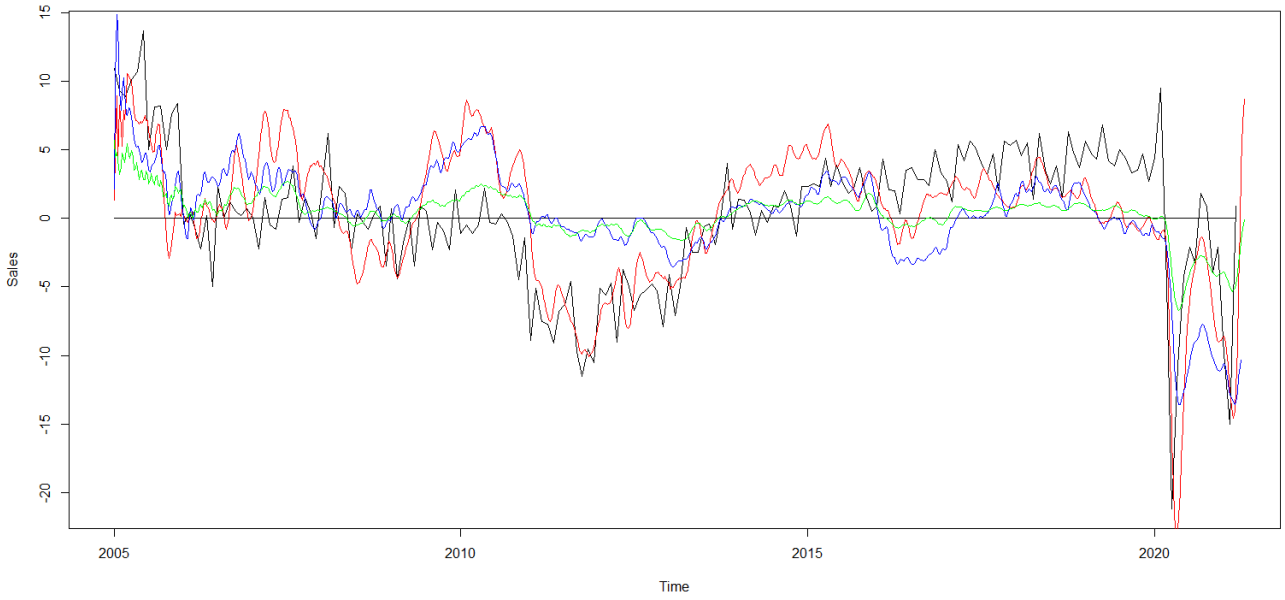


Figure 6 - Comparison of the index from SIS, against the 5 highest (blue) and 25 highest (green) correlations. (DQML, 6 lags)

8) Further Improvements

Although our index presents some successes, we believe it does not yet exploit Gtrends to their fullest.

Some of the improvements that should be made (notably, about data pre-processing) are mentioned in (Woloszko, 2020). The first one is that Google Trends has a multiple sampling problem. Upon importation, we obtain a sample of the actual search data, so there is a sampling variance when obtaining the data. Woloszko solves the problem by importing six times each series and averaging the values. As Google itself imposes restrictions to the importation requests an individual user can do, extracting each series a single time proved too heavy, so we left this potential bias for further research. Another source of concern is the breaks in data collection. The OECD economist warns that, in January 2011 and in January 2016, methodological changes to the data collection process have caused breaks in the series.

The author subtracts the January 2011 (2016) to January 2010 (2015) growth rate's difference from all the subsequent growth rates. This works under the hypothesis that all the difference between January 2010 (2015) and January 2011 (2016) is due to the methodological change. The gain is a reduction of growth rate outliers. Applying such treatment to our dataset renders negative values (problematic to the log transformation that follows). We advance the suggestion that such issue arises in our dataset, contrary to Woloszko's, because we resort to more "lower-level" keywords, that is, some of our search terms present lower search volumes since they are less broad (more specific to the Portuguese context), and therefore more prone to reach negative values in case they get subtracted by a given value. We put this transformation aside for this report. Our estimates might suffer from this bias. However, such treatment is unique to (Woloszko, 2020)[11], and the rest of the literature does not apply it.

Another source of concern is the breaks in data collection. The OECD economist warns that, in January 2011 and in January 2016, methodological changes to the data collection process have caused breaks in the series. The author subtracts the January 2011 (2016) to January 2010 (2015) growth rate's difference from all the subsequent growth rates. This works under the hypothesis that all the difference between January 2010 (2015) and January 2011 (2016) is due to the methodological change. The gain is a reduction of growth rate outliers. Applying such treatment to our dataset renders negative values (problematic to the log transformation that follows). We suppose that this issue arises in our dataset, and not in Woloszko's because we resort to a lower-level of information, i.e. keywords. Some of our search terms present low search volumes since they are less broad than whokle categories, and therefore more prone to reach negative values in case they we subtract a given value. We put this transformation aside for this report. Our estimates might suffer from this bias. However, the treatment is unique to (Woloszko, 2020)[12], and the rest of the literature does not apply it, as far as we know.

A third source of improvement us the extension of our dataset. Indeed, several keywords have been eliminated, and a simple extension would enrich the information set to choose from. Furthermore, we focused only on keywords typed in Portugal. Since 2014, Portugal has known

[12] As well as the actual recognition of the problem.

a boom in tourism. Including keywords from the tourists' countries of origin may solve the poor performance of our index after 2014, for example.

Finally, the importance of each series in the composition of the final index may change over time for several reasons, for example: internet usage may change, crises may affect the preferred goods of consumption and the propensity to save. So one would have to periodically update the dataset and re-run the variable selection procedure A more technical alternative would be to try to apply a Threshold DFM, which could allow a change in the values of the parameters and actually warn the user whenever the dataset is getting outdated.

9) Conclusion

In this project, we tried to answer the traditional problems of timeliness and noisiness facing decision makers. We show that the index obtained gives a coherent story on retail sales, although it is better suited for negative episodes and should not be fully trusted on its upswings. The in-sample prediction exercises show that our index is informative, even when we include lags of the target variable. Although this index is not a perfect standalone tool, it is somewhat informative and can accompany other analyses and indices for cross-checking. Most importantly, the index has two properties the Google Trends' literature had found earlier: it is most useful earlier in a quarter, and in periods of brutal downturns.

Regarding the methods employed, we found that SIS is a useful tool in selecting the relevant features to understand the time series phenomenon at hand, and it beats a basic correlation ranking.

On a more practical side, one asset of our approach is that it requires low computational resources and data access. Anyone with an internet connection and RStudio can track the weekly developments in the Portuguese economy, with a very minimal data collection effort (downloading 24 series). This low entry cost makes the method applicable to other countries, with worse statistical frameworks and workforce than developed countries such as Portugal. Our approach is also quite flexible: we only need to change the targeted variable in the SIS, to produce indices for other variables. Further developments may thus provide a quick and easy, but informative and widely available, nowcasting tracking tool.

10) Bibliography

Bellego, C., & Pape, L.-D. (2019). *Dealing with the log of zero in regression models.* Center for Research in Economics and Statistics.

Bortoli, C., & Combes, S. (2015). *Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues.* Institut national de la statistique et des études économiques.

Eraslan, S., & Götz, T. (2020). *An unconventional weekly economic activity index for Germany.* Deutsche Bundesbank.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*.

Ferrara, L., & Simoni, A. (2019). *When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage.* Center for Research in Economics and Statistics.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS Touch: Mixed Data Sampling Regression Models.* CIRANO.

INSEE. (2020). *"High-frequency" data are especially useful for economic forecasting in periods of devastating crisis.* Institut national de la statistique et des études économiques.

Lewis, D. J., Mertens, K., Stock, J. H., & Trivedi, M. (2020). *Measuring Real Activity Using a Weekly Economic Index.* New York: Federal Reserve Bank.

Rua, A., & Lourenço, N. (2020). *The DEI: tracking economic activity daily during the lockdown.* Bank of Portugal.

Woloszko, N. (2020). *Tracking activity in real time with Google Trends.* Paris: OECD Economics Department Working Papers.

11) Annex

11.1. Outputs of Regressions Excluding Lags of Retail Sales

Table 4 - OLS regression of quarterly YoY growth of retail sales over the quarterly average of our index.

```
> Qreg = lm(Y_Q ~ INDEX_Q) #Create the linear regression
> summary(Qreg)

Call:
lm(formula = Y_Q ~ INDEX_Q)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3465 -2.1415 -0.3222  2.1916  7.5619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10475    0.38054   0.275    0.784
INDEX_Q      0.70544    0.08166   8.639 2.74e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.06 on 63 degrees of freedom
Multiple R-squared:  0.5422,    Adjusted R-squared:  0.535
F-statistic: 74.62 on 1 and 63 DF,  p-value: 2.744e-12
```

Table 5 - MIDAS regression including the first four weeks of our indicator

```
> summary(MIDAS_H0_1stMonth) #Summary from the estimate

MIDAS regression model with "ts" data:
Start = 2005(1), End = 2021(1)

 Formula Y_Q ~ mls(INDEX, 9:12, 13)

 Parameters:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02661    0.75219   0.035   0.9719
INDEX1      -0.75843    0.30506  -2.486   0.0157 *
INDEX2       1.60797    1.31790   1.220   0.2272
INDEX3      -0.78770    2.62406  -0.300   0.7651
INDEX4       0.60237    1.37751   0.437   0.6635
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 3.119 on 60 degrees of freedom
> extract.midas_r(MIDAS_H0_1stMonth, include.rsquared = TRUE, incl


                  coef.      s.e.          p
(Intercept)  0.02661312 0.7521895 0.97189348
INDEX1      -0.75842586 0.3050627 0.01571599
INDEX2       1.60796687 1.3179030 0.22720281
INDEX3      -0.78770142 2.6240579 0.76507406
INDEX4       0.60237377 1.3775124 0.66347029

                 GOF dec. places
R$^2$       0.5470075        TRUE
Num. obs.  65.0000000       FALSE
$\\sigma^2$  3.1191209        TRUE
```

Table 6 - MIDAS regression including the first eight weeks of our indicator

```
> summary(MIDAS_H0_2ndMonth) #Summary from the estimate

MIDAS regression model with "ts" data:
Start = 2005(1), End = 2021(1)

 Formula Y_Q ~ mls(INDEX, 5:12, 13)

 Parameters:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1473     1.0287  -0.143   0.8867
INDEX1        0.6481     1.7895   0.362   0.7186
INDEX2        1.7192     3.2237   0.533   0.5959
INDEX3       -4.5339     3.5471  -1.278   0.2065
INDEX4        4.2364     2.1550   1.966   0.0543 .
INDEX5       -2.0185     1.8038  -1.119   0.2679
INDEX6       -1.2882     1.7744  -0.726   0.4709
INDEX7        2.1578     2.5767   0.837   0.4059
INDEX8       -0.1495     1.2377  -0.121   0.9043
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

 Residual standard error: 2.953 on 56 degrees of freedom
> extract.midas_r(MIDAS_H0_2ndMonth, include.rsquared = T


                 coef.     s.e.          p
(Intercept) -0.1472950 1.028710 0.88665878
INDEX1       0.6480811 1.789512 0.71859909
INDEX2       1.7191737 3.223722 0.59594418
INDEX3      -4.5338626 3.547148 0.20646159
INDEX4       4.2364347 2.154964 0.05427601
INDEX5      -2.0185084 1.803802 0.26790455
INDEX6      -1.2882346 1.774427 0.47086205
INDEX7       2.1578437 2.576667 0.40589667
INDEX8      -0.1495473 1.237686 0.90425965

                   GOF dec. places
R$^2$        0.6210679        TRUE
Num. obs.   65.0000000       FALSE
$\\sigma^2$  2.9529018        TRUE
```

Table 7 - MIDAS regression including all the weeks of our indicator

```
> summary(MIDAS_HO_Full)

MIDAS regression model with "ts" data:
Start = 2005(1), End = 2021(1)

 Formula Y_Q ~ mls(INDEX, 0:12, 13)

 Parameters:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2101     1.0447  -0.201    0.841
INDEX1       -3.4055     2.5143  -1.354    0.182
INDEX2        5.4373     4.2945   1.266    0.211
INDEX3        1.5419     4.3121   0.358    0.722
INDEX4       -4.2599     7.2114  -0.591    0.557
INDEX5       -4.8062     8.2210  -0.585    0.561
INDEX6        7.8994     6.5951   1.198    0.237
INDEX7        3.4656     5.9609   0.581    0.564
INDEX8       -9.3689     8.4088  -1.114    0.270
INDEX9        3.4366     4.6424   0.740    0.463
INDEX10       0.7932     3.1300   0.253    0.801
INDEX11      -1.0049     2.6443  -0.380    0.706
INDEX12       0.5286     3.2971   0.160    0.873
INDEX13       0.4791     1.3245   0.362    0.719

 Residual standard error: 3.015 on 51 degrees of freedom
> extract.midas_r(MIDAS_HO_Full, include.rsquared = TRUE,


                 coef.     s.e.         p
(Intercept) -0.2100813 1.044682 0.8414234
INDEX1      -3.4054520 2.514256 0.1815613
INDEX2       5.4372577 4.294533 0.2112343
INDEX3       1.5419248 4.312137 0.7221348
INDEX4      -4.2598831 7.211436 0.5573221
INDEX5      -4.8061958 8.221003 0.5613781
INDEX6       7.8993577 6.595133 0.2365493
INDEX7       3.4656426 5.960859 0.5635320
INDEX8      -9.3688644 8.408812 0.2704287
INDEX9       3.4366041 4.642409 0.4625348
INDEX10      0.7932334 3.129963 0.8009531
INDEX11     -1.0049185 2.644323 0.7055023
INDEX12      0.5285827 3.297150 0.8732662
INDEX13      0.4791490 1.324528 0.7190329

                   GOF dec. places
R$^2$        0.6402494        TRUE
Num. obs.   65.0000000       FALSE
$\\sigma^2$  3.0149353        TRUE
```

11.2. Outputs of Regressions Including Lags of Retail Sales

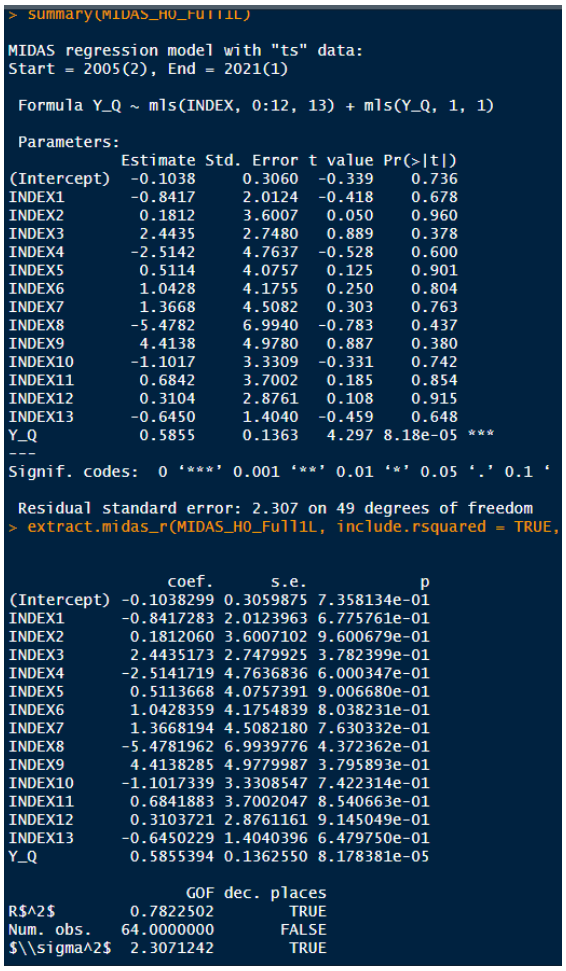
Table 8 - MIDAS regression including the first four weeks of our indicator

```
> summary(MIDAS_H0_1stMonth1L) #Summary from the estimate

MIDAS regression model with "ts" data:
Start = 2005(2), End = 2021(1)

 Formula Y_Q ~ mls(INDEX, 9:12, 13) + mls(Y_Q, 1, 1)

 Parameters:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08016    0.31925  -0.251    0.803
INDEX1       1.25761    1.36229   0.923    0.360
INDEX2      -1.80239    2.55036  -0.707    0.483
INDEX3       1.73382    1.60119   1.083    0.283
INDEX4      -0.83043    0.84702  -0.980    0.331
Y_Q          0.59382    0.11700   5.075 4.27e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 2.235 on 58 degrees of freedom
> extract.midas_r(MIDAS_H0_1stMonth1L, include.rsquared = TRUE,


                  coef.      s.e.            p
(Intercept) -0.08016137 0.3192505 8.026302e-01
INDEX1       1.25761336 1.3622862 3.597465e-01
INDEX2      -1.80239133 2.5503603 4.825700e-01
INDEX3       1.73382403 1.6011868 2.833621e-01
INDEX4      -0.83043015 0.8470159 3.309515e-01
Y_Q          0.59381788 0.1169974 4.272101e-06

                   GOF dec. places
R$^2$        0.7581367        TRUE
Num. obs.   64.0000000       FALSE
$\\sigma^2$  2.2349152        TRUE
```

Table 9 - MIDAS regression including the first eight weeks of our indicator

```
> summary(MIDAS_HO_2ndMonth1L) #Summary from the estimate

MIDAS regression model with "ts" data:
Start = 2005(2), End = 2021(1)

 Formula Y_Q ~ mls(INDEX, 5:12, 13) + mls(Y_Q, 1, 1)

 Parameters:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12062    0.35580  -0.339    0.736
INDEX1       1.11048    1.19649   0.928    0.357
INDEX2      -0.01737    2.90270  -0.006    0.995
INDEX3      -3.40887    4.35733  -0.782    0.437
INDEX4       4.06794    3.32078   1.225    0.226
INDEX5      -1.93070    3.03090  -0.637    0.527
INDEX6       0.48397    2.87189   0.169    0.867
INDEX7       1.00657    2.66096   0.378    0.707
INDEX8      -0.89041    1.58975  -0.560    0.578
Y_Q          0.57389    0.13552   4.235 8.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 2.252 on 54 degrees of freedom
> extract.midas_r(MIDAS_HO_2ndMonth1L, include.rsquared = TRUE,


                  coef.      s.e.           p
(Intercept) -0.12062094 0.3557997 7.359143e-01
INDEX1       1.11047551 1.1964905 3.574803e-01
INDEX2      -0.01736602 2.9027027 9.952486e-01
INDEX3      -3.40887282 4.3573336 4.374379e-01
INDEX4       4.06794128 3.3207767 2.258931e-01
INDEX5      -1.93069797 3.0309036 5.268136e-01
INDEX6       0.48397357 2.8718871 8.668030e-01
INDEX7       1.00656829 2.6609626 7.067114e-01
INDEX8      -0.89040543 1.5897459 5.777323e-01
Y_Q          0.57389387 0.1355155 8.949765e-05

                   GOF dec. places
R$^2$         0.7714458        TRUE
Num. obs.    64.0000000       FALSE
$\\sigma^2$  2.2515824        TRUE
```

Table 10 - MIDAS regression including all the weeks of our indicator



```
> summary(MIDAS_HO_Full1L)

MIDAS regression model with "ts" data:
Start = 2005(2), End = 2021(1)

 Formula Y_Q ~ mls(INDEX, 0:12, 13) + mls(Y_Q, 1, 1)

 Parameters:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1038     0.3060  -0.339    0.736
INDEX1       -0.8417     2.0124  -0.418    0.678
INDEX2        0.1812     3.6007   0.050    0.960
INDEX3        2.4435     2.7480   0.889    0.378
INDEX4       -2.5142     4.7637  -0.528    0.600
INDEX5        0.5114     4.0757   0.125    0.901
INDEX6        1.0428     4.1755   0.250    0.804
INDEX7        1.3668     4.5082   0.303    0.763
INDEX8       -5.4782     6.9940  -0.783    0.437
INDEX9        4.4138     4.9780   0.887    0.380
INDEX10      -1.1017     3.3309  -0.331    0.742
INDEX11       0.6842     3.7002   0.185    0.854
INDEX12       0.3104     2.8761   0.108    0.915
INDEX13      -0.6450     1.4040  -0.459    0.648
Y_Q           0.5855     0.1363   4.297 8.18e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

 Residual standard error: 2.307 on 49 degrees of freedom
> extract.midas_r(MIDAS_HO_Full1L, include.rsquared = TRUE,


                  coef.      s.e.           p
(Intercept) -0.1038299 0.3059875 7.358134e-01
INDEX1      -0.8417283 2.0123963 6.775761e-01
INDEX2       0.1812060 3.6007102 9.600679e-01
INDEX3       2.4435173 2.7479925 3.782399e-01
INDEX4      -2.5141719 4.7636836 6.000347e-01
INDEX5       0.5113668 4.0757391 9.006680e-01
INDEX6       1.0428359 4.1754839 8.038231e-01
INDEX7       1.3668194 4.5082180 7.630332e-01
INDEX8      -5.4781962 6.9939776 4.372362e-01
INDEX9       4.4138285 4.9779987 3.795893e-01
INDEX10     -1.1017339 3.3308547 7.422314e-01
INDEX11      0.6841883 3.7002047 8.540663e-01
INDEX12      0.3103721 2.8761161 9.145049e-01
INDEX13     -0.6450229 1.4040396 6.479750e-01
Y_Q          0.5855394 0.1362550 8.178381e-05

                   GOF dec. places
R$^2$        0.7822502        TRUE
Num. obs.   64.0000000       FALSE
$\\sigma^2$  2.3071242        TRUE
```

Table 11 - OLS regression of quarterly YoY growth rates of retail sales over the quarterly average of our index

```
> summary(QregL)

Call:
lm(formula = Y_Q ~ INDEX_Q + mls(Y_Q, 1, 1))

Residuals:
    Min      1Q  Median      3Q     Max
-5.7443 -1.4914  0.0443  1.6508  6.5236

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.1568     0.2818  -0.556     0.58
INDEX_Q        0.4490     0.0689   6.517 1.55e-08 ***
mls(Y_Q, 1, 1) 0.5249     0.0727   7.218 9.76e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.241 on 61 degrees of freedom
  (1 observation effacée parce que manquante)
Multiple R-squared:  0.7442,    Adjusted R-squared:  0.7358
F-statistic: 88.72 on 2 and 61 DF,  p-value: < 2.2e-16
```