CENTRO DE INVESTIGACIÓN Y DOCENCIA ECONÓMICAS, A.C.

¿CAN GOOGLE TRENDS IMPROVE PREDICTIONS OF CONSUMPTION? EVIDENCE FROM A BAYESIAN STRUCTURAL TIME SERIES MODEL OF ECOMMERCE TRANSACTIONS IN MEXICO

TESINA

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN ECONOMÍA

PRESENTA

JOSÉ LUIS MENDOZA OLAYA

DIRECTOR DE LA TESINA: DR. RICARDO MASSA ROLDÁN

CIUDAD DE MÉXICO 2023

Abstract

In recent years, electronic commerce has emerged as a significant economic driver, making the surveillance of ecommerce markets crucial for various stakeholders including business owners, investors, and policymakers. A large portion of online consumers use search engines like Google to look for specific products, brands, or marketplaces. In this sense, it is relevant to ask the following question. Can data from Google search queries reveal insights into the ecommerce sector in Mexico? Data from ecommerce transactions is obtained from Banxico and data from search queries is available in GoogleTrends.com. The econometric framework present in this study follows the Bayesian Structural Time Series (BSTS) methodology, which combines time series and regression analysis. Results prove that the inclusion of contemporary information from Google Trends in the BSTS framework does reflect a lower cumulative absolute error than simple Structural Time Series models. Particularly, three clusters of significant Google Trends queries are identified: 1) online marketplaces that belong grocery stores, 2) two-sided marketplaces and 3) online apparel stores.

Index

Chapter 1

Introduction

Electronic commerce -or ecommerce- has become an important driver of economic activity in past years. Moreover, monitoring ecommerce activity is pertinent for several stakeholders, such as business owners, investors, and policymakers. Investors and business owners may benefit from analyzing ecommerce transactions to make informed decisions for making profit. Likewise –by studying ecommerce activity– academics may elaborate models to identify patterns in consumer behavior, while in turn public institutions may develop research-based policies that promote the development of the ecommerce sector.

In general, participants in any market economy must take informed decisions about their productive activities to maximize profits. In this sense, there is a common need for reliable information, such as the level and trend of economic activity in certain markets or industries. Commonly, the most trustful sources of public economic information are central banks and official statistics agencies. In many countries, part of these authorities' responsibilities is to keep track of different aspects of the economy and to inform the public in this regard. These organizations collect and publish official data on national ecommerce transactions usually from payment records which they have access to, including ecommerce transactions.

However, there are also some private organizations that provide economically relevant information for the public, specifically related to ecommerce. Google launched Google Trends in 2006, a platform that allows users to analyze web search trends all over the world. Search queries may be assumed to be related to ecommerce activity, as both phenomena occur in the digital space. Particularly, one can think that a proportion of digital costumers use Google to search for the specific product, brand, or marketplace they want access to and finally make the purchase that meets their demands. Thus, Google Trends data provides information about market trends or purchasing intent that might be useful for designing marketing and sales strategies.

Furthermore, it is plausible to identify revealed preferences and characterize the ecommerce market in a particular region, since a proportion of these users might engage in online commercial transactions. In this context, one can ask the following question. Can Google Trends series provide insightful information regarding the ecommerce sector in Mexico? Econometric models can be used to model ecommerce activity using time series and regression models. In this vein, the main hypothesis of this study claims that models that incorporate data provided by Google Trends can outperform simple time series models of ecommerce.

To determine what is the contribution of Google Trends for analyzing ecommerce in Mexico, this study considers an exploratory analysis of ecommerce transactions obtained from Mexico's central bank (Banxico) and a set of 68 series from Google Trends related to online marketplaces and specific goods and services. The selected econometric framework of present in this study is based in the Bayesian Structural Time Series (BSTS) methodology, which combines time series and regression analysis. Time series models use historical data to forecast future trends, while regression models analyze the relationship between ecommerce transactions and other variables, in this case the Google Trends data.

The motivation for studying ecommerce activity in Mexico comes from two main reasons. The first one is that developing countries usually suffer from having a lack of information, or at least reduces sources compared to those that developed countries have. In addition, Mexico stands out as one of the fastest growing ecommerce markets among developing countries. The second one is that literature from the past decade has shown that data from web search queries may be useful to improve predictions in a wide variety of economic contexts. Efforts to analyze Google Trends data using BSTS were originally led by economists Hal Varian alongside Scott and Choi in the past decade.[1] Specifically, Varian argues that even though web search data can sometimes capture irrational tendencies, Google Trends might still be useful for research since it may provide contemporaneous information that would otherwise be missing,[2] something that specially developing countries could take advantage of.

Without further ado, the rest of this article is the following. First, the role of Google Trends in scientific research and Economics is discussed in section II (Literature review); secondly, more information related to data used in this study is presented in section III (Data); subsequently, methodological details are presented in section IV (Methodology); then, outputs

[1] Choi, Hyunyoung, and Hal Varian. "Predicting the Present with Google Trends." *Economic Record* 88 (2012): 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

[2] Choi, Hyunyoung, and Hal Varian. "Predicting the Present with Google Trends." *Economic Record* 88 (2012): 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

of the model are illustrated and interpreted in section V (Results & discussion), and finally conclusions are briefly summarized in the last section VI (Conclusions).

Chapter 2

Motivation & literature review

Google Trends has surprisingly favored the conditions for studying information seeking behavior –mainly for medical and economic purposes– by providing updated information about when and how frequently people are searching for information using Google's search engine for a given region. Even some large international and public organizations have already included this source of data in their research projects. Some examples of these are the following: in 2018, the International Monetary Fund (IMF) published a working paper using Google Trends data to narrow information gaps in low income development countries[3] and, later in 2021, to pursue global economic activity tracking.[4] The latter endeavor is also part of the Organization of Economic Cooperation and Development (OECD) as a topic in their research agenda of 2020.[5]

A fundamental contribution to understand the role of Google Trends in recent scientific research was presented by economists Choi, Jun & Sun Yoo in their article "Ten years of research change using Google Trends: From the perspective of big data utilizations and applications." The authors gather 657 scientific articles related to Google Trends from 2007 to 2016 to run a social network analysis (SNA). In other words, they performed a relational analysis among research papers linked to Google Trends to characterize the use of the platform in science through this decade. One important implication of their results is that the more related research topics tend to eventually combine and produce new research fields, especially if there are incentives provided by commercial opportunities or if it of particular interest for the intellectual communities driven by to particularly successful scientific articles.[6]

The authors also describe the chronological summary of the appearance of the first Google Trends related research papers and what went after that. Apparently, research emerged from the

---

[3] Narita, Futoshi, and Rujun Yin. "In Search of Information:" *IMF Working Papers* 18, no. 286 (2018): 1. https://doi.org/10.5089/9781484390177.001.

[4] Marini, Marco, Paul Austin, James Tebrake, Alberto Sanchez, and Chima Simpson-Bell. "Using the Google Places API and Google Trends Data to Develop High Frequency Indicators of Economic Activity." *IMF Working Papers* 2021, no. 295 (2021): 1. https://doi.org/10.5089/9781616355432.001.

[5] OECD. "Tracking Activity in Real Time with Google Trends." *OECD Economics Department Working Papers*, 2020. https://doi.org/10.1787/6b9c7518-en.

[6] Seung-Pyo Jun, Hyoung Sun Yoo, and San Choi, "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87. https://doi.org/10.1016/j.techfore.2017.11.009.

Computer Science field in 2007, one year after the platform was launched. Later, in 2011, Google Trends gained relevance in the Public Health and Economics fields. Then, after 2013, this interest reached some research communities dedicated to touristic, political, public opinion, environmental and global financial crisis research fields. Since 2016, other sources of big data (such as Android and Twitter data, among other sources) have been combined with data from web search queries to elaborate more sophisticated models; such is the case of sentiment analysis, which captures the subjective content of data (good, bad or neutral).[7]

Prior intellectual efforts in the search behavior field were less productive before Google Trends, since researchers had access to dramatically more limited search engines. In this sense, as suggested by Watts and Porter in 1997,[8] Choi et al. proposes an impact evaluation for evaluating how technology influenced technological progress by assessing if new scientific research based on Google Trends is related to an increased production of patents. The authors provide evidence to sustain that after the platform was launched in 2006 and the first Google Trends based studies were published, there was a reported increase in Marketing and Information Retrieval code patents in the U.S. two years later in 2008.[9] Likewise, the appearance of the first published papers associated to health and psychology keywords after 2012 –such as epidemiology, influenza, and sentiment– preceded increases in code patents for commercialization related with the same words after 2014.[10]

Furthermore, the results of the referred SNA show that there are three large clusters of articles in scientific research that use Google Trends data: Computer Science & Information Systems, Economics & Finance, and Bio Science & Medicine.[11] Results suggest that, mostly in general, Computer Science academics center their attention in comparative analysis of Google

---

[7] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[8] Watts, Robert J., and Alan L. Porter. "Innovation Forecasting." *Technological Forecasting and Social Change* 56, no. 1 (1997): 25–47. https://doi.org/10.1016/s0040-1625(97)00050-4.

[9] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[10] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[11] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

and other online search engines.[12] As of Epidemiological and Medical research, academics tend to focus on Google Trends capacity to monitor and surveil disease transmission.[13] The first and most influential paper within this cluster was published by Ginsberg et al. (2009), who achieved to outperformed predictions of influenza outbreaks made by the Centers for Disease Prevention and Control (CDC) in the United States of America.[14]

Finally, according to Jun et al., the Economics & Finance cluster stands out as the most active one in terms of published articles among Google Trends research papers up to 2016. According to the authors, some of these endeavors even motivated the appearance of new technologies and business initiatives as well.[15] During early stages of research using the platform's data, simple static analysis was the usual standard. However, as research progressed in the economics and finance fields, academics turned their attention to forecasting.[16]

Seemingly, with the right strategy, forecasting models of economic activity may benefit from the inclusion of early and contemporaneous information about changes in search tendencies from Google Trends. Some relevant examples of macroeconomic proven applications include forecasting unemployment and economic crisis,[17] investment,[18] consumer confidence,[19] and stock market movements.[20] At the micro level, the examples of applications

[12] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[13] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[14] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[15] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[16] Jun et al., "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications," *Technological Forecasting and Social Change* 130 (2018): 69–87.

[17] Askitas, Nikolaos, and Klaus F Zimmermann. "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly* 55, no. 2 (2009): 107–20. https://doi.org/10.3790/aeq.55.2.107.

[18] Da, Zhi, Joseph Engelberg, and Pengjie Gao. "In Search of Attention." *The Journal of Finance* 66, no. 5 (2011): 1461–99. https://doi.org/10.1111/j.1540-6261.2011.01679.x.

[19] Vosen, Simeon, and Torsten Schmidt. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends." *Journal of Forecasting* 30, no. 6 (2011): 565–78. https://doi.org/10.1002/for.1213.

[20] Preis, Tobias, Helen Susannah Moat, and H. Eugene Stanley. "Quantifying Trading Behavior in Financial Markets Using Google Trends." *Scientific Reports* 3, no. 1 (2013). https://doi.org/10.1038/srep01684.

vary from estimations of the likelihood of consumer's adoption of certain goods,[21] to monitoring economic activity in the tourism, retail, housing and entertainment industries.[22]

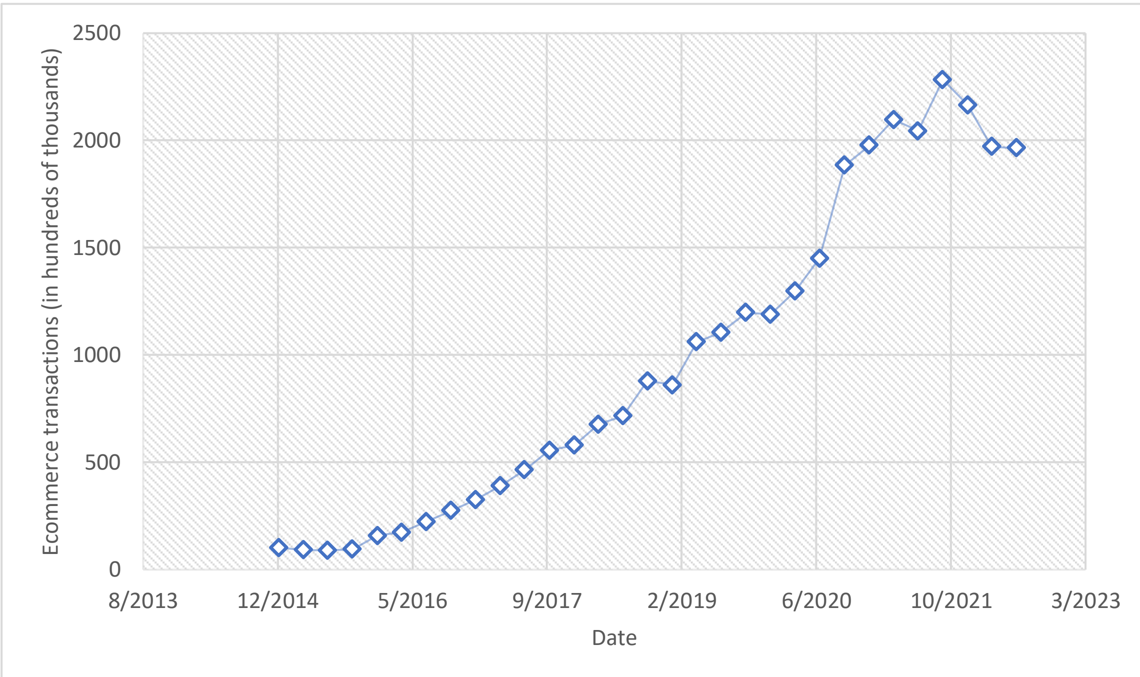
[21] Jun et al., 2014a, 2014b; Jun and Park, 2016; Jun and Park (2017).

[22] Choi, Hyunyoung, and Hal Varian. "Predicting the Present with Google Trends." *Economic Record* 88 (2012): 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Chapter 3

Data

Figure 1. Total amount of ecommerce transactions in Mexico from the 1st quarter of 2015 to the 3rd quarter of 2022.



Source: Economic Information System website of Mexico's central bank (Banxico).

The first source of data for this study is found in the Economic Information System website of Mexican Central Bank. Banxico provides data for authorized payment transactions in websites. The available series consist of the quantity and the monetary amount of authorized transactions, as well as disaggregated series for debit and credit card transactions. The selected series to estimate with the BSTS framework is the total quantity of authorized ecommerce transactions, both credit and debit. Time series data is downloaded for the period of 2015 to 2022. The visual representation of the quantity of ecommerce transactions (in hundreds of thousands) for the period of study is presented in Figure 1.

The second source of information used for the analysis is obtained from the Google Trends website and consists of a set of 68 series indicating the independent relative popularity over time of Google search queries related to marketplaces with presence in Mexico and other queries considered to be relevant. Some of these online marketplaces range from grocery stores, pharmacies, two-sided markets, food chains and airlines. The series downloaded from Google Trends are scaled from 0 to 100, with 0 depicting the period of lowest search frequency and 100 depicting the period of highest search frequency. All the queries included in the dataset are listed in Table 1.

Table 1. List of the analyzed Google Trends series.

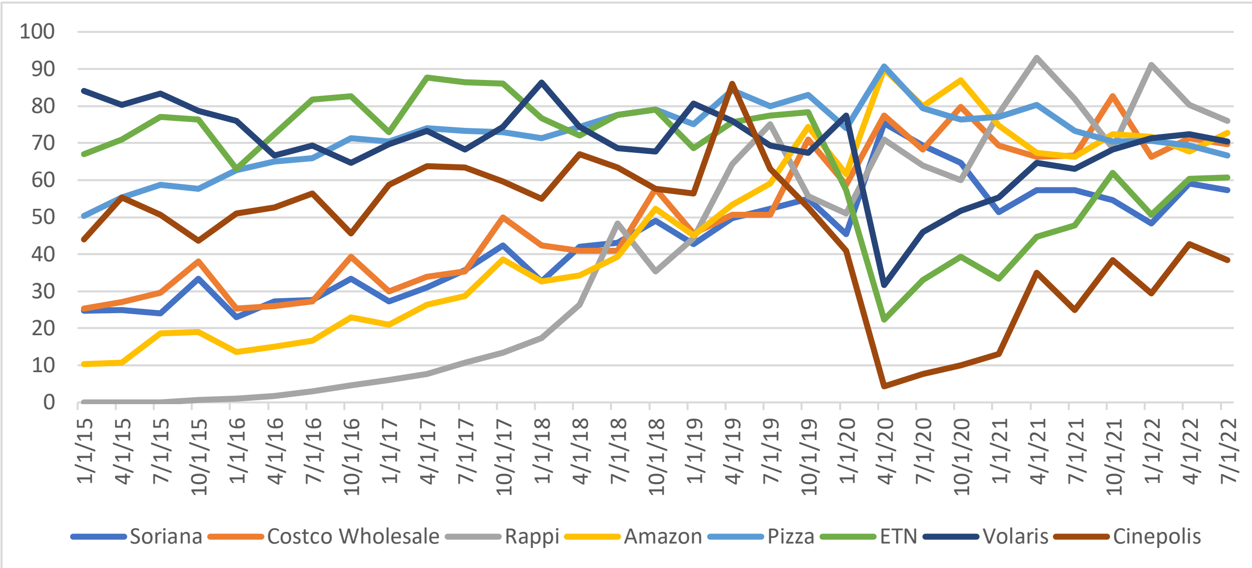
Google Trends Series			
1	ADO	*36*	Viva Aerobus
2	Farmacia San Pablo	*37*	Netflix
3	Farmacias del Ahorro	*38*	Aeromexico
4	H&M	*39*	Volaris
5	Farmacias Guadalajara	*40*	Airbnb
6	Farmacias Similares	*41*	Elektra
7	Samsung	*42*	Shoes
8	Dominos	*43*	Mobile phone accessories
9	Motorola	*44*	Sneakers
10	Alibaba	*45*	Shopping
11	ETN	*46*	Discount
12	Sears	*47*	Buy online
13	Ticketmaster	*48*	Online shopping
14	Cinepolis	*49*	Doctoralia
15	Cinemex	*50*	Shopify
16	Didi	*51*	Dominos
17	Rappi	*52*	Delivery
18	Cornershop	*53*	Pizza
19	Online casino	*54*	Pull&Bear
20	Adidas	*55*	SHEIN
21	Nike	*56*	Zara
22	Xiaomi	*57*	Levi Strauss & Co.
23	Smartwatch	*58*	Price Shoes

24	Liverpool	*59*	Bershka
25	El Palacio de Hierro	*60*	The Home Depot
26	Coppel	*61*	PUMA
27	Costco Wholesale	*62*	Marti
28	Chedraui	*63*	Innovasport
29	Soriana	*64*	STEREN
30	The Home Depot	*65*	Iphone
31	Amazon Prime Video	*66*	Computer keyboard
32	Sams Club	*67*	Computer mouse
33	Walmart	*68*	Laptop
34	Amazon	*69*	Headphones
35	Mercado Libre	*70*	Mask

Source: series downloaded from Google Trends.

One common problem arises at this point: the ecommerce transactions series are arranged quarterly and Google Trends series are arranged monthly. This implies it is not viable to proceed with the analysis due to different dimensionality between datasets. However, to solve this, Google Trends series are simply transformed by calculating quarterly means. A subset of the resulting quarterly Google Trends series is visually represented in Figure 2 to illustrate the relative popularity of some relevant online marketplaces over the period of study.

Figure 2. Google Trends series depicting the scaled (0-100) relative popularity of large companies over time in Mexico.



Source: GoogleTrends.com

Chapter 4

Methodology

This study explores the capacity of Google Trends to explain ecommerce activity by estimating the quantity of authorized ecommerce transactions using Google Trends data in the context of a Bayesian Structural Time Series framework. In this attempt, the analysis may be divided into a series of steps. First, several models of ecommerce transactions are estimated allowing for different parameter settings that adjust the belief about the proportion of series from the Google Trends dataset that are expected to be included in every sampled model –this parameter is called the Expected Model Size (EMS).

Computed results from different model specifications are compared based on the associated cumulative absolute error, which represents the overall fit of the models based on deviation between observed and predicted values. To obtain the cumulative absolute error the absolute difference between predicted values and observed values is calculated. Then, the resulting absolute differences are summed up to obtain the total cumulative absolute error. Also, predictions for future years using different EMS settings are briefly presented and contrasted. Finally, the estimated Marginal Posterior Inclusion Probabilities are furtherly explored to identify which are the most relevant regressors from Google Trends for explaining ecommerce activity.

Now, it is appropriate to present the model specifications. The baseline model consists of pure Structural Time Series (STS), and it does not incorporate any information from Google Trends. The usual expression of STS is written in an additive form that allows simple interpretations, as well as an independent estimation of each component. The observational equation of the model is expressed as:

$$y_t = \mu_t + \tau_t + \varepsilon_t,$$

$$\text{where } \varepsilon_t \sim^{iid} N_m(0, \Sigma_\varepsilon),$$

$$\text{and } t = 1, 2 \ldots, n.$$

The y_t term corresponds to the quantity of ecommerce transactions in each quarter. The first time series component –represented by μ_t– is called the local linear trend component. It captures the underlying level and direction of y_t at period t. The form of the state equation μ_t under stationarity conditions is defined as:

$$\tilde{\mu}_{t+1} = \tilde{\mu}_t + \delta_t + \tilde{u}_t,$$
$$\text{where } \tilde{u}_t \sim N_m(0, \Sigma_\mu),$$
$$\text{and } \delta_{t+1} = D + \tilde{\rho}(\delta_t - D) + \tilde{v}_t,$$
$$\text{where } \tilde{v}_t \sim N_m(0, \Sigma_\delta).$$

The second time series component τ_t captures the seasonal trend of y_t. This component can be thought of as a set of S=4 dummy variables to capture quarterly seasonality. Moreover, there is a constraint imposed over the correspondent dynamic coefficients that forces them to sum zero in expectation for every period of S seasons. The state equation τ_t is defined as:

$$\tau_t = -\sum_{k=0}^{S-12} \tau_{t-k} + w_t.$$

For the inclusion of the Google Trends series, a regression component represented by ξ_t is added to the original expression of the STS baseline model. This is defined as a Bayesian Structural Time Series model and may be expressed by the following equation:

$$y_t = \mu_t + \tau_t + \xi_t + \varepsilon_t,$$
$$\text{where } \varepsilon_t \sim^{iid} N_m(0, \Sigma_\varepsilon),$$
$$\text{and } t = 1, 2 \ldots, n.$$

The regression component ξ_t captures the coefficients associated with the Google Trend series. Most coefficients are expected to be zero, a circumstance referred as sparsity. In this vein, an accurate manner to proceed is through the Spike-and-Slab methodology. The Spike-and-Slab approach incorporates Bayesian inference technique for variable selection and shrinkage of a given regressor pool. It addresses the "fat regression" problem, characterized by a larger number of potential regressors compared to the available observations for the objective series.

The estimation process of the model consists of two steps. In a first stage, a prior belief is assigned to all potential regressors from the Google Trends series based on subjective judgments about their relevance, where predictors are categorized as either "spike" –insignificant– or "slab" –potentially relevant–. In the case of this study, all regressors are assigned the same prior inclusion probability in every estimated model. However, each model has a different correspondent EMS, a hyperparameter that indicates the number of expected significant predictors. This allows to perform a sensitivity analysis allows by comparing the performance of different models with different beliefs about the relevance of the Google Trends series.

For the second stage of the estimation process, Markov Chain Monte Carlo (MCMC) and Bayesian Model Averaging methods are employed. On the one hand, MCMC allows to sample from the posterior distribution of the model parameters. It iteratively updates the coefficient values based on the observed ecommerce transactions data and the assigned prior distributions, resulting in the computation of the posterior distribution. In general, the posterior distribution represents an updated belief about the regressor's relevance, incorporating both the available data and the prior knowledge.

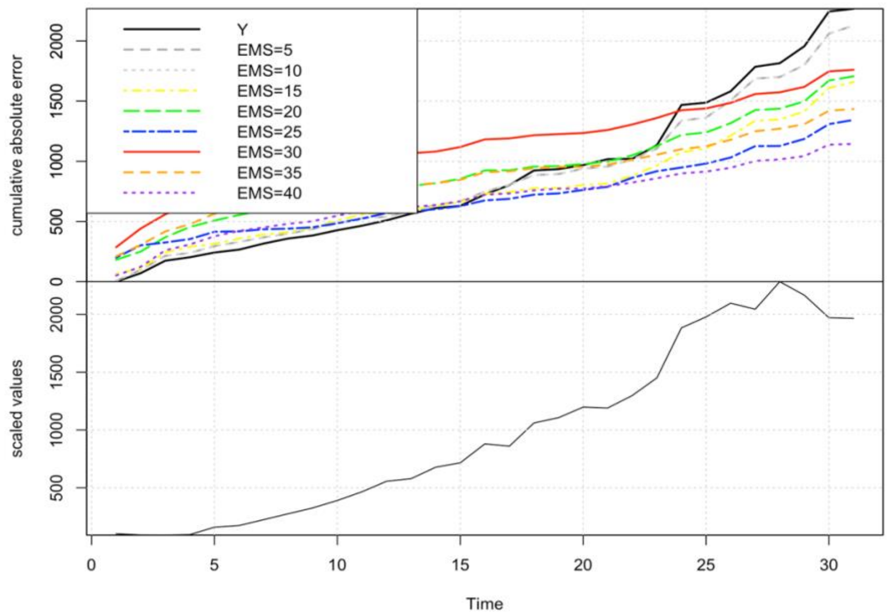
On the other hand, Bayesian model averaging combines the Spike-and-Slab models from the MCMC samples to assess uncertainty and build a more robust model. The purpose of this approach is to approximate the true underlying model. In this process, the coefficient estimates from different models are averaged and weighted based on posterior model probabilities, which are assigned to different models according to their fit and performance given the observed data.

Finally, marginal posterior inclusion probabilities are calculated by integrating the over the resulting posterior distribution of the model parameters from the Bayesian averaged model to measure whether a particular regressor should be included. It is calculated as the proportion of posterior samples in which the regressor's coefficient is non-zero, indicating that the regressor is included in the model. Reggresors associated with higher inclusion probabilities are potentially more likely to have a significant impact on the model and, in this sense, it may be interesting to further analyze them. Thus, it is possible to identify the relative importance of Google Trends series.

Chapter 5

Results & discussion

Figure 3. Cumulative absolute error of models with different expected model size (EMS) I.



Source: calculated with data from Google Trends and Banxico.

Figure 3 depicts the cumulative absolute error calculated for several models with different EMS parameter settings. The curve on the bottom part of the plot corresponds to the observed authorized ecommerce transactions in Mexico for the period of study, from the first quarter of 2015 to the third quarter 2022. Results confirm that complementing pure Structural Time Series –represented by Y– with Google Trends data improves the model fit. The cumulative absolute error associated with the baseline model –Structural Time Series) is the larger than the error associated with the rest of the models that incorporate Google Trends information. Even the most restricted model, with EMS=5, accumulates less error than the baseline model.

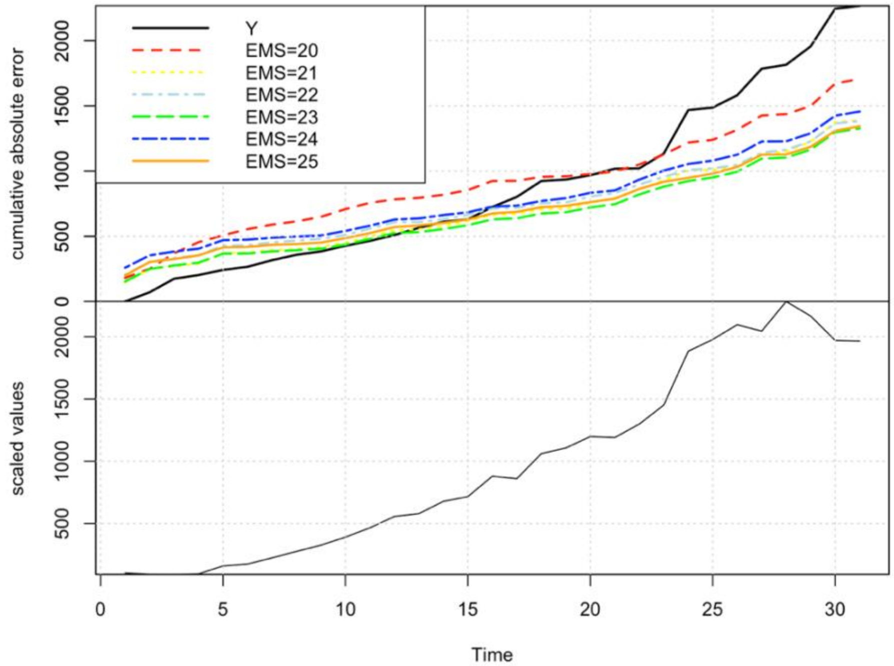
Moreover, the different cumulative absolute error among models seem to indicate that it constantly decreases as the expected model size increases. However, there are some exceptions to this situation. The model associated with the lowest cumulative absolute error is based on an

EMS=40; the next one is based on an EMS=25, while the following one is based on an EMS=35. Similarly, results show that for an EMS equal to 15, 20 and 30, the cumulative absolute error is very similar.

Another important observation is that the baseline model does not adapt to large changes in the rate of growth of ecommerce series as the rest of the models. The last statement may be verified by contrasting the smoothness of the cumulative absolute error curves. This suggest that Google Trends data is particularly useful to reduce cumulative absolute error when the local slope of the changes abruptly; at least, compared to what would be observed using Structural Time Series, which captures only local linear trend and seasonality.

Now, a second round of BSTS models are estimated, but now with a value of EMS equal to 20-25, since there are apparently large improvements in model fit in this interval. An illustration depicting the cumulative absolute error curves is presented in Figure 4. Results indicate that most of these model specifications produce very similar absolute error curves. Parallelly, it is confirmed that allowing a larger EMS parameter setting reduces cumulative absolute error. Here, the EMS=23 parameter setting seems to achieve the best fit, but just by a small difference.

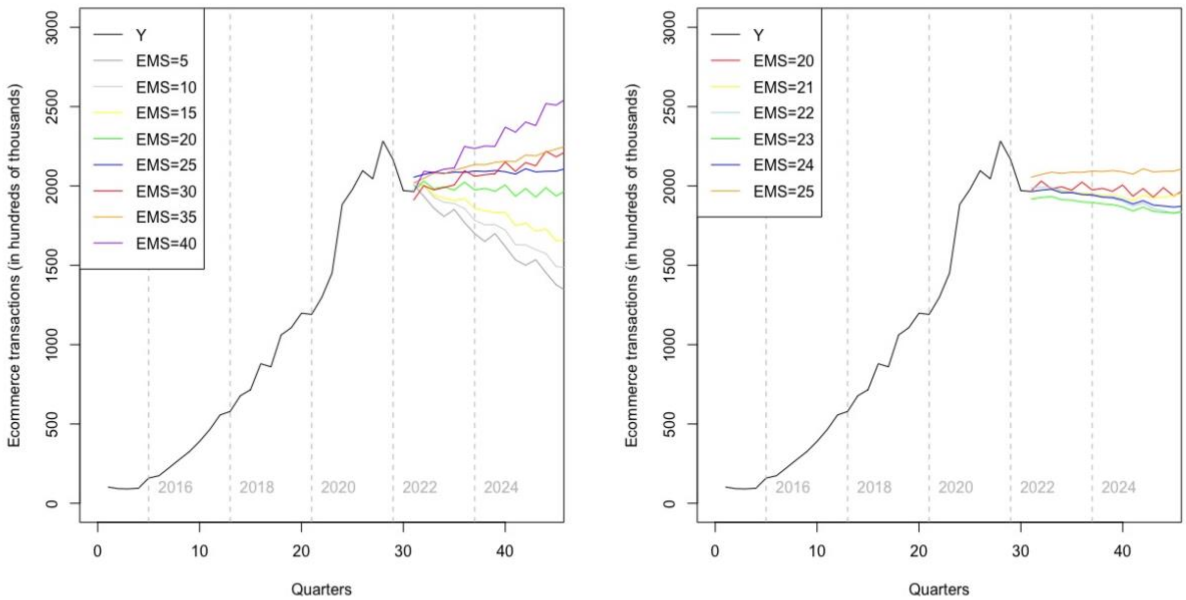
Figure 4. Cumulative absolute error of models with different expected model size (EMS) II.



Source: calculated with data from Google Trends and Banxico.

Figure 5 depicts predictions starting from the fourth quarter of 2022 up to 2026, calculated sets of models. Here, the Y curve represents observations from ecommerce transactions in the period of study. Different trends are observed, from negative to positive by increasing EMS. The second set of models which selected for further analysis of a narrower EMS interval also produces a narrower range of predictive outcomes and a constant level of ecommerce transactions in the future.

Figure 5. Marginal Posterior Inclusion Probabilities

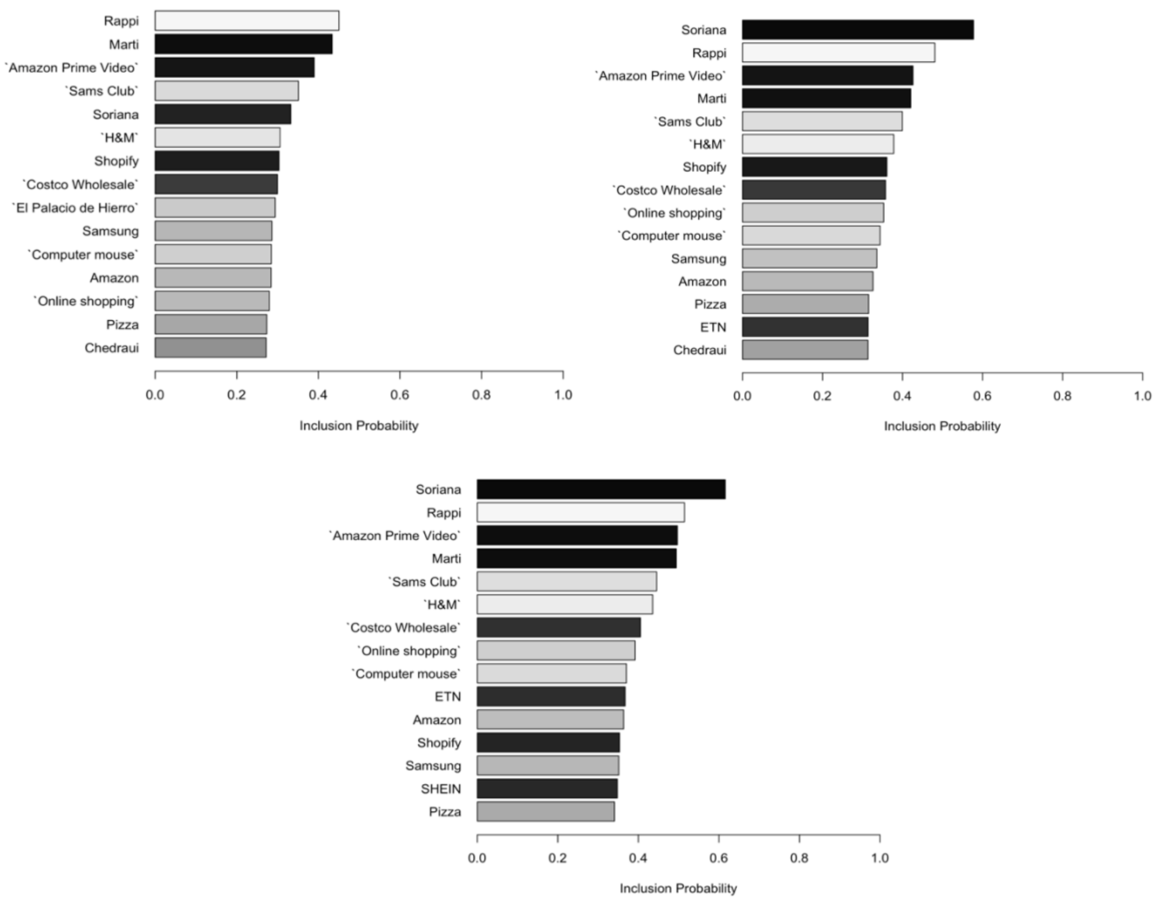


Source: calculated with data from Google Trends and Banxico.

Finally, results from the variable selection are examined to identify the most relevant regressors for explaining ecommerce transactions. Figure 7 presents the marginal posterior inclusion probabilities of the fifteen most relevant regressors in the Bayesian model averaging process. The inclusion probabilities correspond to models with an assigned EMS equal to 20, 30 and 25. Higher probabilities suggest stronger evidence for the relevance of a regressor, while lower probabilities indicate weaker evidence.

It is important to remark that the resulting marginal posterior inclusion probabilities are not high in general. A common criterion for establishing significance among the selected variables is to identify those associated with an inclusion probability above 50%. If this was the case, only 1 out of the 68 variables would be considered as significant. Nevertheless, as prior inclusion probabilities were equally set for all regressors, a more precise manner to proceed is to analyze the parameter estimates with an ordinal perspective to assess the relative importance of the variables and draw conclusions.

Figure 7. Marginal posterior inclusion probabilities based on models with an assigned EMS equal to 20, 23, and 25.



Source: calculated with data from Google Trends and Banxico.

Results can be categorized to identify revealed preferences in the Mexican digital economy. Apparently, grocery stores seem to be of particular importance; other examples prioritized include “Sams Club”, “Chedraui” and “Costco Wholesale”. Thus, it is convenient to identify this set of regressors as a first cluster. The inclusion of other selected variables such as “Rappi”, “Amazon” and “Shopify” suggest there is a second cluster of selected variables that refer two-sided marketplaces. Furthermore, a third cluster of apparel stores is represented by “Marti”, “H&M” and “Palacio de Hierro”. Other variables related to online shopping that do not fit in this clustering exercise refer to electronics, transportation services and food. Table 2 lists the fifteen most relevant variables according to their belonging cluster.

Table 2. List of the 15 Google Trends series mostly related to Ecommerce transactions in México based in results of the BSTS framework.

Google Trends Series		**Cluster**
1	Soriana	Grocery stores
2	Sam's Club	Grocery stores
3	Costco Wholesale	Grocery stores
4	Chedraui	Grocery stores
5	Rappi	Two-sided digital marketplaces
6	Shopify	Two-sided digital marketplaces
7	Amazon	Two-sided digital marketplaces
8	Martí	Brands
9	H&M	Brands
10	Samsung	Brands
11	SHEIN	Brands
12	ETN (bus tickets)	Other
13	Pizza	Other
14	Computer Mouse	Other
15	Online shopping	Other

Source: own formulation with results from estimated models.

Chapter 6

Conclusions

As it was hypothesized, results confirm that the inclusion of Google Trends contemporary information does reflect a lower cumulative absolute error than simple Structural Time Series models for estimating the level of ecommerce transactions in a Bayesian regression framework. Three clusters of relevant Google Trends queries were identified: 1) online marketplaces that belong grocery stores, 2) two-sided marketplaces and 3) online apparel stores. In a broader interpretation, it is possible to assume these results as an illustration of revealed preferences of consumers that engage in online purchasing.

Research from past years show that Google Trends is potentially useful for providing contemporaneous information to keep track of economic, health or other diverse phenomena. This study is an attempt to extend the available literature that incorporates Google Trends data for time series analysis. Seemingly, the results support the statement of Varian that even though web search data can sometimes capture irrational tendencies, GT might still be a powerful resource for research since it provides contemporaneous information that would otherwise be missing.[23]

Nevertheless, it is important to mention that the evidence provided does not confirm that ecommerce transactions can be tracked using solely Google Trends data, or that the results from the feature selection are not subject to uncertainty. To better understand the relationship between Google Trends and ecommerce transactions based on the results of this study, two things must be attended. First, other time series model specifications need to be tested to compare the results from BSTS modelling. Secondly, attempts to combine data from Google Trends and other variables related to ecommerce need to be developed to evaluate the relevance of search queries when the effect of other economic information is considered.

It might even be preferable to develop marketing or economic studies based on Google Trends rather than surveys since models of economic activity that incorporate web search queries might save resources spent on collecting data. For example, evidence provided by

[23] Choi, Hyunyoung, and Hal Varian. "Predicting the Present with Google Trends." *Economic Record* 88 (2012): 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

economists Vosen and Schmidt proved that the predictive power of Google Trends data for estimating private consumption in the U.S. compared to the University of Michigan and the Conference Board indexes of consumer confidence.[24] Jaemin Woo, Ann L. Owen (2017) extended Vosen and Schimdt conclusions with favorable results by using Google Trends for complementing survey-based costumer confidence indexes as well.

However, internet access limitations in other countries, especially developing ones, might not make it possible for Google Trends series to be useful as consumption intent proxies as for the case of Mexico, which is one of the largest developing economies. One should be cautious when trying to extrapolate results from U.S., where a much larger proportion of transactions are made online. More research that pursue to understand the role of Google Trends in explaining ecommerce activity is still needed for the case of other developing countries, both large and small digital economies, to contrast the results among countries and assess the effect of internet penetration. Eventually, with increase of internet penetration and the further digitalization of the economy, Google Trends might turn into a more reliable source for economic studies globally.

[24] Vosen, Simeon, and Torsten Schmidt. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends." *Journal of Forecasting* 30, no. 6 (2011): 565–78. https://doi.org/10.1002/for.1213.

References

Askitas, Nikolaos, and Zimmermann. Klaus F. "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly* 55, no. 2 (2009): 107–20. https://doi.org/10.3790/aeq.55.2.107.

Banbura, Marta, Domenico, Giannone, and Reichlin, Lucrezia. "Nowcasting." *The Oxford Handbook of Economic Forecasting*, 2012, 193–224. https://doi.org/10.1093/oxfordhb/9780195398649.013.0008.

Choi, Hyunyoung, and Varian, Hal. "Predicting the Present with Google Trends." *Economic Record* 88 (2012): 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Da, Zhi, Engelberg, Joseph and Gao, Pengjie. "In Search of Attention." *The Journal of Finance* 66, no. 5 (2011): 1461–99. https://doi.org/10.1111/j.1540-6261.2011.01679.x.

Ginsberg, Jeremy, et al. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457, no. 7232 (2009): 1012–14. https://doi.org/10.1038/nature07634.

Heath, Jonathan. "Lo que indican los indicadores." 2012.

INEGI. Sistema de Cuentas Nacionales de México Fuentes y metodologías Año base 2013.

Jun, Seung-Pyo, Yoo, Hyoung Sun and Choi, San. "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications." *Technological Forecasting and Social Change* 130 (2018): 69–87. https://doi.org/10.1016/j.techfore.2017.11.009.

Marini, Marco, et al. "Using the Google Places API and Google Trends Data to Develop High Frequency Indicators of Economic Activity." *IMF Working Papers* 2021, no. 295 (2021): 1. https://doi.org/10.5089/9781616355432.001.

Narita, Futoshi, and Yin, Rujun. "In Search of Information:" *IMF Working Papers* 18, no. 286 (2018): 1. https://doi.org/10.5089/9781484390177.001.

OECD. "Tracking Activity in Real Time with Google Trends." *OECD Economics Department Working Papers*, 2020. https://doi.org/10.1787/6b9c7518-en.

Preis, Tobias, et al. "Quantifying Trading Behavior in Financial Markets Using Google Trends." *Scientific Reports* 3, no. 1 (2013). https://doi.org/10.1038/srep01684.

Vosen, Simeon, and Torsten Schmidt. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends." *Journal of Forecasting* 30, no. 6 (2011): 565–78. https://doi.org/10.1002/for.1213.

Watts, Robert J., and Alan L. Porter. "Innovation Forecasting." *Technological Forecasting and Social Change* 56, no. 1 (1997): 25–47. https://doi.org/10.1016/s0040-1625(97)00050-4.