WPS7398

Policy Research Working Paper 7398

Nowcasting Prices Using Google Trends

An Application to Central America

Skipper Seabold
Andrea Coppola

WORLD BANK GROUP
Macroeconomics and Fiscal Management Global Practice Group
August 2015

Policy Research Working Paper 7398

Abstract

The objective of this study is to assess the possibility of using Internet search keyword data for forecasting price series in Central America, focusing on Costa Rica, El Salvador, and Honduras. The Internet search data comes from Google Trends. The paper introduces these data and discusses some of the challenges inherent in working with it in the context of developing countries. A new index is introduced for consumer search behavior for these countries using Google Trends data covering a two-week period during a single month. For each country, the study estimates one-step-ahead forecasts for several dozen price series for food and consumer goods categories. The study finds that the addition of the Internet search index improves forecasting over benchmark models in about 20 percent of the series. The paper discusses the reasons for the varied success and potential avenues for future research.

This paper is a product of the Macroeconomics and Fiscal Management Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at jsseabold@gmail.com.

Nowcasting Prices Using Google Trends: An Application to Central America

Skipper Seabold
American University

Andrea Coppola
The World Bank*

JEL codes: E31,C55,C8
Keywords: Macroeconomic modeling and statistics, Inflation, Big Data

*Corresponding author e-mail: jsseabold@gmail.com. This is a preliminary draft.

1 Introduction

It is a well recognized problem that policy makers must make decisions before all data about the current economic environment are available. Given this reality, there is considerable interest in short-term forecasting and nowcasting using intra-period data releases. For example, the forecaster can provide an estimate of GDP this quarter using other data that are available at a monthly frequency. This technique is called nowcasting, or predicting the present. Giannone et al. [2008] lays out three tenets of nowcasting. First, many data series are used. Second, nowcasts are updated as intraperiod data become available. Finally, nowcasting "bridges" higher frequency data releases with the nowcast of the lower frequency series of interest. This study is similar in spirit to that of Giannone et al. [2008]. However, while Giannone et al. [2008] are concerned with nowcasting GDP using a large number of economic data series, this paper nowcasts price series using Internet search keyword data from Google Trends[1]. Furthermore, we do not attempt to "bridge" higher frequency data with lower frequency data explicitly as part of a model. The Google Trends data are not all systematically available at a higher frequency than the series we wish to forecast. Instead, we are more concerned with the efficient aggregation of many series to help improve our nowcasts.

There are three main contributions of this study. First, it focuses on the countries of Central America. Almost the entirety of the nowcasting literature focuses on developed countries with one notable exception in Carrière-Swallow and Labbé [2013]. Second, this is a large scale study, approaching the problem of nowcasting with Google Trends from a data mining perspective rather than one solely grounded in economic theory. This approach gives us insights that will be useful for forecasters who wish to pursue similar ends. Third, we introduce methods from the statistical learning literature to compute the Google Trends keyword search index that are not yet commonly used in forecasting studies.

Given the large number of series included in this study, we rely heavily on automatic model identification procedures. Despite this potential shortcoming, we find that Google Trends can improve our ability to forecast certain series. These findings are notable and may be worth pursuing in more detail. The outline of the paper is as follows. Section 2 reviews some of the literature for nowcasting and the use of Google Trends data in forecasting. Section 3 introduces the data and includes a section that discusses the challenges of working with Google Trends data for the countries of Central America. Section 4 explains the framework used for forecasting and evaluating forecasts. Section 5

[1] http://google.com/trends

discusses the results of this exercise and assesses the usefulness of Google Trends data in forecasting price series for Central American countries. Section 6 concludes, noting several paths for continuing research. While this section deals specifically with ideas for future research there are notes about ongoing research throughout the paper.

2 Literature Review

There is a growing literature that is using Internet search keyword data and Google Trends, in particular, for forecasting and nowcasting. Ettredge et al. [2005] were the first to use search engine keyword data to aid in forecasting. They found keyword-based searches to be helpful in predicting the number of unemployed workers in the United States. The use of Google Trends data, specifically, in forecasting yet to be released macroeconomic series goes back to Choi and Varian [2009, 2012]. They find that Google Trends data help to forecast initial unemployment claims, automobile sales, and consumer confidences in the United States. Since then, there are been numerous efforts to use Google Trends data in forecasting. Schmidt and Vosen [2012] uses search data related to the "cash for clunkers" program to improve forecasts for private consumption in France, Germany, Italy, and the United States. Guzman [2011] uses Google search data to estimate inflation expectations. Suhoy [2009] estimates accurate probabilities of downturn in early 2007 using Google search category data for Israel. The author also finds improvements in estimates of private consumption by employing the search data.

Early results on using Google Trends data as a proxy for consumer sentiment are promising. Traditionally, studies have made use of survey-based sentiment data to provide leading indicators of series of interest. However, this data is not always available, especially in developing countries. Vosen and Schmidt [2011] show that Google Trends outperforms The University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index in predicting private consumption in the United States. One study which is very relevant to our present effort is that of Carrière-Swallow and Labbé [2013]. The authors look at the benefits of using Google Trends data in the context of a developing country, Chile. They develop an index of consumer interest in automobile purchaes and find that it outperforms benchmark specifications that take advantage of the IMACEC index of consumer activity. We will use a similar framework to the one employed in that study in what follows.

3 Data

This section first describes the raw data and then the transformations that are made to each series before estimation. A subsection is dedicated to addressing some of the challenges inherent in working with the search query data from Google in emerging market countries. For each of Costa Rica, El Salvador, and Honduras[2], there are two categories of series that we will forecast – data on aggregate consumer prices and their component series and staple food price data.

We obtained the consumer price data from the statistical office of each country. See Appendix A for details. The raw series are in levels and are not seasonally adjusted.

The food price data was obtained from the Global Information and Early Warning System on Food and Agriculture (GIEWS) from the Food and Agriculture Organization of the United Nations (FAO). The types of food that are available from GIEWS are particular to each country. We obtained every available series. Appendix A gives, for each country, the series names, appropriate region, and the units for which we have data available. These series are not seasonally adjusted.

To augment our forecasts, we have obtained Google Trends data on a number of search keywords. These keywords were chosen *ex ante* with the belief that they contain relevant information that will allow us to use them as a proxy for consumer behavior and beliefs. Obtaining real-time insights into consumer behavior allows us to better predict price changes all other things equal. In some sense, the Trends data takes the place of traditional consumer-sentiment surveys. The keywords that we have chosen are listed in Table 1.

Each individual Google Trend series is relative and not an absolute measures of search volume. That is, the period in which the search interest of a keyword is highest within the dates of inquiry receives a value of 100. All other periods for an individual series are measured relative to this highest period. There is, therefore, no sense of how many people were searching for a term and the terms themselves are not comparable with each other. Furthermore, changes in Internet penetration and the use of Google, in particular, do not matter.

The following transformation is made to each price series before estimation to go from

[2]We could not acquire sufficient data on food prices or on search keywords for Belize, so it is omitted from discussion. Earlier versions of this paper contained every other country in Central America. However, given some of the data challenges discussed below, we chose to narrow our interest to three countries. We chose Costa Rica and El Salvador because they generally have good data availability from Google Trends. The quality of the data for Honduras, on the other hand, was found to be rather poor, so we included it to learn more about how the models perform under adverse data conditions.

Search Keywords

	cr	hn	sv
arroz	x	x	x
azucar	x	x	x
carne	x	x	x
caro	x	x	x
cerdo	x	x	x
combustible	x	x	x
cuesta	x	x	x
diesel -vin	x	x	x
frijoles	x	x	x
gas	x	x	x
gasolina	x	x	x
inflacion	x	x	x
ingresos	x	x	x
maiz	x	x	x
pago	x	x	x
pan	x	x	x
precio	x	x	x
precios	x	x	x
propano	x		x
salario	x	x	x
sueldo	x	x	x
trigo	x	x	x

Table 1: The keywords that are used in the forecasting. We found that the search term "diesel -vin" was more reliable in returning searches related to diesel fuel rather than the actor Vin Diesel. All analysis is based on this term.

levels to month-over-month percentage changes

$$x_t = \frac{p_t - p_{t-1}}{p_{t-1}} \times 100 \tag{1}$$

No series has been seasonally adjusted prior to downloading. Therefore, many of the series exhibit some degree of and sometimes a strong degree of seasonality. As discussed below, we will attempt to model the seasonality explicitly when present.

A few of the GIEWS price data series contain missing observations. The missing observations were replaced using simple linear interpolation before applying this transformation.

The Google Trends data are transformed as follows. Some of the search terms are available at weekly frequencies while other series are only available at monthly frequencies. For those that are available at a weekly frequency, we take the maximum value in each month to be the value for that month. This differs from the approach of Vosen and Schmidt [2011] and Carrière-Swallow and Labbé [2013] who aggregate the weekly data into monthly series by taking the monthly average of the indicators. Since the data are relative, we do not wish to first smooth them in this way. This could mask potentially important, short-lived events. Further transformations to the Trends data are described in the next subsection.

3.1 Challenges in Using Google Trends Data

Several challenges present themselves when working with the Google Trends data in a developing country context. First, as pointed out by Carrière-Swallow and Labbé [2013], Google Trends historical data are not constant over time. Within the same 24-hour period, the results will be the same. However, from day to day the results can be different. Indeed, not only do the values change, but on one day monthly data may be returned. On another biweekly or weekly data for the same keyword search. It is unclear, what exactly is driving these differences – whether different normalizations, sampling considerations, or something else, but for practical purposes we can treat the data as being recorded with sampling error with the same consequences. For the present study, we collected data on all of the keywords for ten days over a period of one month. Figures 1 and 2 show the sampling error for two representative series collected during this period.

These series are chosen to be representative of all of the series used and show the two most salient features for the purposes of this study. First, the sampling error is evident

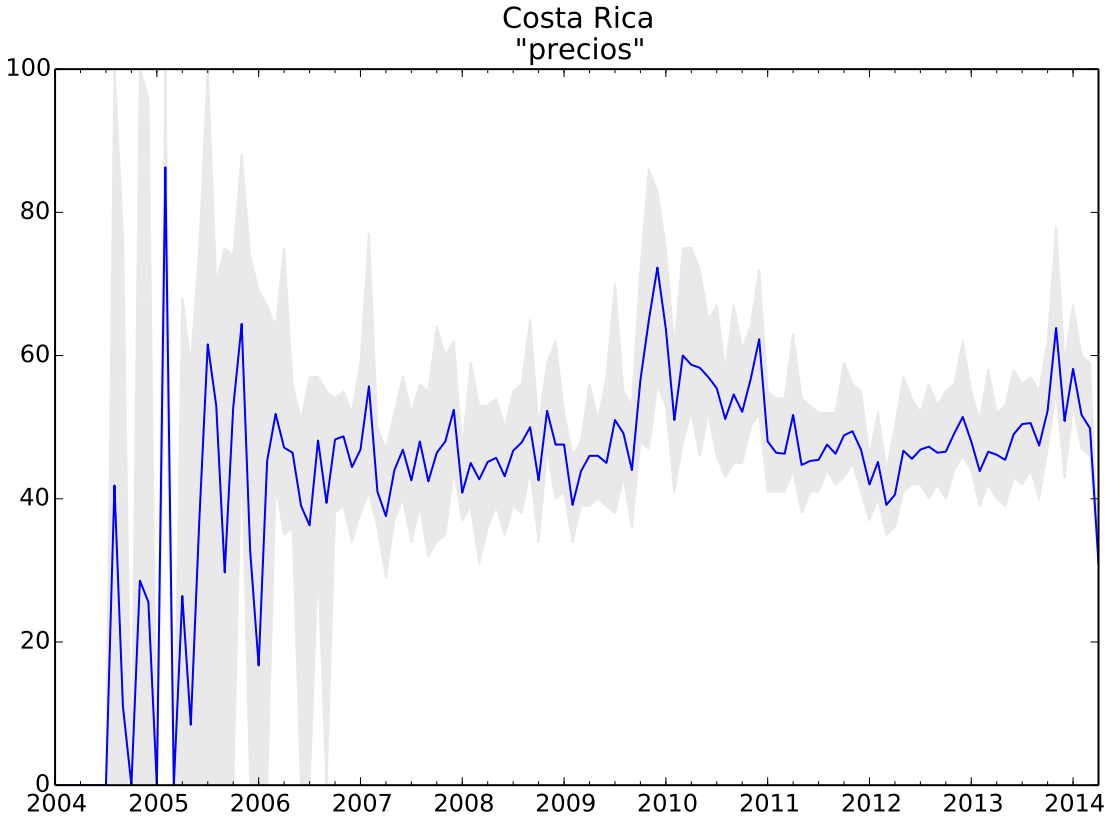

Figure 1: Results for 10 days during the study period for the "precios" keyword in Costa Rica. The dark line is the average. The gray bands are minimum and maximum observed values for that month over the study period.

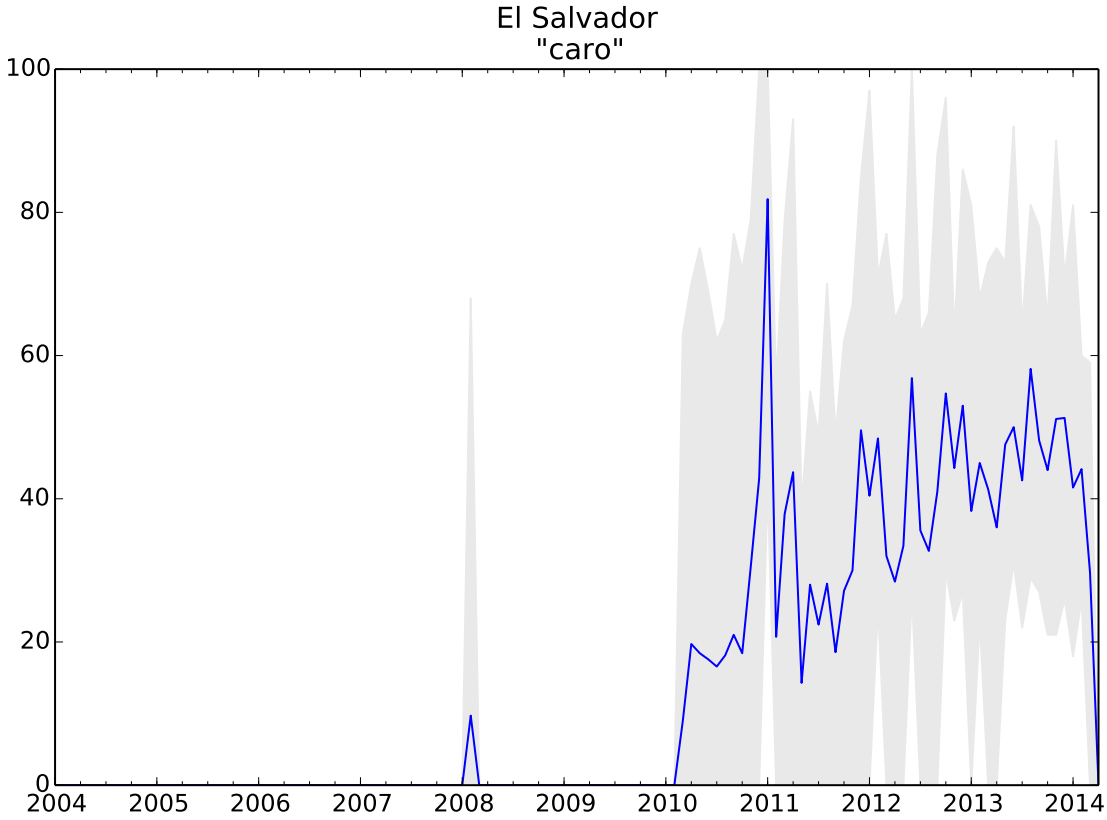

Figure 2: Results for 10 days during the study period for the "caro" keyword in El Salvador. The dark line is the average. The gray bands are minimum and maximum observed values for that month over the study period.

in both figures. Figure 1 is in some sense a best case scenario. Variability is very large for the first two years of the sample but becomes quite a bit more stable after this initial uncertainty. Figure 2, on the other hand, shows high sampling variability throughout the entire period. We will assume that the signal of each series can be well approximated by its average and use the average when referring to the series for a single keyword in what follows unless otherwise indicated.

The second thing to note in figures 1 and 2 are that many of the observations for a single draw of the Google Trends data are exactly zero. These zero observations present two difficulties in particular – one conceptual and one practical. First, conceptually, these zeros suggest a lack of signal where presumably there should be some. As we collect more daily samples of the data, this problem becomes less and less, again assuming that the signal is well approximated by the mean. However, this problem does not disappear. Looking at the early parts of both series, there are still observations which are zero even at the mean.

Second, as a practical problem, some of the Google Trends data contain strong seasonal components. Studies such as Carrière-Swallow and Labbé [2013] alleviate the effects of seasonality in the trends data by using year-over-year percent changes for them as well as the series to forecast. However, if the base year is zero, we would lose this entire year of data.

We employed several techniques in an attempt to overcome these problems, which we will now describe. The Google trends data can be written more formally as $X_{i,j,t}$ where i represents the vintage – a downloaded sample on a particular day, j represents a particular keyword, and t represents the weekly, bi-weekly, or monthly observation of each keyword. The first task is to deal with the i vintage, or sample, index. We took the mean and the median of all the samples. This leaves us with either

$$X_{j,t} = \frac{1}{I} \sum_{i} X_{i,j,t}$$

for the mean, where I is the total number of samples taken or

$$X_{j,t} = \text{med}_i(X_{i,j,t})$$

for the median.

After handling the sampling dimension, we apply transformations to smooth the data for each keyword and attempt to better identify the signal from the noise, given the

nature of the search data. Here, we take several different approaches. First, we apply a simple exponential smoothing model with additive errors to the data. Following the notation of Makridakis et al. as used in Hyndman et al. [2002], this model can be written

$$l_t = \alpha y_t + (1 - \alpha) l_{t-1} \tag{2}$$

We choose to fix $\alpha = .5$. Results typical of this smoothing can be seen in figures 3 and 4. We include both the forecastable part of the series and the unsystematic "surprise" part of the series.

We also tried smoothing the results by applying the Christiano-Fitzgerald (CF) band-pass filter [Christiano and Fitzgerald, 2003]. The CF filter starts from the (false) assumption that the underlying data obeys a unit root process. Using this assumption, the CF filter provides an approximation to an optimal band-pass filter as follows

$$\begin{aligned} \hat{c}_t = & B_0 y_t + B_1 y_{t+1} + \cdots + B_{T-1-t} y_{T-1} + \tilde{B}_{T-t} y_T + \\ & + B_1 y_{t-1} + \cdots + B_{t-2} + \tilde{B}_{t-1} y_1 \end{aligned} \tag{3}$$

where $B_j = \frac{\sin(jb) - \sin(ja)}{\pi j}, j \geq 1$ and $B_0 = \frac{b-a}{\pi}$, $a = \frac{2\pi}{p_u}$, $b = \frac{2\pi}{p_l}$, $\tilde{B}_k = -frac12B_0 - \sum_j = 1^{k-1} B_j$. The parameters p_u and p_l denote the cut-offs for the cycles for the high and low frequency elements, respectively. We remove all stochastic cycles at a periodicity lower than 3 months and higher than 12 months. This has the effect of both smoothing the series and removing long-term seasonality. The results of applying the CF filter to our two selected series can be seen in figures 5 and 6.

One notable advantage of techniques such as exponential smoothing and the CF filter is that they provide us with real-time estimates at the ends of our series so that we do not need to truncate our observed series at the beginning or the end as would be necessary if we used a simple moving averages, seasonal differences, or, another filter such as the Baxter-King.[3]

[3]Of course, we could estimate a model and forecast and backcast then apply a filter that truncates, using these extra data points. However, this is another form of uncertainty that we would like to avoid introducing. Instead, we prefer to use only the information we have.

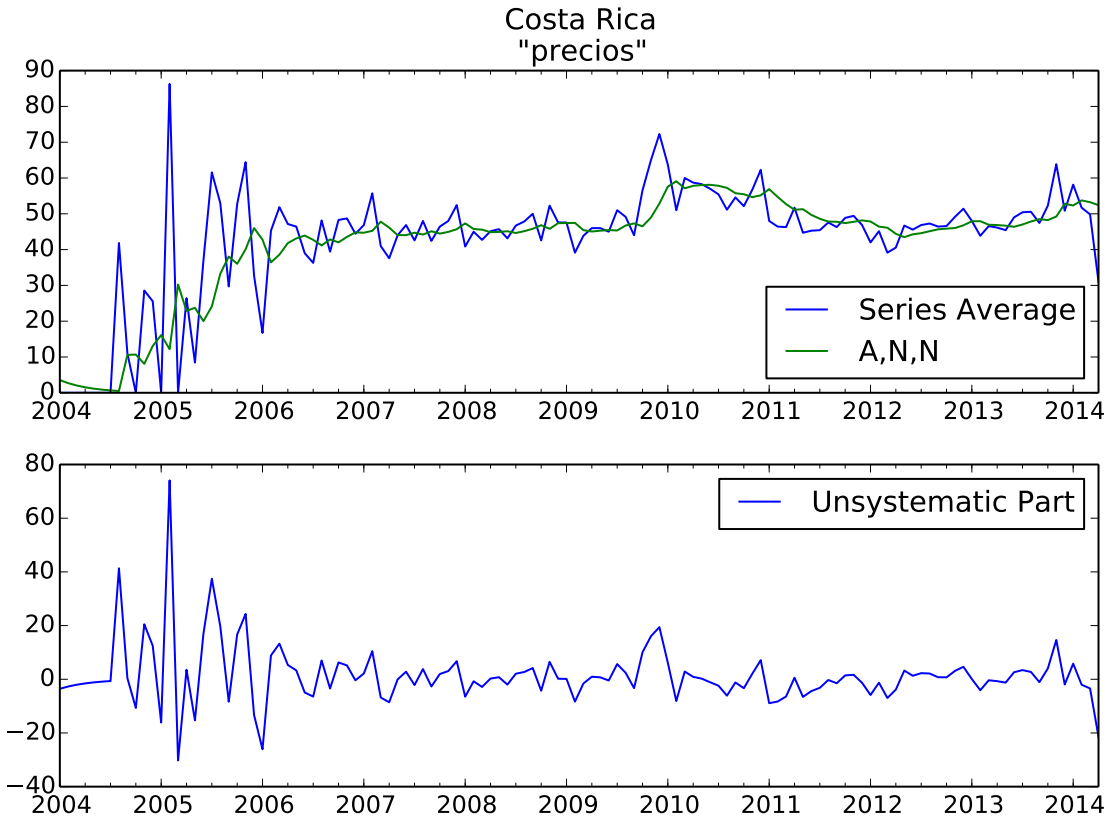

Figure 3: Smoothed results for the average of the "precios" keyword in Costa Rica. The top pane contains the original series and the smoothed, in-sample forecasted series. The forecasted series is labeled A, N, N indicating additive errors, no trend, and no seasonality according to the Hyndman et al. [2002] taxonomy. The bottom pane contains the unsystematic or "surprise" component of the series.

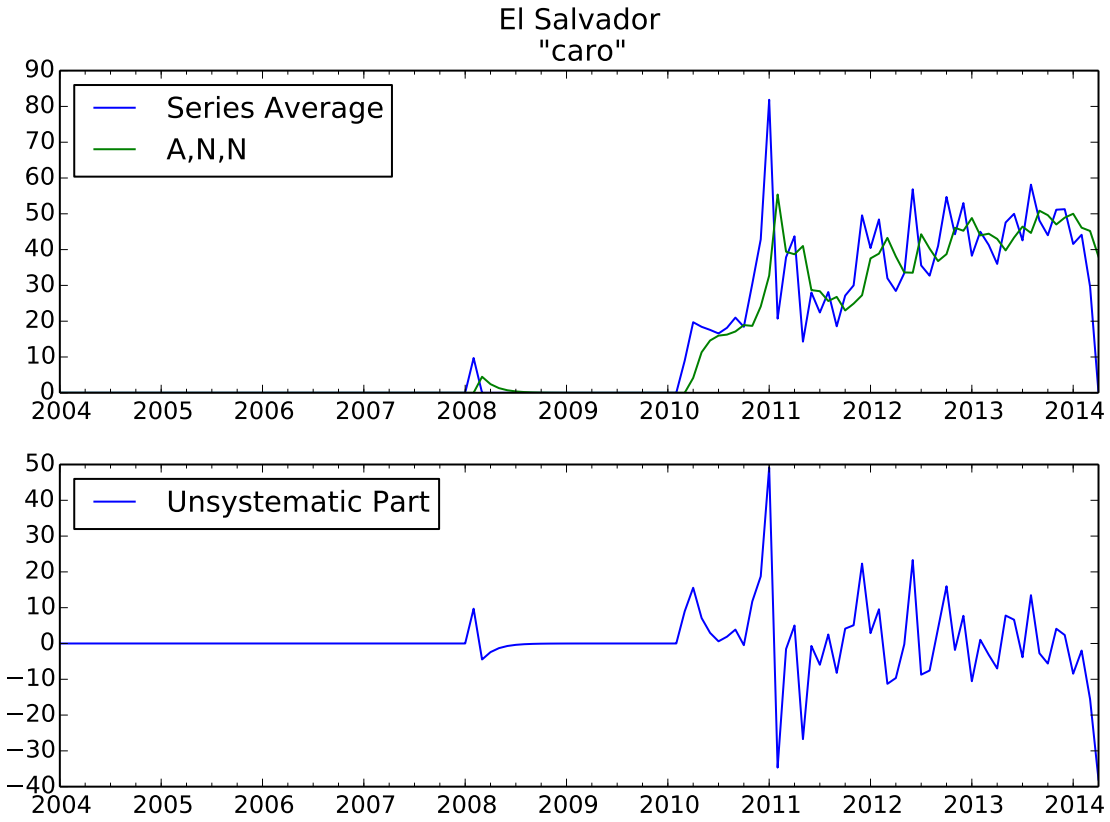

Figure 4: Smoothed results for the average of the "caro" keyword in El Salvador. The top pane contains the original series and the smoothed, in-sample forecasted series. The forecasted series is labeled A, N, N indicating additive errors, no trend, and no seasonality according to the Hyndman et al. [2002] taxonomy. The bottom pane contains the unsystematic of "surprise" component of the series.

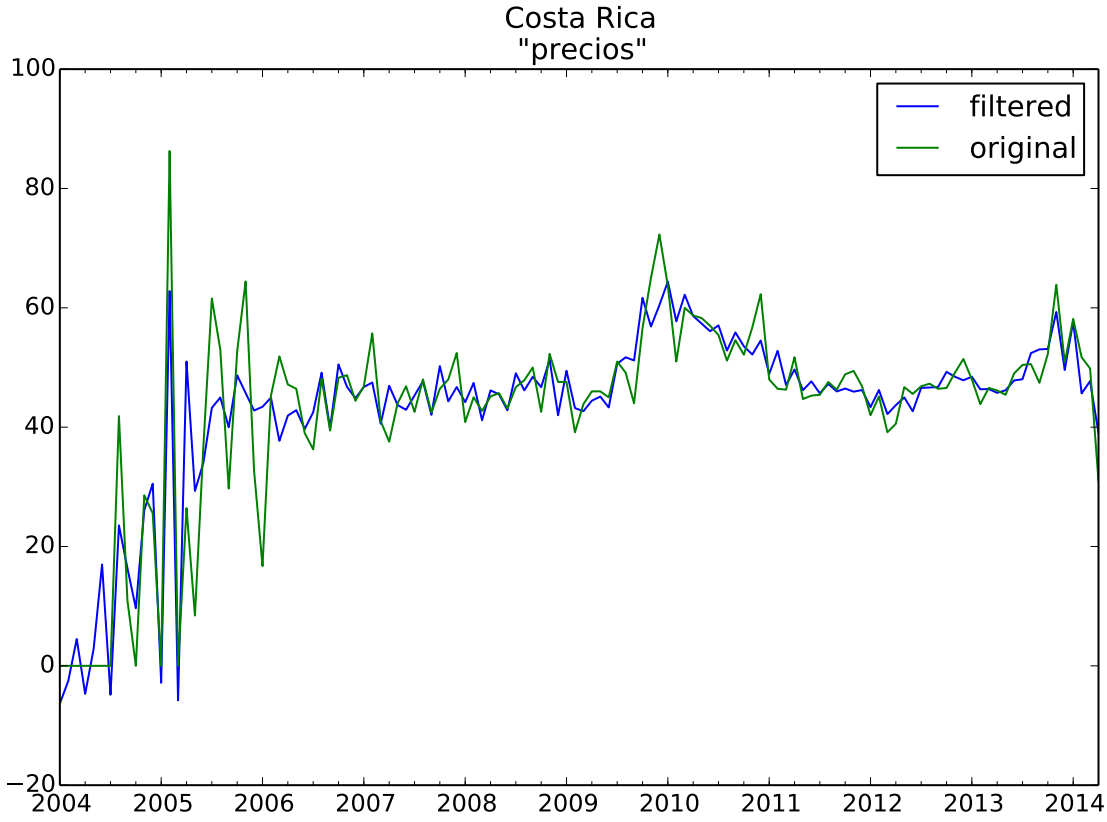

Figure 5: Smoothed results for the average of the "precios" keyword in Costa Rica. The smoothed series is computed using the Christiano-Fitzgerald filter with all stochastic cycles at a periodicity lower than 3 months and higher than 12 months removed.

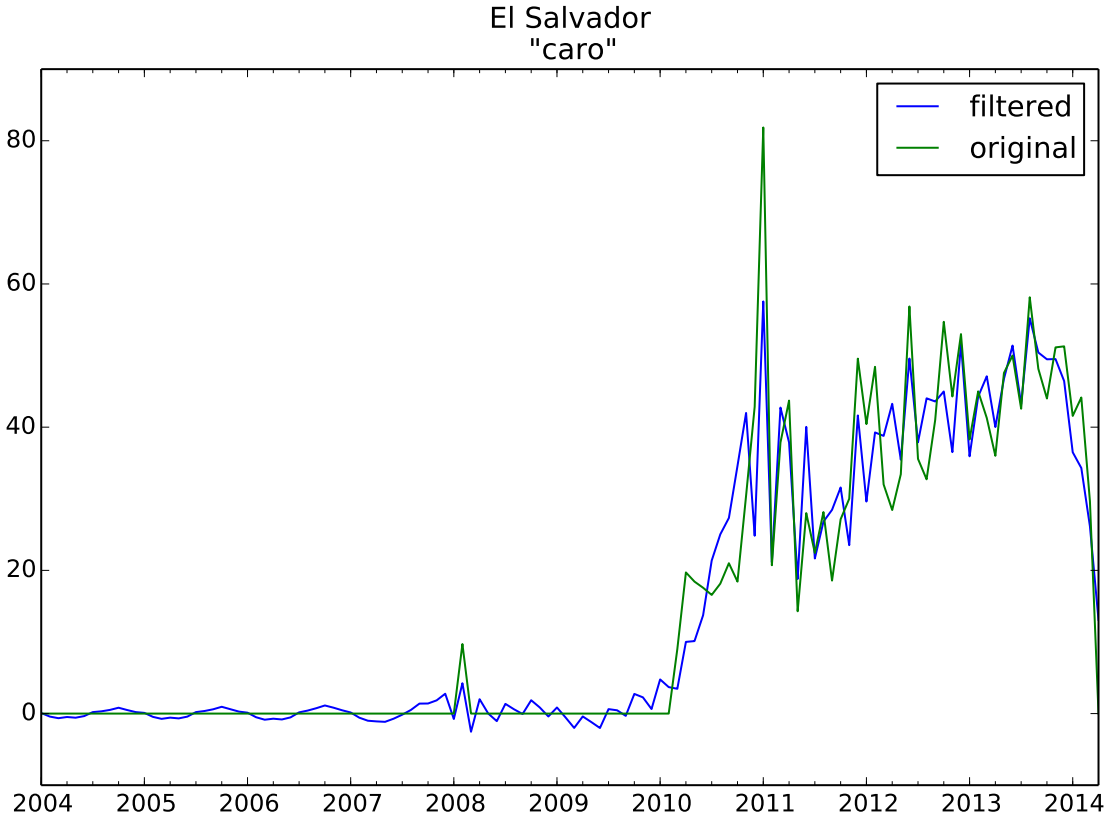

Figure 6: Smoothed results for the average of the "caro" keyword in El Salvador. The smoothed series is computed using the Christiano-Fitzgerald filter with all stochastic cycles at a periodicity lower than 3 months and higher than 12 months removed.

4 Methodology

To nowcast a series at a particular point in time, we produce an estimate of the series before that variable has been observed but when other contemporaneous variables in our information set have been observed. For instance, we might use data available to us now to get an estimate for economic growth or inflation before official statistics are released. As a concrete example, suppose that in mid-April 2014, we have either a few weeks of Google Trends data or perhaps some preliminary monthly estimate of a search term, but we do not yet know the current inflation. Lags in publication of inflation could mean that we only have estimates for inflation through March or even February 2013. If a policymaker is interested in knowing inflation today, we would nowcast at a monthly m horizon of $h_m \geq 1$.

Our strategy for this exercise is as follows. For each series in each country we will compare nowcasts using Google Trends data and one-step ahead forecasts from a best effort ARIMA model to some benchmark models to assess if the information available from Google Trends data improves our forecasting ability. We now introduce our benchmark models. In the following subsection, we discuss what we mean by a "best effort" ARIMA model.

4.1 Benchmark Models

Five simple models are estimated to provide a baseline for the candidate models described below. The estimated baseline models are the simple mean of the series, the median of the series, the value of the series in the previous period, an AR(1) model, and an A, A, N exponential smoothing model. This exponential smoothing model written in its recursive form is given by

$$\begin{aligned} l_t &= \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \end{aligned} \tag{4}$$

where l_t and b_t are the level and growth rate, respectively, and the parameters along with the initial states are estimated as described in section 3.1. This model is otherwise known as Holt's linear method with additive errors and is equivalent to an ARIMA $(0, 1, 1)$ model [Hyndman et al., 2008]. Our one-step ahead point forecasts are given by

Benchmark Results

Benchmark Model	Total
ar	23
ets	4
mean	11
median	20

Table 2: The total number of series for which each benchmark model is deemed the best by the MSE criterion.

$$\hat{y}_{t+1} = \frac{1}{t}\sum_{i=1}^{t} y_i \tag{5a}$$

$$\hat{y}_{t+1} = \text{median}(\{y_i\}) \forall i = 1, \ldots, t \tag{5b}$$

$$\hat{y}_{t+1} = y_t \tag{5c}$$

$$\hat{y}_{t+1} = \hat{\rho} y_t + \epsilon_t \tag{5d}$$

$$\hat{y}_{t+1} = l_t + b_t \tag{5e}$$

where $\epsilon_t \sim N(0, \sigma)$ in (5d). We choose the baseline model for each series based on mean squared error (MSE). MSE is defined as usual

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t - Y_t)^2$$

where $\hat{Y}_t$ is our forecast estimate, Y_t is the true observation at time t, and T is the total number of observations.

To compute the MSE for the benchmarks, we start with two years of data and compute one-step ahead forecasts using each benchmark model until time $T-1$ where T is the last period for which we have data that we wish to forecast. We then choose the model that has the best performance in all periods as the benchmark model for that series. Table 2 presents an overview of which benchmark model is best in an MSE sense. The AR(1) and median are preferred the most often. The benchmark for the individual series is presented with the full results in section section 5 for ease of comparison.

4.2 Forecasting and Nowcasting Models

To attempt to improve over these baseline models, we first estimate a possibly seasonal Autoregressive integrated moving-average (ARIMA) model for each monthly series.

$$\varphi(L)^p\varphi_{12}(L^{12})^P(1-L)^d(1-L^{12})^D(y_t-\mu)=\theta(L)^q\theta_{12}(L^{12})^Q\epsilon_t \tag{6}$$

where y_t is the series we wish to forecast, ϵ_t follows a white noise process, L is the lag operator $L^i y_t = y_{t-i}$, $\varphi\left(L\right)^p = (1-\phi_1 L-\cdots-\phi_p L^p)$ is the non-seasonal polynomial of order p in the lag operator that describes the autoregressive component of the model and $\varphi_{12}\left(L^{12}\right)^P = \left(1-\phi_{12,1}L-\cdots-\phi_{12,P}L^P\right)$ is the seasonal polynomial of order P in the lag operator that describes the seasonal autoregressive component of the model. The polynomial of order q that denotes the non-seasonal MA component of the model is $\theta(L)^q$, and likewise the seasonal MA component of order Q is denoted $\theta(L^{12})^Q$. The non-seasonal and seasonal orders of differencing are denoted d and D, respectively.

We use the auto.arima function from the forecast package in R[4] for order identification for each series. See Hyndman and Khandakar [2008] for more information on the model identification procedure.[5] The auto.arima automatic model identification procedures allows parameters to be zero, so in principal, for example, the model is only differenced or includes a seasonal component when it is appropriate.

To test whether there is information in the Google Trends data that will help us forecast each series, we use a possibly seasonal ARIMAX model where the Trends data is used as an exogenous variable.[6]

The seasonal ARIMAX model estimated is specified

[4]We used the 5.4 develpment version obtained from https://github.com/robjhyndman/forecast/

[5]We also performed order identification using the AUTOMDL procedure from X-13ARIMA-SEATS [Staff, 2013] as well as using (seasonal) unit root tests to identify the order of (seasonal) differencing and then using the Bayesian information criteria (BIC) to select the best model. None of the procedures used produced identical results, nor did any procedure do unambiguously better than any other. The auto.arima function was the most computationally performant and is thus the basis for the results below. We used the default arguments for this function.

[6]This model is sometimes referred to as a regression model with ARMA errors. Ignoring seasonality, it may be written

$$\begin{aligned} y_t &= \beta' \boldsymbol{x}_t + z_t \\ \varphi(L) z_t &= \theta(L) z_t \end{aligned} \tag{7}$$

This is to contrast it with the ARMAX model which is written

$$\varphi(L)^p(y_t) = X_t\beta + \theta(L)^q\epsilon_t \tag{8}$$

$$\varphi(L)^p\varphi_{12}(L^{12})^P(1-L)^d(1-L^{12})^D(y_t-\beta' x_t)=\theta(L)^q\theta_{12}(L^{12})^Q\epsilon_t \tag{9}$$

where everything is as in (6) and x_t contains the Google Trends Index that we describe in the next section. The addition of this term allows us to model the information contained in the Google Trends data as a time-varying mean.

4.3 Index Construction

In order to incorporate the information from the various Google Trends search keywords, it is desirable to synthesize the information in all of the Google Trends data into something more manageable. Formerly, authors used the Google Insights search categories data. This data is used in many of the studies referenced in section 2. However, previously such an index from Google Insights was usually not available outside of large, developed countries, so studies such as Carrière-Swallow and Labbé [2013] estimate their own. The advantage of having an index is mainly parsimony of information. Indeed, such an index may be of interest in its own right. Furthermore, in September 2012 Google merged some features of Google Insights with Trends and discontinued the aggregate search categories entirely.[7]

To solve the keyword aggregation problem Carrière-Swallow and Labbé [2013] creates an index from multiple search terms by use of an expanding linear regression model described below. Other approaches rely on factor analysis techniques for dimension reduction such as unweighted least squares [Vosen and Schmidt, 2011] or principal components analysis [Stock and Watson, 2002]. These methods assume that there are some underlying, unobserved common factors for all of the series. We describe our use of statistical learning techniques for variable selection below.

We took several approaches to constructing our search indices. First, we applied the linear index approach of Carrière-Swallow and Labbé [2013]. This is a common approach in the literature and is an attractive choice mainly for its simplicity. Let $\boldsymbol{X}$ be our matrix of year-over-year percent changes for the Google Trends terms. We construct an index I_t for these terms, for each series y_t that we wish to forecast in the following way. In each period, we estimate the weights $\hat{\beta}$ by using the observations up to time $t-1$ and fitting a linear model

[7] http://insidesearch.blogspot.com/2012/09/insights-into-what-world-is-searching.html One may only speculate that it was discontinued because this task is very difficult to automate.

$$y_t = \alpha + \beta X_t + \epsilon_t$$

The index for period t is

$$I_t = \hat{\mathrm{E}}[\beta \mid y_{t-1}, X_{t-1}]X_t$$

Given that y and $\boldsymbol{X}$ contain monthly percent changes, we can interpret I_t as the linear combination of search terms which best explains the series that we are forecasting, in a linear least squares sense. The expanding nature of the construction of the index allows for the factors in the trends that explain the changes in our price series to change over time. This is certainly something we might be interested in given the heterogenous character of the included terms. Figure 7 contains an example of an index created using the expanding linear OLS. That is, this is the last out-of-sample fitted value of each index created for a single price series.

We anticipate two potential problems with this approach for the current exercise, and we construct this linear index using two other methods from the statistical learning literature. Both of these techniques were implemented using the scikit-learn Python package [Pedregosa et al., 2011]. First, we have a high number of variables relative to the number of observations, especially in the early years of the index. To improve the degrees of freedom of our fit, we are interested in obtaining sparse models. To this end, we applied the lasso technique introduced in Tibshirani [1996].[8] The lasso is a penalized least squares method that allows both continuous shrinkage and variable selection through the imposition of an L_1-penalty on the regression coefficients β. That is, the coefficients are pushed both towards and to zero when appropriate. The optimization function for the lasso is

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2 + \alpha\|\beta\|_1$$

where α is chosen via $K-$folds cross-validation with $K = 5$ and the L_P-norm is defined $\|X\|_p = \sum_{i=1}^{n}(|x_i|^p)^{1/p}$. Figure 8 contains an example of an index created using the expanding lasso linear model. The fit is much more conservative than the linear OLS fit given the sparse nature of the solution.

[8] We also considered the more general LARS estimator introduced by Efron et al. [2004]. The results of this estimator were comparable, though slightly worse than the lasso. It should also be noted that we ran the computationally efficient LARS algorithm variant for the lasso solution path.

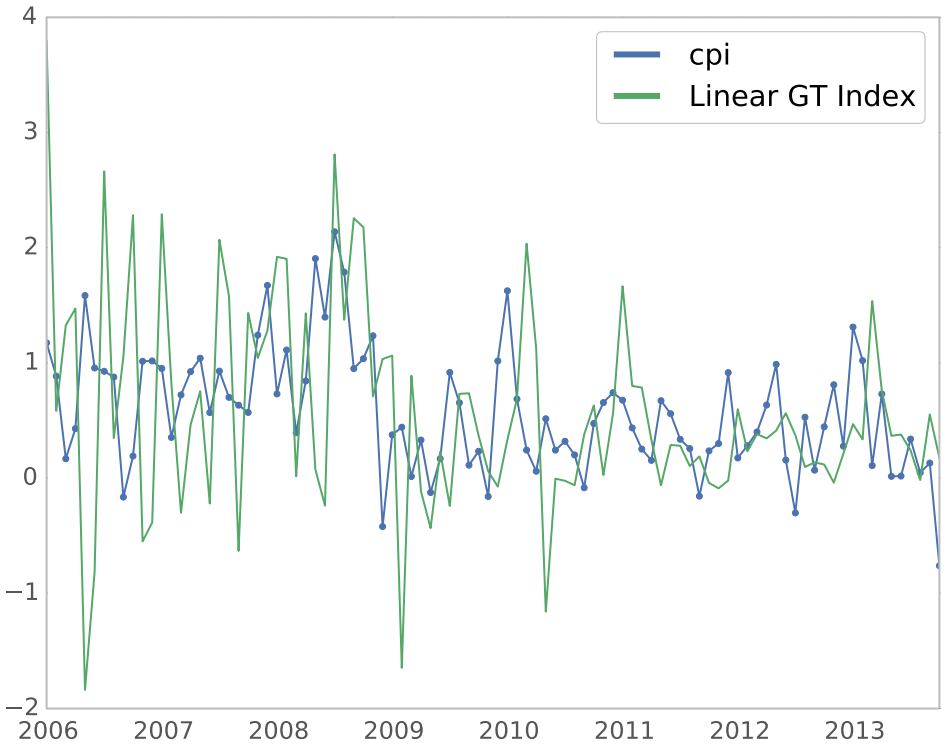

Figure 7: Linear OLS index created for CPI series from Costa Rica. Displays some evidence of overfitting.

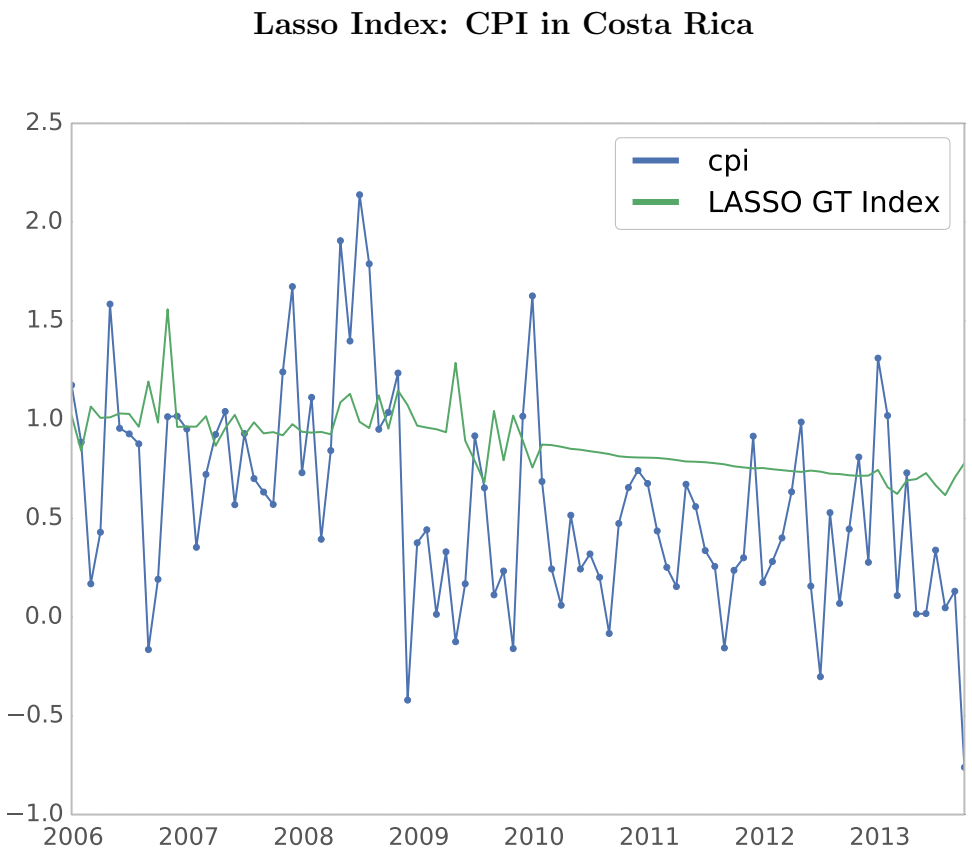

Figure 8: Lasso model index created for CPI series from Costa Rica. Conservative fit. Does not vary much.

Both Zou and Hastie [2005] and Tibshirani [1996] point out that the lasso may not perform well empirically in the cases where the number of variables is higher than the number of observations[9], there are groups of variables with high pairwise correlation, or there are high correlations between all predictors. These are all possible concerns for our keywords from Google Trends. To account for these issues, we employ the elastic net estimator of Zou and Hastie [2005].

The elastic net estimator is a linear combination of the L_1-penalty of the lasso and the L_2-penalty of ridge regression [Hoerl and Kennard, 1970]. The objective function of the elastic net is

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2 + \alpha\rho\|\beta\|_1 + \frac{1}{2}\alpha(1-\rho)\|\beta\|_2$$

where α and ρ are chosen via K-folds cross-validation with $K = 5$. Using both the lasso and the elastic net, we compute the index in the same way as the linear OLS index except that the β coefficients are obtained from the two new estimators. Figure 9 contains an example of an index created using the expanding elastic net model. The fit is somewhere between the high variance OLS model and the conservative lasso model. In the following section we describe the empirical results of using these indices.

5 Forecasting Results

Our hypothesis is that there is additional information in our transformations of the Google Trends data that allows improved nowcasts of the series of interest before their respective data releases have been made versus an ARMA model and our respective benchmarks. To test this hypothesis we compute one-step ahead forecasts using (6) and (9) and compare them to the chosen models from (5). Just like for the benchmarks, we start with two years of monthly data and then estimating expanding window models until time $T-1$ where T is the last period for which we have data for the series we wish to forecast and for which we have T Google Trends index values. At each time t in $t = 24 \ldots T+1$ we recompute the order of the seasonal ARIMA(X) model as described above. This is to emulate what a practitioner would do in any given period. For each forecast (and nowcast) we compute the one-step ahead forecast error

[9]This is not the case in the current analysis, though we do have the case where the number of variables is only slightly smaller than the number of observations in the early periods of our index construction

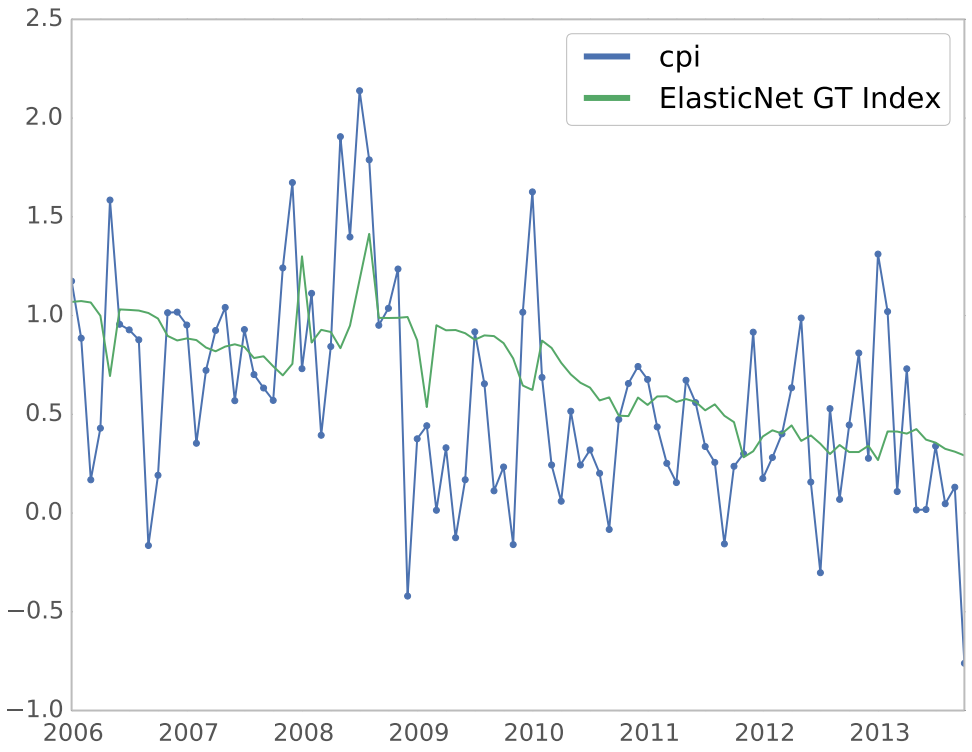

Figure 9: Elastic net model index created for CPI series from Costa Rica. Somewhere in between high variance OLS and low variance lasso.

$$\hat{e}_{k,t+1} \equiv y_{k,t+1} - \mathrm{E}_t[\hat{y}_{k,t+1}] \tag{10}$$

for model k. We compute the relative MSE for each series combination method defined in section 4.3. That is, for the original data $X_{i,j,t}$ we computed the results for each of reduction methods of the i sampling dimension – mean, median, applying the CF filter after taking the mean, and ETS smoothing after taking the mean – and for each i reduction we also computed the three linear indices over the j keywords – linear OLS, lasso, and elastic net. We found first that the linear OLS trend preformed unambiguously the worst. We were we unable to beat both the benchmark model and the ARMA model even once regardless of the smoothing technique that we applied. This is not wholly suprising given that we did not apply any variable selection of the keywords beforehand. Inclusion of inappropriate keywords appears to have led to overfitting and poor out of sample performance.

Moving to the lasso and the elastic net, for each respective estimator the ETS smoothed results performed best. Between the classifiers, the elastic net performed marginally better, beating both the benchmarks and the ARMA models in a few more cases. Again, this is not wholly surprising given the documented better empirical performance of the elastic net estimator when there is high pairwise correlations among the regressors. Due to the large number of results, Table 3 contains only the results of the ETS smoothed data and the trend computed via the elastic net estimator.

Using our best performing method, the ARMAX model outperforms the benchmark in 28% of the cases or for 16 out of the 58 series. In each of these cases, the ARMAX also outperforms the ARIMA model. The ARMA model fairs only slightly better versus the benchmark, outperforming the benchmark in 22% of the series or for 13 out of 58 series. However, the ARIMA model is only the best model versus the ARIMAX model in 7 of these cases. The food price series appear to be particularly difficult to forecast. If we consider only the consumer price series, then the ARMAX model is the best model in 24% of the cases while the ARMA model is only the best in 14% of the cases. The difficulty in forecasting food prices is likely due to the food price crisis during the period. The price US dollar price fluctuations during the time under consideration were due largely to events external to the countries of Central America.

These results, though only partially successful, indicate that there may be some benefit to exploring the further use of Google Trends data in forecasting economic series in Central America. We use the concluding section to speculate on some of the reasons for this success or lack thereof and give suggestions for future research.

Table 3: The results for each series using the ETS smoothed data and the elastic net estimator. Relative MSE (1) is the MSE for the expanding window ARMA model results versus the benchmark model given in columns 6. Relative MSE (2) is the MSE for the elastic net model versus the benchmark in column 6. A relative MSE less than 1 indicates that the proposed model beat the benchmark.

Country	Series	Relative MSE (1)	Relative MSE (2)	N	Benchmark
cr	food01	1.04644	1.09103	94	ar
cr	food02	1.00313	1.09517	94	median
cr	food03	1.08606	0.977452	94	mean
cr	food04	1.0755	1.06406	94	median
cr	food05	1.32068	1.32068	38	median
cr	food06	1.01586	1.04028	94	ar
cr	food07	1.08315	1.1684	94	mean
cr	food08	1.08143	1.17251	94	mean
cr	food09	0.886634	1.20473	57	median
cr	infl01	0.980986	1.09914	94	ar
cr	infl02	1	1.00755	63	mean
cr	infl03	0.993006	2.13919	63	median
cr	infl04	1.08064	1.08917	63	mean
cr	infl05	0.981035	1.11166	95	ets
cr	infl06	0.840378	0.808711	63	ar
cr	infl07	1.03188	1.00261	63	ar
cr	infl08	1.03867	0.860798	63	ar
cr	infl09	1.0767	1.01002	63	median
cr	infl10	1.11333	1.23137	63	ets
cr	infl11	1.10306	1.15871	63	median
cr	infl12	1.17467	1.18608	63	ets
cr	infl13	1.14892	1.35561	63	ets
cr	infl14	1.0491	1.04081	63	ar
hn	food01	1.11477	1.23652	94	median
hn	food02	1.06082	1.02454	94	mean

Continued on next page

Table 3: The results for each series using the ETS smoothed data and the elastic net estimator. Relative MSE (1) is the MSE for the expanding window ARMA model results versus the benchmark model given in columns 6. Relative MSE (2) is the MSE for the elastic net model versus the benchmark in column 6. A relative MSE less than 1 indicates that the proposed model beat the benchmark.

Country	Series	Relative MSE (1)	Relative MSE (2)	N	Benchmark
hn	food03	1.02453	1.22845	56	ar
hn	food04	1.05794	1.02891	56	ar
hn	food05	1.50041	1.41338	56	median
hn	food06	1.06789	1.34548	56	ar
hn	food07	1.02272	1.19456	56	ar
hn	food08	1.02289	1.02308	56	ar
hn	infl01	1.14475	1.07673	94	ar
hn	infl02	0.954699	1.13202	94	ar
sv	food01	1.32029	1.34929	69	ar
sv	food02	0.981255	1.17397	69	ar
sv	food03	0.999156	1.4412	69	ar
sv	food04	1.00938	1.37656	69	ar
sv	food05	1.0587	1.22251	69	median
sv	food06	1.01563	1.2812	69	ar
sv	food07	1.14023	1.42201	69	median
sv	food08	1.26369	1.31265	69	median
sv	food09	1.08291	1.07086	69	median
sv	food10	1.05501	1.22104	69	ar
sv	food11	1	0.998684	69	median
sv	food12	1.24149	1.24152	69	median
sv	infl01	1.44499	1.33092	34	median
sv	infl02	0.986749	0.834845	34	ar
sv	infl03	1.10362	1.51751	34	median
sv	infl04	0.97742	0.97742	34	median
sv	infl05	1.11776	1.15664	34	ar
sv	infl06	1	0.999088	34	mean

Continued on next page

Table 3: The results for each series using the ETS smoothed data and the elastic net estimator. Relative MSE (1) is the MSE for the expanding window ARMA model results versus the benchmark model given in columns 6. Relative MSE (2) is the MSE for the elastic net model versus the benchmark in column 6. A relative MSE less than 1 indicates that the proposed model beat the benchmark.

Country	Series	Relative MSE (1)	Relative MSE (2)	N	Benchmark
sv	infl07	1.29662	0.986124	34	ar
sv	infl08	0.948536	0.948536	34	median
sv	infl09	1.0029	1.00293	34	mean
sv	infl10	1.00788	1.13624	34	mean
sv	infl11	1.18043	1.03241	34	mean
sv	infl12	1.03771	1.01186	34	median
sv	infl13	1.04429	1.29648	34	mean

6 Conclusion

In this paper, we studied the possibility of using Internet search keyword data to nowcast price changes in Central America. We gathered price data for Costa Rica, El Salvador, and Honduras. We also identified several search keywords and downloaded data for them from Google Trends over a period of weeks. We tried several aggregation, smoothing, and linear index construction methods for this Internet search data and were partially successful in improving nowcasts for Costa Rica and for El Salvador, countries for which the search data were of higher quality.

As part of the exercise, we were able to identify several important points for practictioners who wish to forecast using high-dimensional Internet search keyword time series. First, variable selection is of upmost importance. Many, if not most, of the successful forecasting studies that use Internet search keyword data are based on some theory of consumer behavior. This may be the idea that consumers use the Internet to do research before the purchase of a consumer durable as in Carrière-Swallow and Labbé [2013] or searches for jobs and unemployment and welfare as in Choi and Varian [2009]. In the absence of a strong model of consumer behavior, one should incorporate some kind of variable selection mechanism. Naively including a large number of search keyword terms into a model for a search index with the hope that the coefficients on unimportant terms will be small leads to very poor results. However, by employing some variable selection methods from the statistical learning literature we were able to substantially improve all of our forecasts and beat both the ARMA models and benchmarks in several instances.

The second takeaway is the importance of order identification in ARIMA modeling. This is perhaps not a surprise for any forecaster, but the successful results here using automatic techniques are encouraging. If a forecaster were to focus on fewer series and apply the Box-Jenkins methodology rather than relying on automatic model selection procedures it might be possible to outperform the benchmark models further.

Finally, this study suggests several avenues for further research. We might consider more estimators such as the TS-LARS, which is LARS estimator explicitly written with time-series data in mind [Gelper and Croux, 2008]. It allows selection of distributed lags and ranking of predictors. Ranking of predictors will be of particular interest for those who use an exercise such as the one in this paper to generate ideas about consumer behavior and the search for keywords and categories that help forecast price changes. One might also explore using dynamic linear models or a structural model where the Internet search information stands in explicitly for some aspect of the theoretical model. There are also a number of different smoothing techniques and variable selection methods

that might be explored.

In conclusion, the study of the manifestation of consumer sentiment via Internet search behavior is very much still in its infancy. It certainly presents a number of challenges, but the potential insights and use cases are varied and exciting. It may be tempting to dismiss this excitement as hype. All the same, it is difficult to deny the possible benefits of real-time consumer sentiment to future economics research and forecasting studies.

References

Yan Carrière-Swallow and Felipe Labbé. Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298, 2013.

Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment benefits. *Technical Report*, 2009.

Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

Lawrence J Christiano and Terry J Fitzgerald. The band pass filter*. *international economic review*, 44(2):435–465, 2003.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Michael Ettredge, John Gerdes, and Gilbert Karuga. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11):87–92, 2005.

Sarah Gelper and Christophe Croux. Least angle regression for time series forecasting with many predictors. *FBE Research Report KBI_0801*, 2008.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4): 665–676, 2008.

Giselle Guzman. Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, 36(3):119–167, 2011.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer, 2008.

Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, 26(3), 2008.

Rob J Hyndman, Anne B Koehler, Ralph D Snyder, and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3):439–454, 2002.

Spyros Makridakis, SC Wheelwright, and Rob J Hyndman. Forecasting: methods and applications. 1998.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Torsten Schmidt and Simeon Vosen. Using internet data to account for special events in economic forecasting. *Ruhr Economic Paper*, (382), 2012.

Time Series Research Staff. *X-13ARIMA-SEATS Reference Manual*. Statistical Research Division U.S. Census Bureau, 1.1 edition, 2013.

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460): 1167–1179, 2002.

Tanya Suhoy. *Query indices and a 2008 downturn: Israeli data*. Research Department, Bank of Israel, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Simeon Vosen and Torsten Schmidt. Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6):565–578, 2011.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

A Appendix

This Appendix contains extra information on the variables used in this study. Table A1 provides full series names for the abbreviations used for the forecasted series. More information is provided in Table A2.

Table A1: This table contains the abbreviation used in the main tables and the full series name.

Country	Abbreviation	Series
cr	food01	Costa Rica, National Average, Beans (black), Retail...
cr	food02	Costa Rica, National Average, Beans (black), Wholes...
cr	food03	Costa Rica, National Average, Beans (red), Retail, ...
cr	food04	Costa Rica, National Average, Beans (red), Wholesal...
cr	food05	Costa Rica, National Average, Maize (white), Retail...
cr	food06	Costa Rica, National Average, Maize (white), Wholes...
cr	food07	Costa Rica, National Average, Rice (first quality),...
cr	food08	Costa Rica, National Average, Rice (second quality)...
cr	food09	Costa Rica, National Average, Wheat (flour), Retail...
cr	infl01	cpi
cr	infl02	cpi_alc
cr	infl03	cpi_clothes
cr	infl04	cpi_comm
cr	infl05	cpi_core
cr	infl06	cpi_educ
cr	infl07	cpi_entertain
cr	infl08	cpi_food
cr	infl09	cpi_health
cr	infl10	cpi_household
cr	infl11	cpi_housing
cr	infl12	cpi_misc
cr	infl13	cpi_restaurant
cr	infl14	cpi_trans
hn	food01	Honduras, National Average, Beans (red), Wholesale,...
hn	food02	Honduras, National Average, Maize (white), Wholesal...
hn	food03	Honduras, San Pedro Sula, Beans (red), Wholesale, (...
hn	food04	Honduras, San Pedro Sula, Maize (white), Wholesale,...
hn	food05	Honduras, San Pedro Sula, Rice (second quality), Wh...
hn	food06	Honduras, Tegucigalpa, Beans (red), Wholesale, (USD...
hn	food07	Honduras, Tegucigalpa, Maize (white), Wholesale, (U...
hn	food08	Honduras, Tegucigalpa, Rice (second quality), Whole...

Continued on next page

Table A1: This table contains the abbreviation used in the main tables and the full series name.

Country	Abbreviation	Series
hn	infl01	cpi
hn	infl02	cpi_food
sv	food01	El Salvador, San Salvador, Beans (red), Retail, (US...
sv	food02	El Salvador, San Salvador, Beans (red), Wholesale, ...
sv	food03	El Salvador, San Salvador, Beans (red, seda), Retai...
sv	food04	El Salvador, San Salvador, Beans (red, seda), Whole...
sv	food05	El Salvador, San Salvador, Maize (white), Retail, (...
sv	food06	El Salvador, San Salvador, Maize (white), Wholesale...
sv	food07	El Salvador, San Salvador, Rice, Retail, (USD/Kg)
sv	food08	El Salvador, San Salvador, Rice, Wholesale, (USD/Kg)
sv	food09	El Salvador, San Salvador, Sorghum (Maicillo), Reta...
sv	food10	El Salvador, San Salvador, Sorghum (Maicillo), Whol...
sv	food11	El Salvador, San Salvador, Wheat (flour), Retail, (...
sv	food12	El Salvador, San Salvador, Wheat (flour), Wholesale...
sv	infl01	cpi
sv	infl02	cpi_alc
sv	infl03	cpi_clothes
sv	infl04	cpi_comm
sv	infl05	cpi_educ
sv	infl06	cpi_entertain
sv	infl07	cpi_food
sv	infl08	cpi_furniture
sv	infl09	cpi_health
sv	infl10	cpi_house_fuel
sv	infl11	cpi_misc
sv	infl12	cpi_restaurant
sv	infl13	cpi_trans

Table A2: Full information for all of the food price series used throughout the study.

Country	Region	Series	Units
CR	National Average	Beans (black), Retail	(USD/Kg)
CR	National Average	Beans (black), Wholesale	(USD/Kg)
CR	National Average	Beans (red), Retail	(USD/Kg)
CR	National Average	Beans (red), Wholesale	(USD/Kg)
CR	National Average	Maize (white), Retail	(USD/Kg)
CR	National Average	Maize (white), Wholesale	(USD/Kg)
CR	National Average	Rice (first quality), Retail	(USD/Kg)
CR	National Average	Rice (second quality), Retail	(USD/Kg)
CR	National Average	Wheat (flour), Retail	(USD/Kg)
HN	National Average	Beans (red), Wholesale	(USD/Kg)
HN	National Average	Maize (white), Wholesale	(USD/Kg)
HN	San Pedro Sula	Beans (red), Wholesale	(USD/Kg)
HN	San Pedro Sula	Maize (white), Wholesale	(USD/Kg)
HN	San Pedro Sula	Rice (second quality), Wholesale	(USD/Kg)
HN	Tegucigalpa	Beans (red), Wholesale	(USD/Kg)
HN	Tegucigalpa	Maize (white), Wholesale	(USD/Kg)
HN	Tegucigalpa	Rice (second quality), Wholesale	(USD/Kg)
SV	San Salvador	Beans (red), Retail	(USD/Kg)
SV	San Salvador	Beans (red), Wholesale	(USD/Kg)
SV	San Salvador	Beans (red, seda), Retail	(USD/Kg)
SV	San Salvador	Beans (red, seda), Wholesale	(USD/Kg)
SV	San Salvador	Maize (white), Retail	(USD/Kg)
SV	San Salvador	Maize (white), Wholesale	(USD/Kg)
SV	San Salvador	Rice, Retail	(USD/Kg)
SV	San Salvador	Rice, Wholesale	(USD/Kg)
SV	San Salvador	Sorghum (Maicillo), Retail	(USD/Kg)
SV	San Salvador	Sorghum (Maicillo), Wholesale	(USD/Kg)
SV	San Salvador	Wheat (flour), Retail	(USD/Kg)
SV	San Salvador	Wheat (flour), Wholesale	(USD/Kg)

Country	Series	Source
CR	All CPI Data	Banco Central de Costa Rica
HN	All CPI Data	Banco Central de Honduras
SV	All CPI Data	Banco Central de Reserva de El Salvador

Table A3: Sources for the CPI data for each country considered in the study.

Table A3 lists the sources for the CPI data used for each country. The food price series were all obtained from FAO-GIEWS.