Turun yliopisto
University of Turku

NOWCASTING UNEMPLOYMENT AND RETAIL SALES WITH GOOGLE DATA

Master's Thesis in Economics

Author:
Nikolai Myllymäki

Supervisors:
Prof. Heikki Kauppi
Prof. Jouko Vilmunen

22nd October 2018
Turku

Turun Kauppakorkeakoulu · Turku School of Economics

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Contents

1 Introduction **7**

2 Web data in economics **8**

2.1 Nowcasting with big data 8

2.2 Search engine data in econometrics 9

2.3 Google search queries 10

2.4 Problems and limitations with search engine data 11

2.5 Replication issues in scientific literature: why open data matters 11

3 Bayesian structural time series **14**

3.1 The structural time series model 14

3.2 Bayesian econometrics 15

3.3 Obtaining the posterior 17

3.4 Prior selection 17

3.5 Variable selection: the spike and slab prior 18

3.6 Pure time series prior 20

4 Does Google data help in nowcasting economic time series? **21**

4.1 Evaluation of forecasting model performance 21

4.2 Retail sales prediction accuracy 23

4.3 Unemployment claims prediction accuracy 28

5 Performance disparities with Scott and Varian (2014) **38**

5.1 Inflated errors in the U.S. retail sales pure time series model . 38

5.2 Modifying the regression model priors 39

5.3 Replicating the U.S. initial unemployment claims example . . 43

6 Conclusions **46**

References **46**

List of Figures

1 Original U.S. retail sales analysis: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel), from Scott and Varian (2013) 22

2 Diagnostics plots for the retail sales ARIMA(2, 1, 0) model from January 2004 to August 2012 25

3 Diagnostics plots for the unemployment claims ARIMA(1, 1, 2)[0, 1, 1] model on the training period from January 2004 to August 2012 29

4 Residuals for the unemployment claims ARIMA and correlate models, 2012-2017 forecasts 32

5 Density plots of the residuals for the different unemployment claims models, 2012-2017 forecasts 33

6 Autocorrelation function (upper panel) and partial autocorrelation function (lower panel) on the residuals of the unemployment claims ARIMA model, 2012-2017 forecasts 34

7 Autocorrelation function (upper panel) and partial autocorrelation function (lower panel) on the residuals of the unemployment claims pure time series model, 2012-2017 forecasts . 35

8 Autocorrelation function (upper panel) and partial autocorrelation function (lower panel) on the residuals of the unemployment claims correlate model, 2012-2017 forecasts 36

9 Replication of the U.S. retail sales analysis with both informative and noninformative priors in the pure time series models: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel) 40

10 U.S. retail sales analysis with modified priors on regression models: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel) 42

11 Original U.S. unemployment analysis: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel) from Scott and Varian, 2013 44

12 Replication of U.S. unemployment analysis: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel) 45

List of Tables

1 U.S. retail sales ARIMA(2, 1, 0) training period coefficient estimates, standard errors and p-values 24

2 Prediction root mean squared errors (RMSE) for September 2012 to March 2017 U.S. retail sales 27

3 U.S. unemployment claims ARIMA(1, 1, 2)[0, 1, 1] training period coefficient estimates, standard errors and p-values . . . 30

4 Prediction root mean squared errors (RMSE) for 2012-2017 U.S. unemployment claims forecasting models 31

5 Google Trends categories used in all `econ` models 41

1 Introduction

Official economic time series require significant effort to produce and as a result are typically released with a lag measured not in days or even weeks but months. This lag in availability presents difficulties for decision-makers in government and business as judgement has to be made with outdated information.

Scott and Varian (2014) present a Bayesian structural time series method for short-term forecasting or "nowcasting" economic time series with the help of hundreds of explanatory variables obtained from search engine query data. The presented main advantage of their approach is that a large number of potential explanatory variables can be used to forecast the immediate future, potentially improving both the timeliness of forecasts and accuracy of the predictions in the case of structural breaks such as recessions.

In this thesis the predictive performance of the original models nowcasting U.S. initial unemployment claims and U.S. retail sales are examined for the five years of data accumulated since the publication of Scott and Varian (2014). When nowcasting initial unemployment claims, Google data is found to reduce prediction root mean squared error when compared to a standard ARIMA model out of the original sample. Conversely, no notable performance improvements are found from using Google search query data in a Bayesian structural time series model to nowcast retail sales.

Possible causes for these disparities in prediction performance are studied, and the most compelling explanation found is that a highly informative prior was used in the original study in the case of the reference time series without any regression component. It seems that the original prior decreased the predictive performance of the baseline time series model and thus increased the comparative advantage of the models augmented with Google data. When a noninformative prior is used in the retail sales analysis over the original study period, the predictive performance of the baseline model increases to be on par with the models with Google data, eliminating the perceived advantage of models with Google data over a pure time series model.

2 Web data in economics

2.1 Nowcasting with big data

To avoid using yesterday's data for today's inference and decisions, researchers have found ways of utilising newly available data sources, often referred to in the popular media with the term "big data". One of the more explored avenues has been using private-sector search engine data to augment official statistics. This data is available with a lag of days, which is a considerable improvement to typical government statistics, which often have release lags of months and are routinely updated long after publication. The term "nowcasting" is used to emphasise the near-real-time availability and to distinguish from longer forecasting periods.

In addition to rapid after-the-fact availability, statistics have several features that are almost always desirable:, completeness, low cost, granularity, representativeness and, in the case of time series, temporal length. While this thesis focuses on studying how economic indicators can be updated faster, researchers have suggested and demonstrated several other ways in which non-traditional data sources can help analysis.

Mohebbi et al. (2011) note that search query data is often available in time series with shorter aggregation intervals and more fine-grained geographical information than official statistics. Thus missing time periods and regions can be augmented with search data. Query data is also available publicly at no cost to researchers, which contrasts starkly with the expenses of traditional surveys and other data collection methods. This free availability eases replication efforts (see section 2.5).

More generally, private-sector data can allow researchers to examine the inner workings of companies and markets and presents opportunities for randomised experiments. On the public side, moving from relying on surveys to using administrative data gives researchers access to the complete or nearly complete population, allows them to research variation in characteristics such as wages, education, health and productivity in different subpopulations and makes it easier to create more consistent and longer index time-series. New

quasi-experimental research and tracking outcomes of both natural and controlled experiments becomes more feasible. (Einav and Levin 2014.)

2.2 Search engine data in econometrics

Search engine data, often from Google, has been used in numerous studies to nowcast or examine economic phenomena. Kristoufek (2013) uses search query data for stock portfolio diversification. He uses searches of the stock's ticker as a measure of its popularity and decreases the weight of popular stocks in the portfolio to decrease portfolio riskiness. Similarly Da et al. (2011) examine the search volumes of companies' stock tickers and main products. They employ a vector autoregression model finding that search volumes predict stock prices. Goel et al. (2010) predict movie revenues and video game sales with a linear regression model incorporating Yahoo search volume as an explanatory variable. Preis et al. (2010) investigate the links between internet searches and stock market transaction volumes. Mohebbi et al. (2011) research the rate of mortgage refinancing in the United States.

In addition to the retail sales nowcasting in Scott and Varian (2014), private consumption has been forecasted with Google data by Kholodilin et al. (2010) in Germany and Vosen and Schmidt (2011) in the U.S. Both study year-on-year growth rates of private consumption with autoregressive models augmented with Google data.

While diverse in subject matter, the literature is also geographically heterogeneous: Artola et al. (2015) forecast tourism inflows to Spain, Askitas and Zimmermann (2009) examine German unemployment and Carrière-Swallow and Labbé (2013) nowcast automobile sales in Chile.

In addition to economic problems, search engine data has been used for predicting and nowcasting the rates of numerous diseases such as influenza (Lampos et al. 2015). Mohebbi et al. (2011) list several other examples.

In earlier literature, the predictive queries were usually picked by the researchers. With large datasets or numerous models this becomes impractical. The Bayesian structural time series model examined in this thesis was introduced by Scott and Varian (2014) for nowcasting economic time series with

Google Correlate data characterised by numerous individual keywords. One of the attractive features of this model is that it performs variable selection, which is useful when using search engine data with hundreds of potential predictors. In addition to nowcasting economic time series, Bayesian structural time series have been used for causality inference in advertising and studying the effects of security patrolling on crime (Brodersen et al. 2015; Liu and Fabbri 2016).

2.3 Google search queries

Google offers a service, Google Correlate, for finding the 100 queries aggregated by week or month that have the highest correlations with an input time series. Google Correlate ceased updating new data in 2017, but data from January 2004 to March 2017 is still accessible through the website. The service is based on an approximate nearest neighbour algorithm, where each week is represented as a dimension in space. The system finds the approximate nearest neighbours (scaled query volumes that correlate the most with the scaled input series) and calculates the exact correlation coefficients on the top results (Mohebbi et al. 2011). The results can be narrowed down by region and time range, and are scaled to have mean of 0 and a standard deviation of 1.

Google Trends is the main Google portal for accessing Google query data. Individual keyword search volumes can be obtained, and the service also has over a thousand different categories or "verticals" for grouping search queries. Searches can be narrowed down by region as well, and similarly to Google Correlate each output is scaled, this time by normalising and scaling to have values between 0 and 100 (Scott and Varian 2014, 7). Because of the normalisation, the output of the system is dependent on when the data is fetched and varies slightly over time. Both Trends and Correlate data are used in this thesis.

2.4 Problems and limitations with search engine data

While search engine query data can be utilised in econometrics, the limitations have to be borne in mind. As always, correlation does not imply causation, and spurious correlations are abundant in the results of Google Correlate. There is a high risk of overfitting when a large dataset of highly correlating variables is used as a predictor for a time series, and multicollinearity in predictors might cause problems with inference.

Mohebbi et al. (2011) list several other reasons for caution. User search behaviour may change over time, altering or overturning a previous pattern of correlation. An example of this is the Google Flu model which has had mixed results in different years (Lampos et al. 2015). The underlying motivation of a person entering a search query can also be quite different from the explanation that seems obvious to the researcher. The data is not a random sample from the whole population, and time series with low or regular variation typically yield no meaningful correlations. More generally, Einav and Levin (2014) state that possibly the greatest challenge in using big data is finding ways to respect privacy and confidentiality.

New data sources and sophisticated methodologies can also trick us to believe that complexity automatically offers increased performance in comparison to simpler models. In some cases, this hypothesis is quite probably true (the first example in Scott and Varian (2014) might well be one of those cases), but complexity in methodology can also increase the opacity of analyses and make replication of previous studies more difficult even when freely available data eases replication efforts.

2.5 Replication issues in scientific literature: why open data matters

This thesis started as an attempt to use Bayesian structural time series, introduced by Scott and Varian (2014), to nowcast Finnish economic time series with the aid of Google search data. To familiarise myself with the method and software I tried to reproduce the second example, nowcasting

U.S. retail sales, of the original paper. My efforts were unsuccessful until a probable misspecification of prior hyperparameters in the original study was identified.

The encountered problems in replicating the original study were unexpected and, at first, puzzling. However, recent research into the level of replicability in science has shown that failures in reproducing the results of peer-reviewed studies are quite common.

To rigorously assess the replicability of published research in well-regarded journals, the Open Science Collaboration (2015) shows massive effort in replicating one hundred experimental and correlational studies in psychology with new experiments and data. While replication success is assessed in several ways, only 39% of the new studies are subjectively rated as having reproduced the original findings. The general conclusion is that the majority of replications offer weaker evidence supporting the original conclusions than the original results.

Economics is not immune to replication problems. In fact, Duvendack et al. (2017) posit that economics is behind several other fields in the practice of replicating studies. They cite psychology and, somewhat less, political science as spearheading replication practices in social sciences. Measuring the rate of reproducibility in economics, Chang and Li (2015) have a success percentage of 49 when trying to replicate the original analysis using author-provided replication code and data. Their measure of replication success is different from the Open Science Collaboration (2015) or this thesis, a failure being an instance where original code or data cannot be obtained, but the study highlights problems in reproduction even when the publishing journals supposedly require complete replication files as a prerequisite for publication.

In light of the frequency of replication problems, the issues presented in the section 5 are not so surprising as first. I want to thank both Mr. Scott and Mr. Varian for generously answering my questions and Mr. Scott in particular in suggesting the possibility that the default values of priors in the `bsts` package might have changed since the article's publication. Priors had occurred to me as a possible source for the error, but before Mr. Scott's suggestion I had dismissed the idea as the paper stated that default values

in `bsts` were used.

I also applaud the authors for supplying the code and data (Scott 2017b) for replicating the first analysis in their paper. This practice of publishing the code and data used in scientific papers is exemplary and helped me in producing my analysis. Finally, Google should be thanked as well for providing access to its data, easing both replication efforts and new studies.

3 Bayesian structural time series

3.1 The structural time series model

Scott and Varian (2015) describe Bayesian structural time series as a combination of three statistical methods: a basic structural time series model, variable selection with spike and slab regression and Bayesian model averaging. The authors die not invent the underlying concepts, but rather combined them together in a novel way in a comprehensive method and supplied software as well as practical examples for its usage.

The model takes as its inputs the time series to be predicted and time series of several hundred Google search queries correlated with it. The introduction of a regression component into the analysis has two potential benefits: the earlier availability of search query data can help make predictions faster, and introduction of search activity data can aid in the detection of structural breaks in the time series being predicted. The method is designed to reduce the number of utilised explanatory query variables from hundreds to a handful, which is valuable when using Google data where the amount of available variables is large.

Scott and Varian (2014, 8) present their model in state space form:

$$y_t = Z_t^T \alpha_t + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma_\epsilon) \tag{1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \qquad \eta_t \sim N(0, Q). \tag{2}$$

Here (1) is the *observation* equation, linking the observations y_t with the unobserved latent state α_t. Equation (2) is the *transition* equation, explaining how the latent state vector $\boldsymbol{\alpha}$ evolves over time. Z_t, T_t and R_t are model matrices the arrangements of which can achieve different specifications of the model.

Durbin and Koopman (2001, 1–2) define structural time series as models in which the observations consist of several components: trend, seasonal, cycle, regression and error terms. The Bayesian structural time series model being inserted into equations (1) and (2) can be presented by the components forming a local linear trend model, augmented with a regression component:

$$
\begin{aligned}
y_t &= \mu_t + \beta^T \mathbf{x}_t + \epsilon_t \\
\mu_t &= \mu_{t-1} + \delta_{t-1} + u_i \\
\delta_t &= \delta_{t-1} + v_t
\end{aligned} \tag{3}
$$

Here μ_t is the present level of the trend, δ_t is the slope of the trend and $\beta^T \mathbf{x}_t$ is the regression component. ϵ_t, u_t and v_t are independent random noise components, initially assumed to be Gaussian with zero expected value. The Gaussian assumption is relaxed in section 4.

In equation (3) the basic local linear trend model is expanded by the regression vector component $\beta^T \mathbf{x}_t$. In our application the regression component contains the search queries from Google Correlate and categories from Google Trends. The parameters to be estimated in equation (3) are the variances of the noise terms $\sigma_\epsilon^2, \sigma_u^2$ and σ_v^2 (with $Q = \text{diag}(\sigma_u^2, \sigma_v^2)$) and the regression coefficients β. More detailed information on the model specification can be found in Scott and Varian (2014), and the Bayesian aspect of the model is explained in the following sections.

3.2 Bayesian econometrics

In the Bayesian framework data is fixed, while the parameters of the model are the random variables the distributions of which are estimated. Bayesian inference rests on the Bayes theorem:

$$
P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}. \tag{4}
$$

Substituting probabilities with conditional density functions we have

$$
p(\theta \mid y) = \frac{f(y \mid \theta)\, p(\theta)}{f(y)}, \tag{5}
$$

where θ is the parameter (or vector of parameters) we wish to estimate and y is the observed data. Here $p(\theta \mid y)$ is called the *posterior density function* of θ given observations y. $f(y \mid \theta)$ is the *likelihood function*, which is the density function of observations y given the parameter value θ. When vectors of

parameters are estimated, we speak of *joint* functions. (Greenberg 2013, 13-15, 23.)

$p(\theta)$ is called the *prior density function*, or prior in short, which introduces the researcher's prior beliefs and knowledge about the distribution of θ. These subjective priors are not arbitrary and can be based on preceding empirical work. The updating of existing information, introduced to the analysis with the prior density function, is a central theme in Bayesian analysis.

The denominator $f(y)$ in equation (5) is equal to $\int f(y \mid \theta)p(\theta)\,d\theta$. Its function is to normalise the posterior distribution so that the integral $\int p(\theta \mid y)\,d\theta$ (or the sum in the case of discrete distributions) is equal to one as dictated by Kolmogorov's second axiom.

Of the elements of equation (5), the prior is the subject to controversy because of its subjective nature. Sivia and Skilling (2006, 12) state that while the assigning of prior probability distributions might seem like it would make a large impact on the results, in practice the effect is usually not crucial. Generally the impact of the prior on the posterior distribution decreases when the size of the sample grows (Greenberg 2013, 17). These statements, however, refer to *noninformative*[1] priors that are designed to impact the posterior distribution as little as possible. While the statements about the lowering impact of the prior are true in the limit, when a highly *informative* prior is used results can be distorted substantially even with moderately high observation counts. This phenomena is central to the issues analysed in section 5, and it is why highly informative priors are typically only used explicitly and when there is a need to include in the analysis a large amount of earlier evidence or research that suggests a specific value for the parameter in question.

An important but less controversial decision is also made when choosing the likelihood function, where an explicit function is required to be able to derive the posterior distribution. This decision of functional form can also be seen as a part of prior information but it is represented in the likelihood

[1]The term "noninformative" is used throughout the thesis in the sense of being as little informative as possible, to be distinguished from noninformative priors in the strict mathematical sense as described by Greenberg (2013)

function by convention. Because the posterior distribution is proportional to the prior distribution and the likelihood function, the chosen form of the likelihood is an important factor in the final inference. Nevertheless, as the sample size grows, the posterior distribution converges towards a Gaussian distribution and the impact of the chosen functional form of the likelihood function decreases. (Greenberg 2013, 21–23; 29.)

3.3 Obtaining the posterior

Equation (5) gives us the foundation of Bayesian inference. Combining the prior and the likelihood function yields us the posterior density function. In the case of a multivariate distribution, say that we are interested primarily in the parameter θ_1. In this situation we would want to find out the marginal density $p(\theta_1 \mid y)$ by integrating out the other parameters from the joint posterior distribution (Greenberg 2013, 23):

$$p(\theta_1 \mid y) = \int p(\theta_1, ..., \theta_d \mid y) d\theta_2 \cdots d\theta_d. \tag{6}$$

This seems straightforward, but unfortunately in many cases the integrals cannot be found analytically. In more complex instances such as Bayesian structural time series, Markov chain Monte Carlo (MCMC) methods are used instead to simulate the posterior distribution. Regardless of the computational method the principles of inference remain the same.

3.4 Prior selection

Time series often have relatively small sample sizes, as the number of observations is limited by the length of the data collection period and serial correlation reduces the effective sample size (Stigler 2016, 60). In the Bayesian case this has implications to the impact of the prior on the results, as the posterior distribution is greatly influenced by the prior in small samples (Greenberg 2013, 43). An example of the dangers of unmindful prior selection is given in section 5, where the results of Scott and Varian (2014) are found to be substantially inflated by a prior distribution unsuitable to the situation.

Any probability distribution can be chosen as the prior distribution as long as it can be integrated to unity. If a prior has this property it is called *proper*. If possible, the prior distribution is typically chosen from the same family (same form but different parameters) as the posterior distribution. These priors are called *conjugate* to the posterior distribution. (Greenberg 2013, 45-47.) The parameters of the prior distribution are called *hyperparameters* to differentiate from the parameters of the posterior distribution we are calculating.

In Bayesian structural time series the prior depends on whether or not a regression component is present. In a pure time series model, the prior is defined with an inverse gamma distribution. The gamma function and its sibling inverse gamma function have desirable properties for a prior distribution, as they have great diversity in the shapes obtainable by changing the hyperparameters. The exponential and χ^2 distributions are special cases of the gamma distribution, for example. The normal distribution is in the same family of exponential distributions as the gamma distribution (Nielsen and Garcia 2009), making them conjugate and simplifying analytical solutions in Bayesian inference.

3.5 Variable selection: the spike and slab prior

Because there are thousands of Google search verticals and queries that can be inserted into the regression component in equation (3), some kind of automated variable selection has to be performed. In the Bayesian paradigm the preferred way to do this is to use a "spike and slab" prior on the regression coefficients. The name is indicative of its central idea and graphical representation: for each regression coefficient's inclusion in the model, there is a "spike" of high probability mass at zero. The "slab" indicates that the prior distribution for the value of each regression coefficient β_k is very weakly informative and thus close to flat. (Scott and Varian 2014, 10.)

The spike and slab prior is formally represented by

$$p(\beta, \gamma, \sigma_\epsilon^2) = p(\beta_\gamma | \gamma, \sigma_\epsilon^2) p(\sigma_\epsilon^2 | \gamma) p(\gamma). \tag{7}$$

Here γ is a vector of zeroes and ones indicating the inclusion of parameters in β to the regression, with $\gamma_k = 1$ if $\beta_k \neq 0$ and $\gamma_k = 0$ if $\beta_k = 0$. β_γ is the subset of elements where $\gamma_k = 1$ and $\beta_k \neq 0$.

The marginal distribution $p(\gamma)$ represents the spike and is Bernoulli distributed:

$$\gamma \sim \prod_{k=1}^{K} \pi^{\gamma_k}(1-\pi)^{1-\gamma_k}. \tag{8}$$

π in equation (8) represents the expected model size, so that $\pi = p/K$ when p is the expected number of predictors that have $\beta_k \neq 0$ and K is the number of total input regression variables.

The "slab" portion of the prior is expressed by a pair of conditionally conjugate distributions:

$$\beta_\gamma|\sigma_\epsilon^2, \gamma \sim N(0, \sigma_\epsilon^2(\Omega_\gamma^{-1})^{-1}) \qquad \frac{1}{\sigma_\epsilon^2}|\gamma \sim Ga(\frac{\nu}{2}, \frac{ss}{2}). \tag{9}$$

To obtain Ω_γ^{-1}, we need to define Ω^{-1} as the symmetric full-model prior information matrix obtained by setting $\Omega^{-1} = \kappa(w\mathbf{X}^T\mathbf{X}+(1-w)\text{diag}(\mathbf{X}^T\mathbf{X}))/n$. Here $\mathbf{X}$ is the observation matrix where predictor $\mathbf{x}_t$ is on row t, $w = 1/2$ and $\kappa = 1$ are hyperparameters with the values used by Scott and Varian (2014) and $\text{diag}(\mathbf{X}^T\mathbf{X})$ denotes the diagonal matrix with $\text{diag}(\mathbf{X}^T\mathbf{X})_{ii} = (\mathbf{X}^T\mathbf{X})_{ii}$. Now, Ω_γ^{-1} in equation (9) denotes the rows and columns of Ω^{-1} that correspond to $\gamma_k = 1$.

$GA(r, s)$ is the gamma distribution with mean r/s and variance of r/s^2. The prior sample size or weight given to the prior is represented by the hyperparameter ν, and the prior sum of squares of the regression by ss. The prior sum can be set with the expected R^2 from the regression and ν by setting $ss/\nu = (1 - R^2)s_y^2$, where s_y^2 is the marginal standard deviation of the dependent variable.

The spike and slab prior, presented here in slightly simplified form, has numerous hyperparameters and can seem complicated; however, the the most important hyperparameters in this thesis are the expected model size p, expected R^2 and sample size ν. The values of these hyperparameters and their

impact on the results of the analysis are discussed further in section 5.2. The MCMC algorithm that utilises equations (7) to (9) to simulate draws from the posterior distribution $p(\beta, \sigma_\epsilon^2, \sigma_u^2, \sigma_v^2, \boldsymbol{\alpha} \mid \mathbf{y})$ is described in more detail by Scott and Varian (2014).

3.6 Pure time series prior

When the regression component $\beta^T \mathbf{x}_t$ is dropped from equation (3) for pure time series analysis, the prior can be expressed in simpler terms. Now,

$$\frac{1}{\sigma_\epsilon^2} \sim Ga(\upsilon, \lambda). \tag{10}$$

In the absence of a regression component, prior information is entered to the `bsts` function as a prior guess at the value of the residual standard deviation and a weight given to this guess, interpretable as the number of prior observations (Scott 2017a; Scott 2018). These inputs are then translated into hyperparameters υ and λ of an inverse gamma distribution, analogous with equation (10):

$$\sigma_\epsilon^2 \sim InvGa(\upsilon, \lambda). \tag{11}$$

There is no explicit explanation in any of the relevant software packages or the original paper of how the software input values of the prior guess at the size on the residual standard deviation and the weight given to it translate into values of υ and λ in equation (11). Derivation of the exact analytical form of this relationship is outside the scope for this thesis, so in subsequent sections the intuitively easily interpretable "prior guess at the value of the residual standard deviation" and "prior observation count" are used. This convention has the added benefit of being synonymous with the `sigma.guess` and `sample.size` argument values in the `SdPrior` function of the `Boom` package used in the analysis, simplifying programming if the reader wishes to test the results. Unless otherwise stated, pure time series models in this thesis set the prior on the residual standard errors to 1 and the prior observation count to 0.01, the default values in the current `bsts` package.

4 Does Google data help in nowcasting economic time series?

4.1 Evaluation of forecasting model performance

Scott and Varian (2014) use cumulative absolute one step ahead prediction errors to illustrate the in-sample performance of Bayesian structural time series in nowcasting monthly seasonally adjusted retail sales excluding food services in the United States (figure 1). The upper panel of the graph shows the cumulative errors, while the lower panel shows the scaled values of the actual time series that is being predicted.

In figure 1 the "pure.time.series" graph denotes a pure time series model on the retail sales data where the regression component from equation (3) is omitted. The "correlate" model has as its regression component the one hundred Google Correlate search queries matching the retail sales time series most closely, as explained in section 2.3. The "econ" model utilises both Google Correlate and Google Trends data, where the Trends categories have been hand-picked by the authors according to their potential relevancy to retail sales. The "all" model utilises the Correlate data and all (over 600) Google Trends verticals available at the time of the release of the study. The deseasonalised versions of the models have the same data with the difference that the regression components have been deseasonalised before fitting the models.

The approach in figure 1 compares only the in-sample performance of the models. The question of how well the predictions would fare in out-of-sample forecasting is not answered. Luckily, almost five years of Google Correlate data was collected between the final observation of the original analysis and the ceasing of the Correlate service's updating. This creates a natural way of testing how a model trained using data from the original period from January 2004 to August 2012 would have fared in predicting retail sales in consequent years.

Inoue and Kilian (2006) compare two forecasting model selection methods: simulated out-of-sample (SOOS) prediction root mean squared error

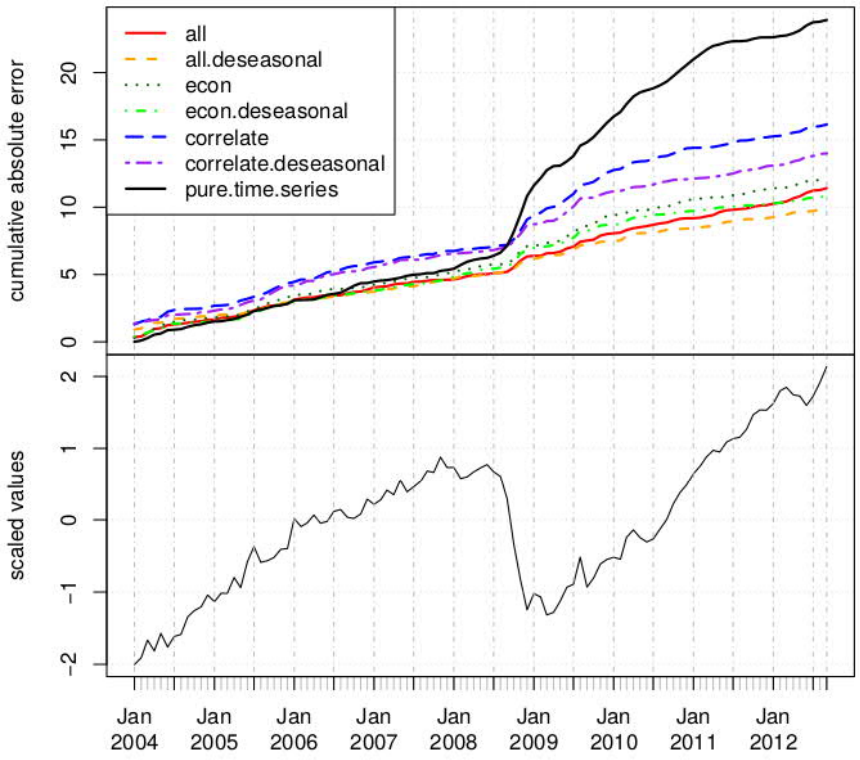

Figure 1: Original U.S. retail sales analysis: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel), from Scott and Varian (2013)

(RMSE) and information criteria based approaches. They find that under specific conditions the information criteria based approach is consistent and preferable to the more standard SOOS method.

However, the information criteria approach uses the number of parameters in the forecast model as an input. Because Bayesian structural time series already performs variable selection by placing most regression coefficients' probability mass at zero (section 3.5), it is not immediately clear which number should represent the correct number of parameters. For example, should the researcher use the number of input explanatory factors, the predicted model size or the mode or median of the posterior distribution

of model size? The RMSE method is used here because it is less ambiguous in the present case. Most importantly, since we have five years of new data at our disposal, we can perform true out-of-sample measurements of prediction errors and compare different models directly with the RMSE.

The prediction RMSE is obtained by fitting each model on the first S observations, using the fit to forecast observation $S + h$ and repeating this for $S = R, R+1, R+2, \ldots, T-h$ where T is the number of total observation periods, 159. (Inoue and Kilian 2006.) In our nowcasting example, the forecast horizon h is equal to 1. As we are interested in the model performance after the original study was conducted, the value for the starting sample size R is chosen to be the number of observations in the original study period from January 2004 to August 2012: $R = 104$. The method is computationally intensive, with the calculations taking over four hours in the retail sales case and over fifteen hours in the unemployment claims example on a four-core 1.7 GHz laptop running GNU/Linux Debian 9 and R version 3.3.3.

It should be noted that fitting a Bayesian structural time series model multiple times with different training period lengths results in $T - R = 55$ posterior distributions that have different probability densities for their regression parameters. This approach is thus not quite identical to a process where variables of model i are predetermined and thus have a inclusion probability of 1. However, for our purpose of assessing the prediction performance of different models, the RMSE method is adequate and easy to understand.

4.2 Retail sales prediction accuracy

The prediction accuracy of the different models are compared using the RMSE method described in section 4.1. The training period is that of the original study, from January 2004 to August 2012. The test period ranges from September 2012 to the end of Google Correlate data in March 2017, a total of 55 monthly observations.

An ARIMA model is also fit to obtain a baseline for model comparison. The Box-Jenkins method was used for model selection and an ARIMA(2, 1, 0) model produced a good fit to the training data. Here an ARIMA(p, d, q)

model means an autoregressive integrated moving average model where p denotes the order of the autogergressive (AR) term, d is the order of differencing and q is the order of the moving average (MA) term. Diagnostics plots for the ARIMA model fit on the training period can be found in figure 2 and a table with coefficient estimates, standard errors and p-values can be found in table 1. None of the coefficients are statistically significant at the $p = 0.05$ level, but as the function of the ARIMA model is to serve as a point of comparison for the Bayesian models, its statistical properties are not of primary interest in any case. The model could possibly be improved by, for example, using year-on-year growth rates instead of absolute values as in Vosen and Schmidt (2011). For comparison purposes, however, using the same unmodified time series as the Bayesian models is attractive, and in any case the ARIMA model is sufficient to provide a competitive alternative to the Bayesian models.

Table 1: U.S. retail sales ARIMA(2, 1, 0) training period coefficient estimates, standard errors and p-values

	Estimate	SE	p.value
AR1	0.1396	0.0965	0.1510
AR2	0.1915	0.0969	0.0509
constant	0.0216	0.0137	0.1175

All the bayesian models were based on equation (3), omitting the regression component $\beta^T \mathbf{x}_t$ in the pure time series case. The default, noninformative prior hyperparameter values in `bsts` were used, as specified in section 3. Two new models are added: only Trends and only Trends with deseasonalised regressors. These include only the economically relevant Google Trends verticals and no Google Correlate queries. The two models are added for testing how well just Trends data can perform since Correlate ceased updating in March 2017. The models with all Trends categories and the full Correlate dataset were not obtained due to the download limitations of Google Trends. There are over 1000 categories in Google Trends, and obtaining all of them

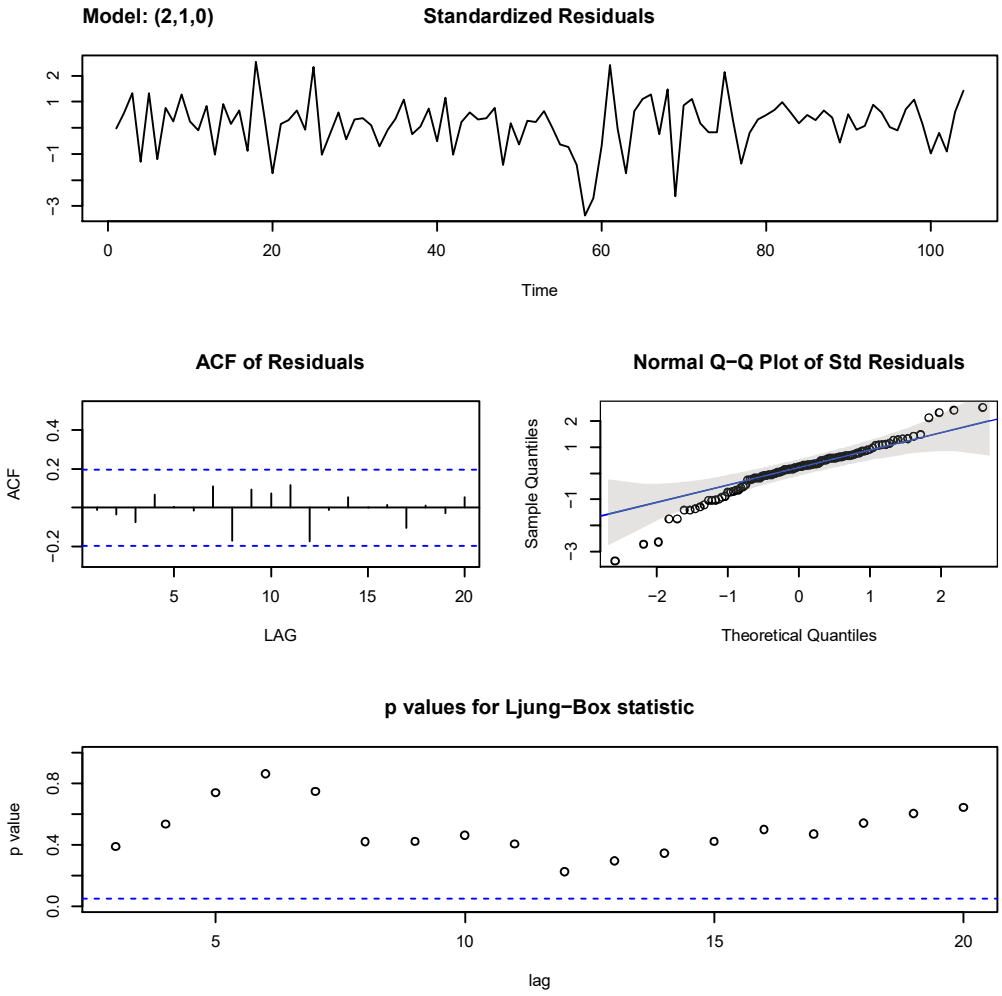

Figure 2: Diagnostics plots for the retail sales ARIMA(2, 1, 0) model from January 2004 to August 2012

proved to be impractical. The omission of the last two models should not radically influence the conclusions presented here.

As before, the "Correlate and Trends" models utilise both Trends and Correlate data, and the Correlate models have as regression components the Google Correlate query volumes from January 2004 to March 2017 that correlated with the U.S. retail sales data most highly during the original period of study, from 2004 to August 2012. The Correlate queries are selected this way purposefully to test how well the queries selected by Google Correlate during the original period would have predicted consequent out-of-sample retail sales volumes. If retail sales data from the whole period of Google Correlate's operation time would have been used as an input to the service, the forecasting performance would have been inflated because by definition *Google Correlate finds the searches that are most highly correlated with the input time series.* The regression components would have had a very high correlation and probably a lower RMSE with retail sales data for the whole period, but a significant part of this performance would very likely have been spurious. Another approach would have been to obtain a separate Correlate dataset for each S, but as Correlate has ceased operating and this method is not applicable in future studies, this approach is not chosen. The retail sales data is provided by the U.S. Bureau of the Census.

The performance of different models is compared in table 2. A lower RMSE indicates better performance. As the table shows, the models with Google data do not generally outperform the baseline ARIMA model, and performance is quite similar between different models. The lowest overall RMSE is obtained with the newly added model that only uses Google Trends data, dropping all Correlate variables. The model with just Trends data also performs slightly better than the ARIMA(2, 1, 0) model, but the difference is so small that the utility of the added complexity of the regression component is questionable. Moreover, the Correlate and Trends model that has the same Trends verticals as the only Trends model as a subset of its regressors has the poorest performance of all the available models. This suggests that if some predictors do indeed include information about retail sales, the Bayesian structural time series model is not efficient enough to separate the signal

Table 2: Prediction root mean squared errors (RMSE) for September 2012 to March 2017 U.S. retail sales

Model	RMSE	Difference to baseline
Baseline ARIMA(2, 1, 0)	0.0645	
Pure time series	0.0658	+1.98 %
Correlate	0.0658	+1.99 %
Correlate, deseasonalised	0.0654	+1.45 %
Correlate and Trends	0.0668	+3.62 %
Correlate and Trends, deseasonalised	0.0654	+1.44 %
Only Trends	0.0643	−0.36 %
Only Trends, deseasonalised	0.0668	+3.60 %

from the noise in this case.

The effect of deseasonalising the predictors is not as uniformly positive as in the original study. Both of the models with Correlate data perform better with deseasonalised predictors, but the model with just Trends data performs better with raw data. In addition to possible different seasonal mechanics of Trends and Correlate data, another possible reason might be in a different deseasonalising method even though the original study is followed as closely as is possible with the information at hand. In this analysis the whole dataset was desasonalised before any modeling, which is a minor violation of the iterative one-step-ahead prediction approach. The theoretical direction of the effect of this would be to improve the results of deseasonalised data as the seasonal decomposition is conducted with the maximum time series length and should be more accurate than when done with S observations on each iteration.

The differences in RMSE do not provide clear evidence that introducing Google data can help in predicting U.S. retail sales, regardless of whether the regression models are compared to the pure time series model or the ARIMA model. The differences between models are so small that for example switching the performance metric from the root mean squared error to mean

absolute error flipped many of the relative advantages and disadvantages that can be seen in table 2. This demonstrates that the models perform quite similarly, which is in line with figure 9 in that Google data does not seem to offer a significant performance boost for predictions. The identified probable reason for the difference in performance with Scott and Varian (2014), namely the prior used in the pure time series model, is examined in detail in section 5.

4.3 Unemployment claims prediction accuracy

To examine whether or not the inability to improve nowcasting with Google data extends beyond retail sales, the U.S. initial unemployment claims example is examined with the same approach using data from the U.S. Employment and Training Administration. The model from equation (12) is used. R is set to the value 456, letting the final observation of the training period to fall on 23.9.2012. This corresponds to the last date in the `iclaims` dataset from the `bsts` package that is quite likely the exact data used in the original study. $T = 688$, so there are a total of 232 one-step-ahead nowcasts for each model with the final observation of the test period falling on 05.03.2017. Because of the seasonality present in the data, a seasonal ARIMA$(p, d, q)[P, D, Q]$ model was used where the length of of the season was set to 52 weeks to catch annual patterns in the time series. As before, p denotes the order of the autogergressive term, d is the order of differencing and q is the order of the moving average term. The seasonal components are denoted inside the square brackets where P denotes the order of the seasonal autogergressive term, D is the order of seasonal differencing and Q is the order of the seasonal moving average term. An ARIMA(1, 1, 2)[0, 1, 1] model was selected as a baseline. The ARIMA model diagnostics are provided in table 3 and figure 3.

For the bayesian models, equation (3) is expanded with a seasonal component τ_t to obtain equation (12). As before, the regression component $\beta^T \mathbf{x}_t$ is omitted in the pure time series model.

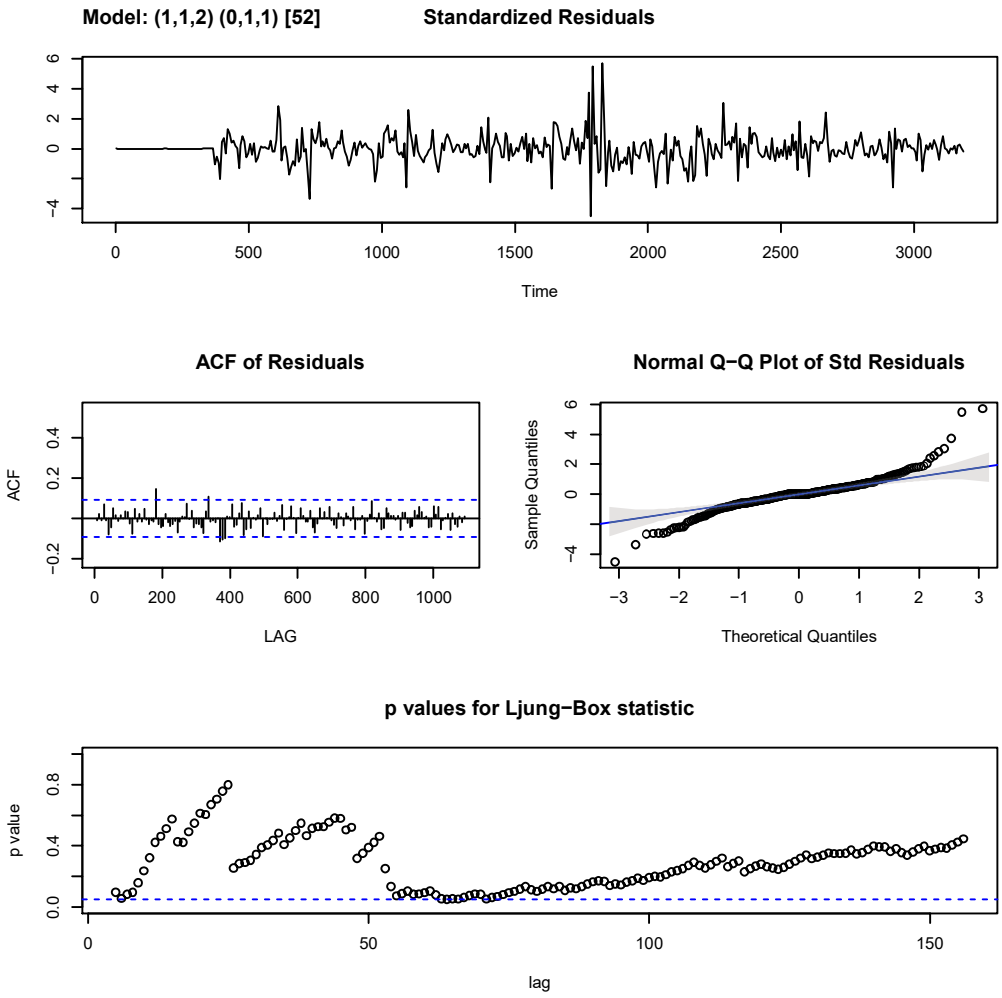

Figure 3: Diagnostics plots for the unemployment claims ARIMA(1, 1, 2)[0, 1, 1] model on the training period from January 2004 to August 2012

Table 3: U.S. unemployment claims ARIMA(1, 1, 2)[0, 1, 1] training period coefficient estimates, standard errors and p-values

	Estimate	SE	p.value
AR1	0.9159	0.0463	0.0000
MA1	-1.6014	0.0529	0.0000
MA2	0.6631	0.0414	0.0000
SMA1	-0.1555	0.0495	0.0018

$$
\begin{aligned}
y_t &= \mu_t + \tau_t + \beta^T \mathbf{x}_t + \epsilon_t \\
\mu_t &= \mu_{t-1} + \delta_{t-1} + u_t \\
\delta_t &= \delta_{t-1} + v_t \\
\tau_t &= -\sum_{s=1}^{S-1} \tau_{t-s} + w_t, \qquad S = 52.
\end{aligned}
\tag{12}
$$

The results of comparing the ARIMA and Bayesian models for nowcasting unemployment claims data are presented in table 4. As can be seen from the percentage differences in RMSE, the Bayesian model without a regression component performed somewhat worse than the baseline ARIMA. However, the Bayesian model augmented with Correlate data outperformed both models substantially, with close to 14% smaller RMSE than the next-best ARIMA forecasts. The difference to the retail sales predictions is sizeable, and the results seem to suggest that Google Correlate data and Bayesian structural time series are indeed helpful in nowcasting U.S. unemployment claims.

A closer examination of the residuals brings some nuances to the analysis. Figure 4 shows the residuals for the ARIMA and Correlate models, omitting the pure time series model for readability. Visual inspection suggests that the Correlate model has fewer extreme mispredictions than the ARIMA forecasts, indicating that Google data can perhaps help in indicating when the underlying series is subject to structural shocks that cause larger deviations from the trend.

Table 4: Prediction root mean squared errors (RMSE) for 2012-2017 U.S. unemployment claims forecasting models

Model	RMSE	Difference to baseline
Baseline ARIMA(1, 1, 2)[0, 1, 1]	0.2534	
Pure time series	0.2644	+4.34 %
Correlate	0.2188	−13.64 %

Conversely, the density plots in figure 5 show that while there are more extreme values in the ARIMA residuals, they are still more closely concentrated around zero than in the Bayesian models when the absolute values of the residuals are small. Overall, neither figure 4 or figure 5 can be claimed to offer unambiguous evidence for the supremacy of either of the models, but they do highlight one of the reasons for the better RMSE of the Correlate model: the squaring in the process of calculating the RMSE makes the measure react porpotionately more to larger absolute values of residuals. This is, of course, a deliberate choice in the development of RMSE, as a few massive forecast errors can often be disproportionately worse for decisionmakers than many marginally larger but still small forecast errors. This analysis might be a case where the type of error the analyst wants to minimise guides the analysis method.

Could the models be improved upon? Figures 6 to 8 show the autocorrelation funcions and partial autocorrelation functions for the residuals of three models. While a p-value analysis on the residuals of a Bayesian model might not be theoretically quite consistent, such a review was nevertheless conducted and indicates that there is statistically significant autocorrelation between residuals in the first and third degrees ($p = 0.05$) in the Google Correlate model but not in the ARIMA or pure time series models. This suggests that there might still be room for improvement in the model, although the differences in the ACF and PACF results are not drastic between models nor are the significant p-values much under the $p = 0.05$ treshold.

There are several possible reasons for why Bayesian structural time series

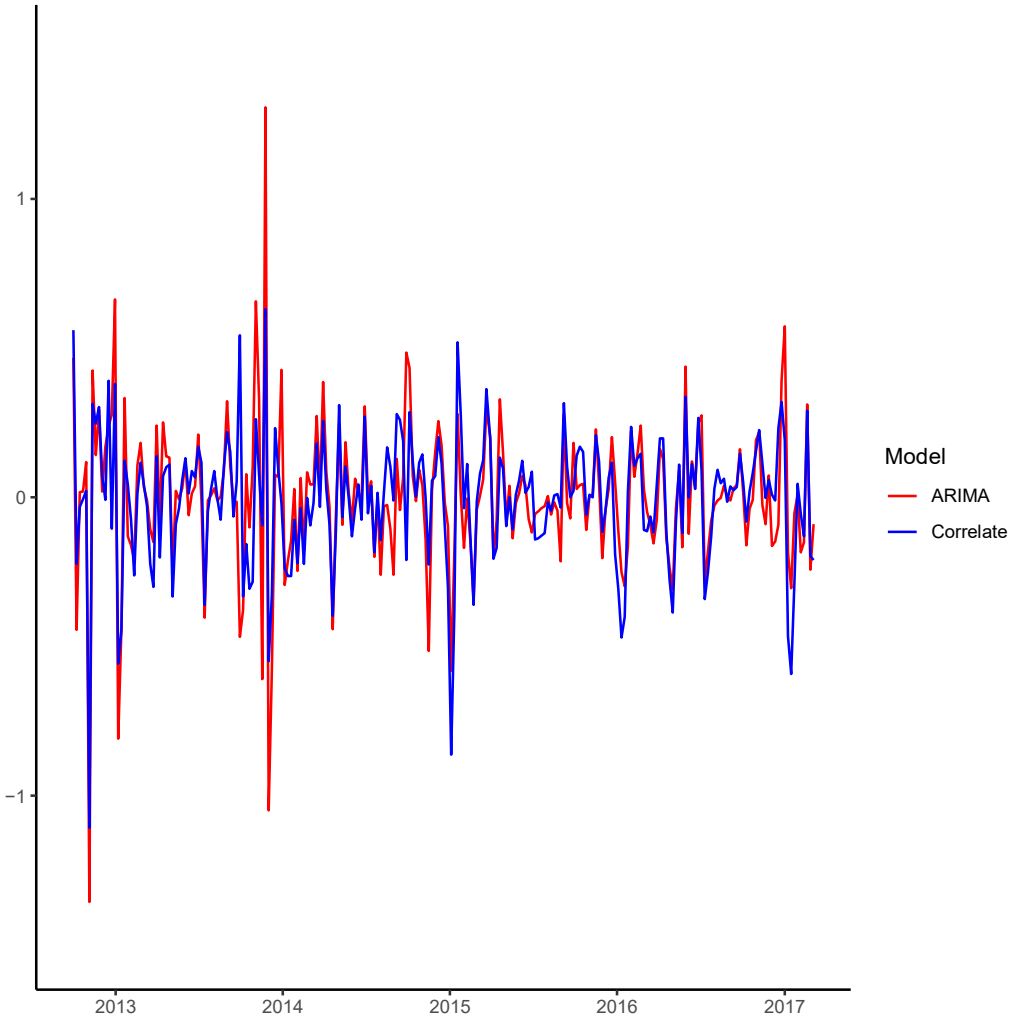

Figure 4: Residuals for the unemployment claims ARIMA and correlate models, 2012-2017 forecasts

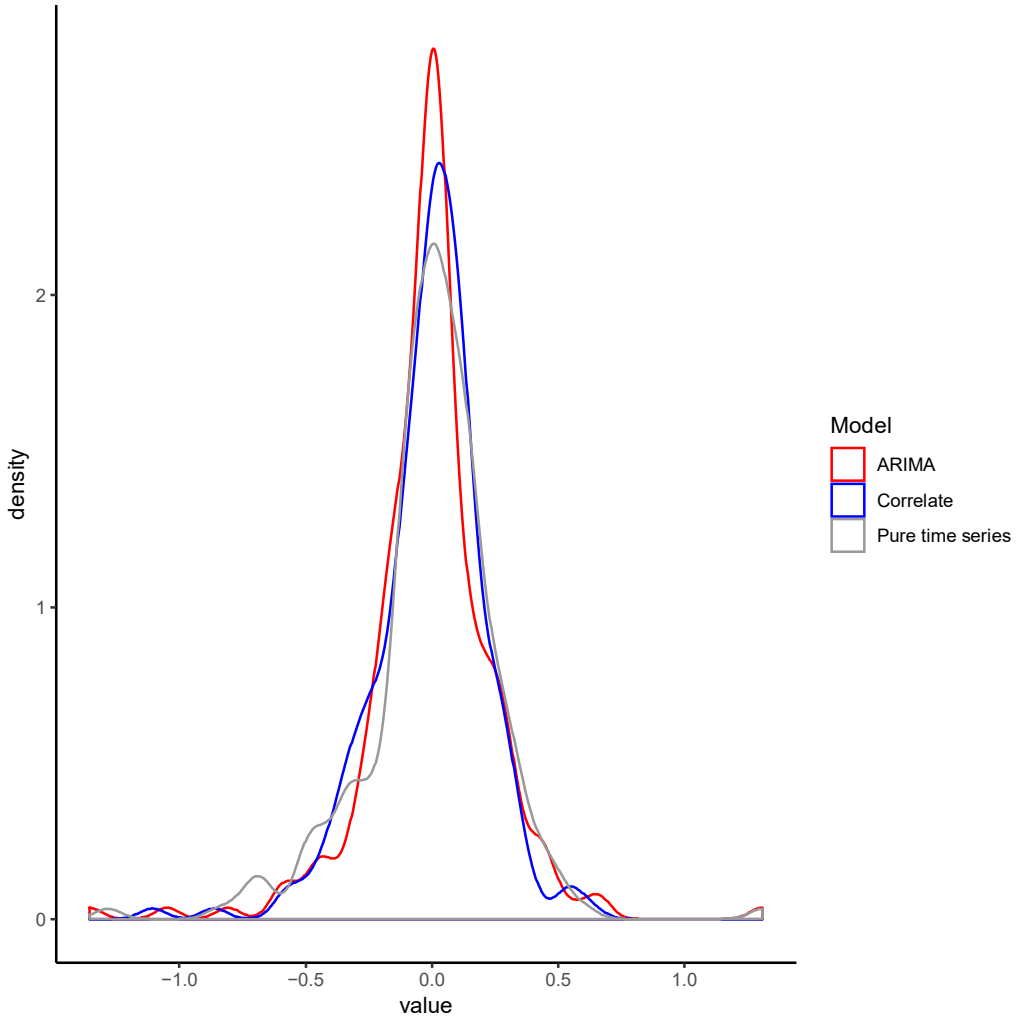

Figure 5: Density plots of the residuals for the different unemployment claims models, 2012-2017 forecasts

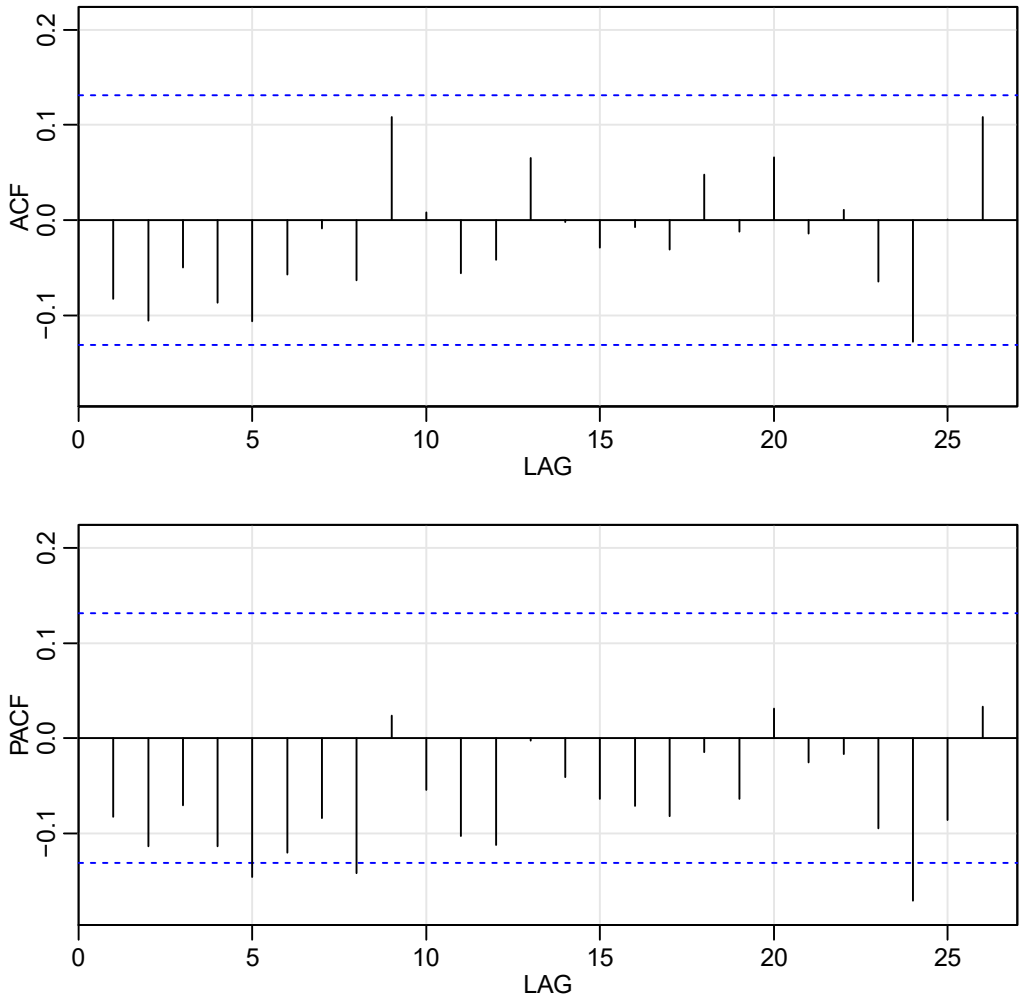

Figure 6: Autocorrelation function (upper panel) and partial autocorrelation function (lower panel) on the residuals of the unemployment claims ARIMA model, 2012-2017 forecasts

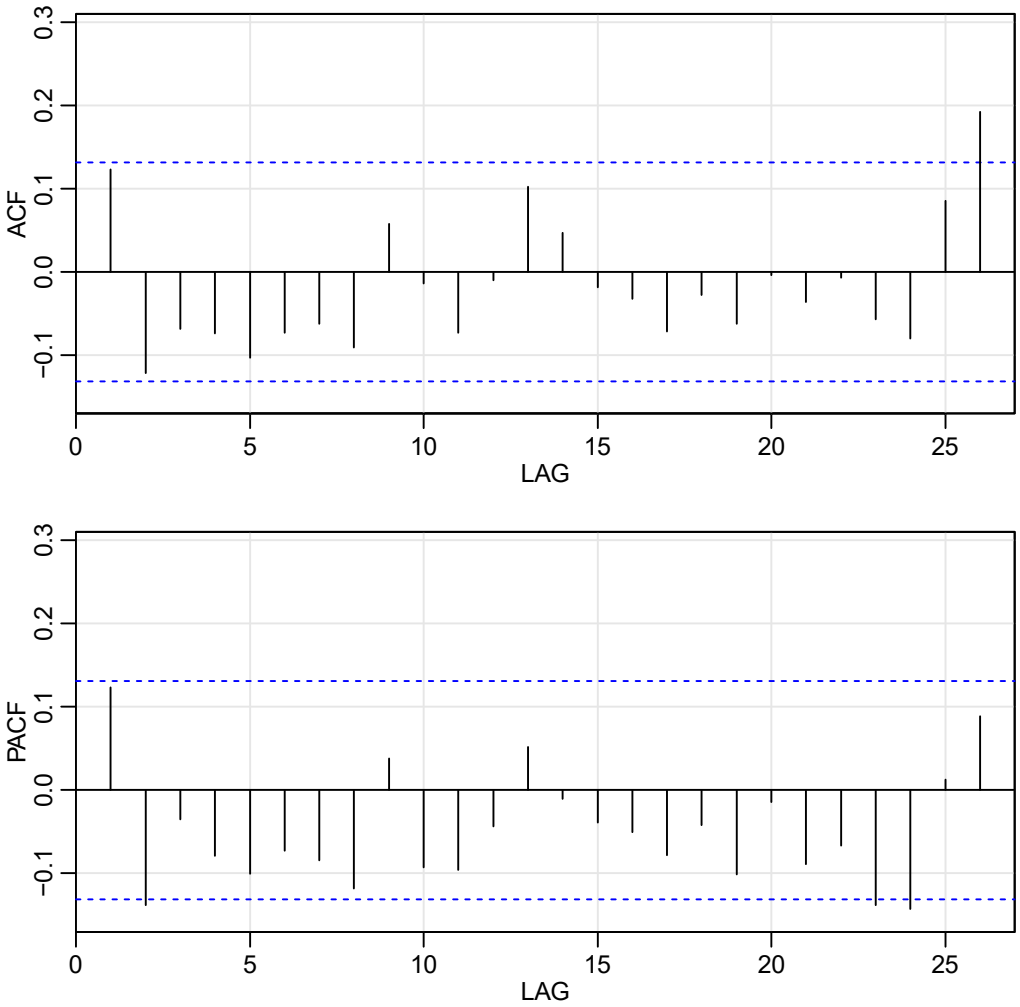

Figure 7: Autocorrelation function (upper panel) and partial autocorrelation function (lower panel) on the residuals of the unemployment claims pure time series model, 2012-2017 forecasts

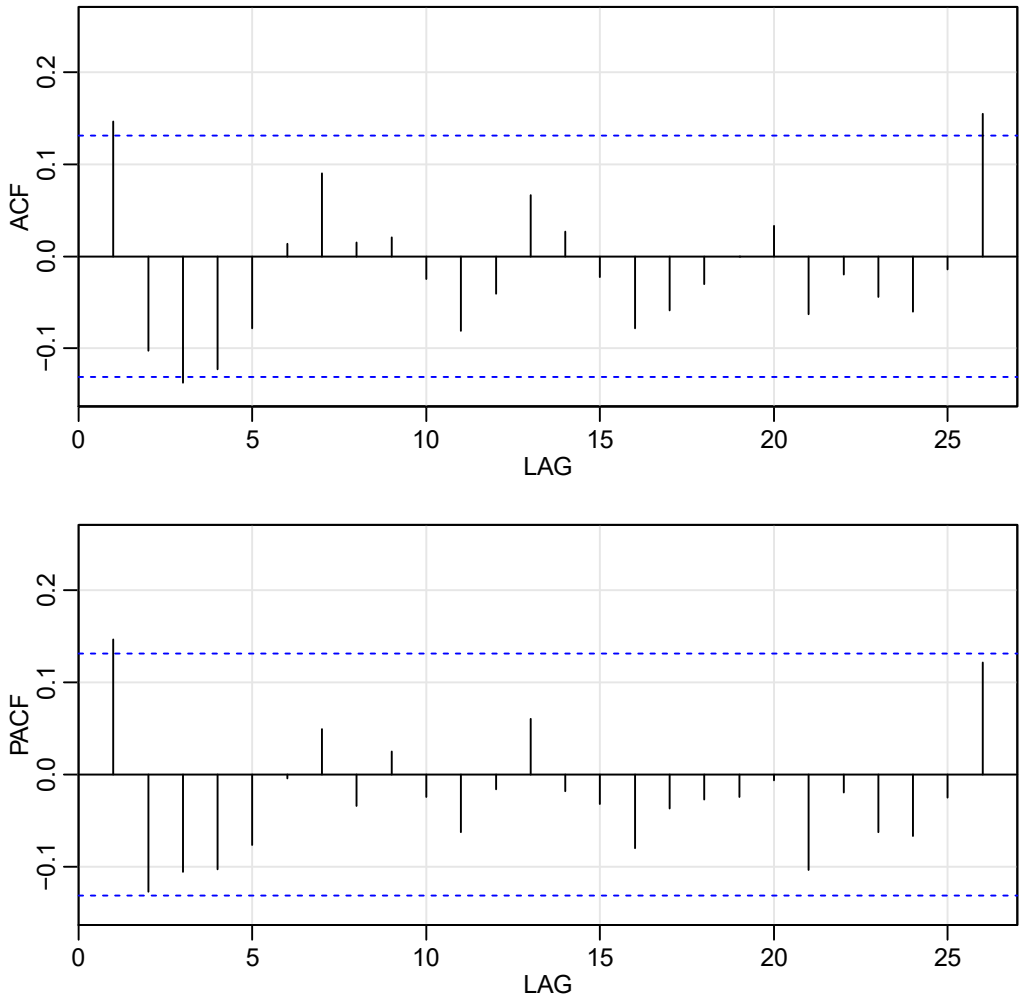

Figure 8: Autocorrelation function (upper panel) and partial autocorrelation function (lower panel) on the residuals of the unemployment claims correlate model, 2012-2017 forecasts

seem to perform worse in the retail sales case than in the U.S. unemployment claims example. The most important is the granularity of data. The unemployment data is weekly, whereas retail sales data is monthly. The finer resolution of the data allows for more observations, creating a larger effective sample size.

Another possible reason is that filing for an unemployment claim is a specific action that has a narrower set of possible Correlate search queries related to it. There are millions of different products that contribute to retail sales and can be sought online, whereas thinking of even a hundred different ways to search for information on how to file for unemployment benefits would prove a challenge to most people.

The third variable is temporal. People might search online for different products months before the purchase to gain information for the purchase decision or after years of ownership for service instructions. In contrast, there seems to be few intuitive reasons for people to conduct unemployment-related searches if they are not either economists or expecting to be unemployed in the immediate future. In this light the relatively poor performance of Google data in predicting retail sales is not as surprising as at first. But why did the retail sales analysis show good results in Scott and Varian (2014), when search engine data does not improve predictions in subsequent years? The next chapter examines a potential reason why the results of the original retail sales analysis might have appeared more impressive than the actual prediction performance of Google data would have warranted.

5 Performance disparities with Scott and Varian (2014)

5.1 Inflated errors in the U.S. retail sales pure time series model

The inability of Google data to help predict U.S. retail sales presented in section 4.2 contrasts with the results of Scott and Varian (2014). The question is, what is the cause of the performance differences? After all, figure 1 from the original study shows significant decrease in prediction errors when regression components with Google data are added to the pure time series model. A reasonable expectation would be that this difference in prediction performance would hold in out-of-sample nowcasting as well.

However, using the `rsxfs` dataset from the `bsts` R package (released by the same authors and at least visually the exact same data as that used by Scott and Varian) as an input to Google Correlate and the pure time series model produces results resembling the original study *only if a highly informative prior is used on the pure time series model.* It seems probable that the original study used a highly informative prior where a noninformative would have been more appropiate, inflating the prediction errors and increasing the perceived comparative advantage of the models with regression components.

Scott and Varian (2014, 11) mention the default hyperparameter values the `bsts` package used at the time of the writing of the paper in the case of the models with regression components. However, they do not mention any specific values for pure time series prior hyperparameters, only that the defaults in `bsts` were used (Scott and Varian 2014, 17). Using the current default priors in the `bsts` package greatly decreases the differences in performance of the different models, suggesting that at the time of the original paper's writing `bsts` used a different and highly informative prior.

In figure 9 the black pure time series graph with the informative prior, practically identical to the original pure time series graph in figure 1, is recreated by setting the prior on the residual standard errors to 1 and the

observation count of the prior to 100 as explained in section 3.6. Testing with different hyperparameter values showed that an observation count of 50, for example, was not sufficiently large to closely mimic figure 1. The model from equation (3) is used with the dropping of the regression component $\beta^T \mathbf{x}_t$.

The pure time series model with the noninformative prior is created by setting the prior guess on the residual standard errors to 1 and the prior observation count to 0.01, the default values in the current `bsts` package. As can be seen, as the impact of the prior on the posterior is decreased by making the prior observation count small the performance of the pure time series model is increased drastically and ends up on par with the models with regression components. As of the models with Google data, similarly noninformative prior hyperparameter values of expected $R^2 = 0.5$, $\nu = 0.01$ and $p = 5$ (the default values in the current version of the `bsts` package; see section 3.5) are used.

In light of figure 9 it seems that the prior used for the original pure time series model was highly informative and increased its prediction errors significantly. A large amount of the perceived advantage of using Google Correlate and Trends data looks like it was in fact caused by the prior on the pure time series model.

5.2 Modifying the regression model priors

Interestingly, figure 9 shows much smaller differences between the models with regression components as well, even though these models have the exact hyperparameter values cited by Scott and Varian (2014, 11). This can be caused by different input data. As explained in section 2.3, Google Correlate and Trends output differ somewhat from week to week, so reproducing the exact data used by the original research is impossible. There is now more Trends categories than when the original study was written, and the Trends categories used for the original study were not listed comprehensively, so the categories used for analysis here were selected either if they were explicitly mentioned by Scott and Varian (2014) or if they seemed to have economic relevance. The categories used were fetched with the `gtrendsR` package and

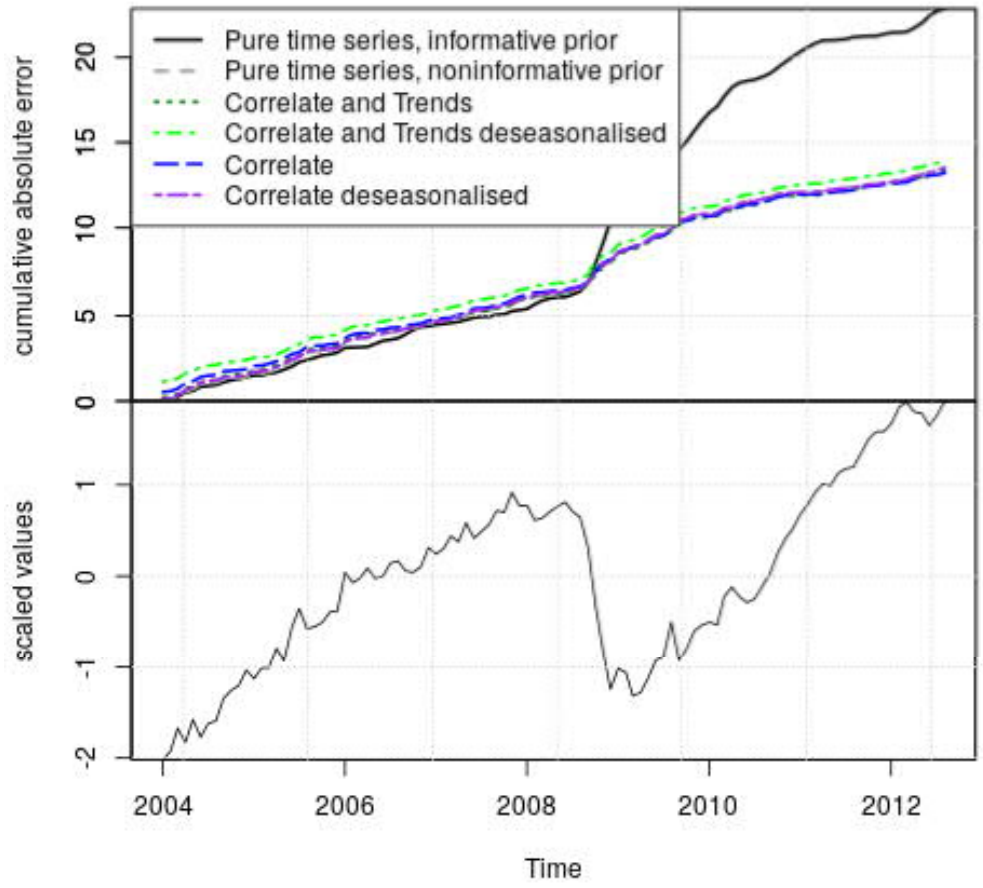

Figure 9: Replication of the U.S. retail sales analysis with both informative and noninformative priors in the pure time series models: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel)

are listed in table 5.

Table 5: Google Trends categories used in all `econ` models

Home Financing	Welfare & Unemployment
Social Services	Spas & Beauty Services
Real Estate	Computers & Electronics
Shopping	Shopping Portals & Search Engines
Autos & Vehicles	Auto Financing
Business Education	Home Insurance
Hybrid & Alternative Vehicles	Off-Road Vehicles
Motorcycles	Finance
Insurance	

Still, the problems with the priors in the pure time series case raise questions about the priors on the models with regression components as well. To illustrate that figure 1 could be more closely approximated by manipulating the hyperparameters of the prior distributions of the regression models, figure 10 is created by setting the expected R^2 of the model to 0.001, prior degrees of freedom ν to 10 and prior information weight κ to 2 (see section 3.5).

Figure 10 is by no means a perfect match to the original figure 1, but nonetheless resembles it more closely than figure 9. There are wider differences between different models, and deseasonalising seems to have a larger impact on performance. If desired, with individual tailoring of hyperparameters to each different model the match could be made closer still. Figure 10 serves the purpose of highlighting that the hyperparameters in the regression priors could also have been different at the time of the writing of the original study, but because (conversely to the pure time series model) the regression input data has changed substantially between the original analysis and the replication, further inference on the reasons of discrepancies in results is impossible.

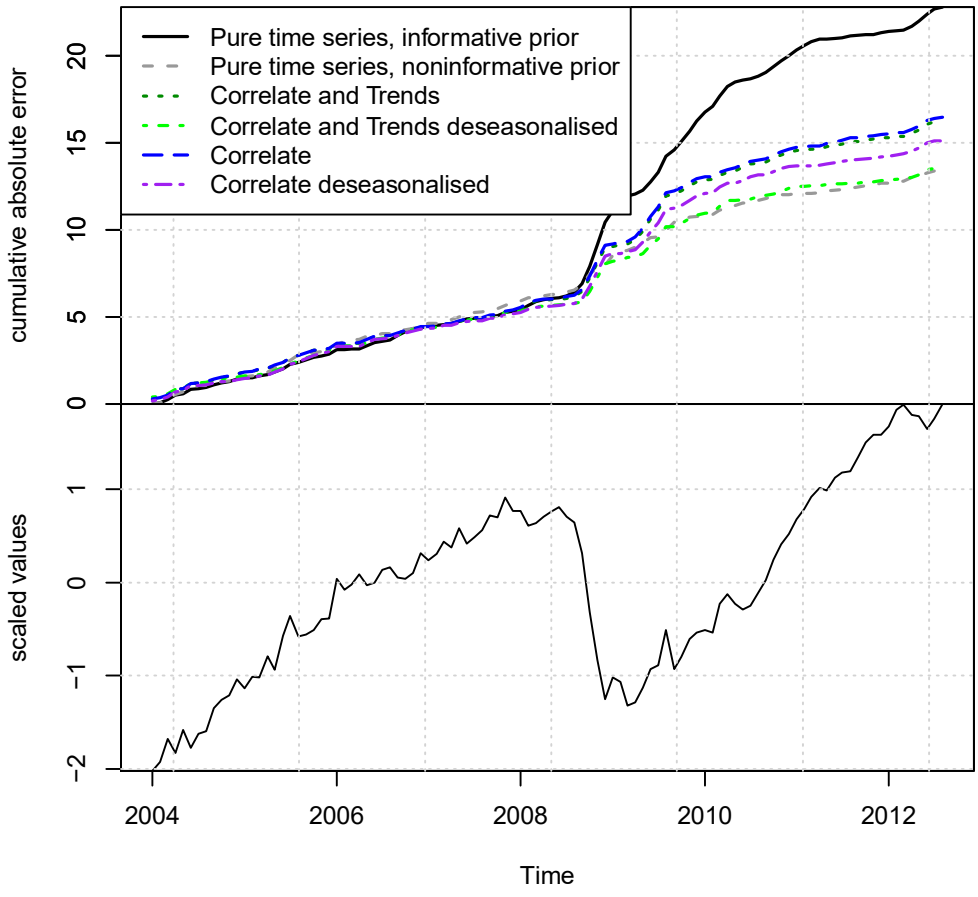

Figure 10: U.S. retail sales analysis with modified priors on regression models: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel)

5.3 Replicating the U.S. initial unemployment claims example

To demonstrate that the issues with the priors explained above are indeed in the original source and not in the code written for this paper, the first example in Scott and Varian (2014), nowcasting weekly initial U.S. unemployment claims with Google Correlate, is replicated basing the code on Scott (2017b) as in the retail sales analysis above. The data is obtained from the `bsts` package directly, but using data from the U.S. Employment and Training Administration as an input to Google Correlate yields identical results.

The original graph is shown in figure 11 and the replication in figure 12. As before, the black solid line in the upper part of the graphs shows the cumulative absolute errors of the pure time series model, and the dotted red line denoted "Google Trends" shows the cumulative absolute errors of the model with a Google Correlate regression component.

The replication seems to follow the original closely, but careful examination shows that the cumulative absolute errors of the pure time series model are indeed somewhat lower in figure 12 where an noninformative prior is used. This seems to communicate two things: the original analysis did indeed use informative priors that different from those that are currently in use in `bsts`, and the initial unemployment claims example is less sensitive to the prior than the retail sales example. The smaller sensitivity to the prior is probably explained by the larger sample size of the unemployment example (456 observations versus 104 in the retail sales example).

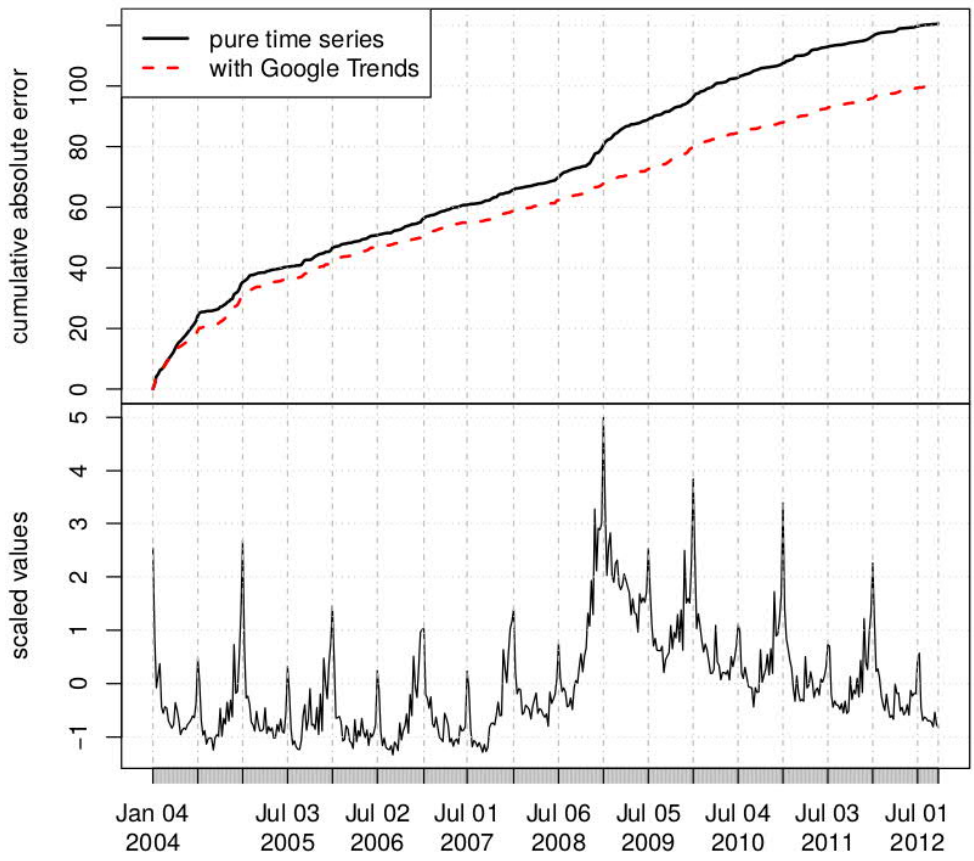

Figure 11: Original U.S. unemployment analysis: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel) from Scott and Varian, 2013

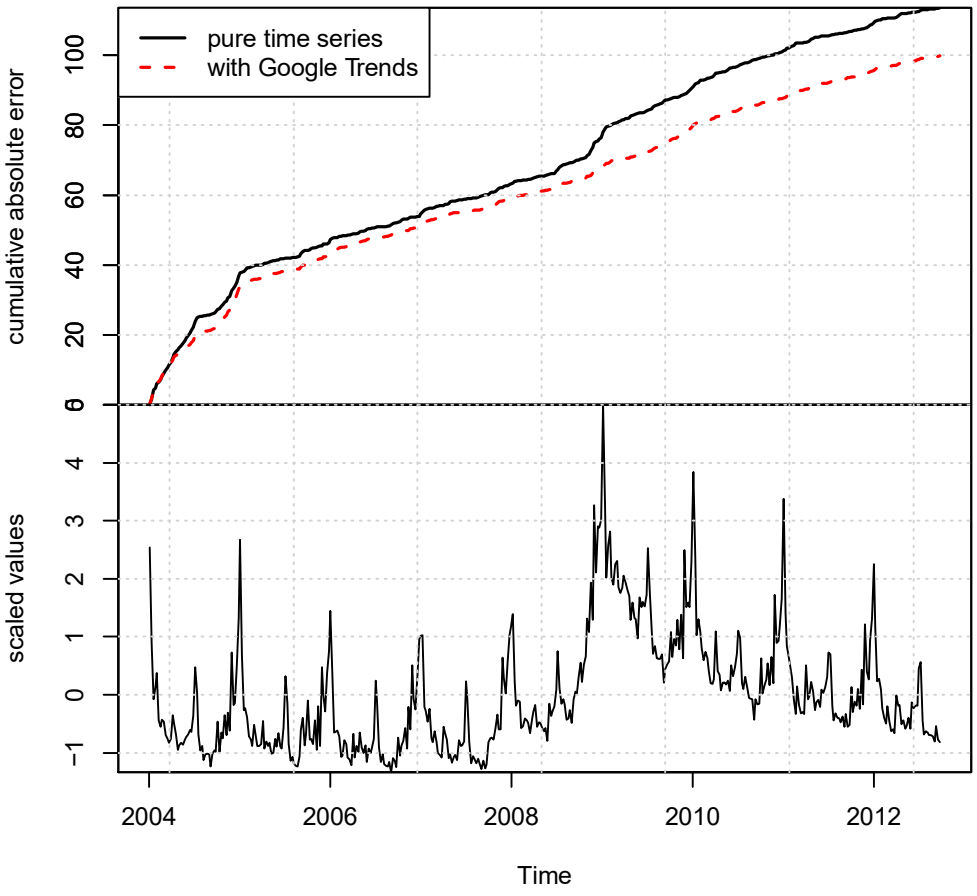

Figure 12: Replication of U.S. unemployment analysis: cumulative absolute prediction errors (upper panel) and the actual time series (lower panel)

6 Conclusions

In this thesis Bayesian structural time series using Google search query data is tested as a method for nowcasting U.S. retail sales and U.S. initial unemployment claims. Google data seems to improve prediction performance in the case of unemployment data, but no meaningful evidence that Google data can improve predicting U.S. retail sales is found. A potential cause, the usage of a highly informative prior in Scott and Varian (2014), is examined and the differences between the original analysis and this thesis are attributed to this hypothesized usage of a highly informative prior in the former.

Several possible opportunities for further research can be found. One possible avenue is examining refining the U.S. initial unemployment claims model, as the analysis in 4.3 identified autocorrelation in the residuals of the model over the test period. Another potential research area could be testing how well the positive results of Kholodilin et al. (2010) and Vosen and Schmidt (2011) in nowcasting private consumption would fare after the release of the studies and whether or not their methods would improve upon the ones detailed in this thesis. While they examine private consumption, not retail sales, their methods could still be useful in the latter case.

There are numerous additional studies where Google search data is used to forecast different time series, and the efficacy of Bayesian structural time series could be compared to the methods used in these papers as well. In addition to replication efforts, novel ways to employ Bayesian structural time series and Google data can be found in any subject where a connection between the statistic and search engine usage is at least plausible.

References

Artola, Concha – Pinto, Fernando – Pedraza Garcìa, Pablo de (2015) Can internet searches forecast tourism inflows? *International Journal of Manpower*, vol. 36 (1), 103–116.

Askitas, Nikolaos – Zimmermann, Klaus F. (2009) Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, vol. 55 (2), 107–120.

Brodersen, Kay H. – Gallusser, Fabian – Koehler, Jim – Remy, Nicolas – Scott, Steven L. (2015) Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, vol. 9 (1), 247–274.

Carrière-Swallow, Yan – Labbé, Felipe (2013) Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, vol. 32 (4), 289–298.

Chang, Andrew C. – Li, Phillip (2015) Is economics research replicable? sixty published papers from thirteen journals say "usually not". *Board of Governors of the Federal Reserve System (U.S.), Finance and Economics Discussion Series: 2015-83, 2015, 25 pp.*, 25.

Da, Zhi – Engelberg, Joseph – Gao, Pengjie (2011) In search of attention. *The Journal of Finance*, vol. 66 (5), 1461–1499.

Durbin, James – Koopman, Siem Jan (2001) *Time series analysis by state space methods*. Oxford: Oxford University Press.

Duvendack, Maren – Palmer-Jones, Richard – Reed, W. Robert (2017) What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, vol. 107 (5), 46–51.

Einav, Liran – Levin, Jonathan (2014) Economics in the age of big data. *Science*, vol. 346 (6210).

Goel, Sharad – Hofman, Jake M. – Lahaie, Sébastien – Pennock, David M. – Watts, Duncan J. (2010) Predicting consumer behavior with web search. *Proceedings of the National academy of sciences*, vol. 107 (41), 17486–17490.

Google Correlate (2011). Google Inc. `https://www.google.com/trends/correlate`.

Google Trends. Google Inc. `https://trends.google.com/trends/`.

Greenberg, Edward (2013) *Introduction to Bayesian econometrics*. 2nd ed. New York: Cambridge University Press.

Inoue, Atsushi – Kilian, Lutz (2006) On the selection of forecasting models. *Journal of Econometrics*, vol. 130 (2), 273–306.

Kholodilin, Konstantin A. – Podstawski, Maximilian – Siliverstovs, Boriss (2010) Do Google searches help in nowcasting private consumption? a real-time evidence for the US. *KOF Swiss Economic Institute Working Paper No. 256; DIW Berlin Discussion Paper No. 997*.

Kristoufek, Ladislav (2013) Can Google Trends search queries contribute to risk diversification? *Scientific Reports*, vol. 3 (2713).

Lampos, Vasileios – Miller, Andrew C. – Crossan, Steve – Stefansen, Christian (2015) Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, vol. 5 (12760).

Liu, Paul – Fabbri, Marco (2016) *More eyes, (no guns,) less crime: estimating the effects of unarmed private patrols on crime using a Bayesian structural time-series model*. Social Sience Research Network. `https://ssrn.com/abstract=2739270`, accessed 27.12.2017.

Massicotte, Philippe – Eddelbuettel, Dirk (2018) *gtrendsR: perform and display Google Trends queries*. The Comprehensive R Archive Network. `https://cran.r-project.org/web/packages/gtrendsR/index.html`, accessed 26.2.2018.

Mohebbi, Matt – Vanderkam, Dan – Kodysh, Julia – Schonberger, Rob – Kumar, Hyunyoung Choi Sanjiv (2011) *Google Correlate whitepaper*. Google Inc. `https://www.google.com/trends/correlate/whitepaper.pdf`, accessed 7.1.2018.

Nielsen, Frank – Garcia, Vincent (2009) Statistical exponential families: a digest with flash cards. *CoRR*, vol. abs/0911.4863.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, vol. 349 (6251).

Preis, Tobias – Reith, Daniel – Stanley, H. Eugene (2010) Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 368 (1933), 5707–5719.

Scott, Steven L. (2017a) *Bsts: Bayesian structural time series R package, version 0.7.1*. The Comprehensive R Archive Network. `https://cran.r-project.org/package=bsts`, accessed 10.12.2017.

Scott, Steven L. (2017b) *Fitting Bayesian structural time series with the bsts R package*. The Unofficial Google Data Science Blog. `http://www.unofficialgoogledatascience.com/2017/07/fitting-bayesian-structural-time-series.html`, accessed 20.11.2017.

Scott, Steven L. (2018) *Bayesian object oriented modeling*. The Comprehensive R Archive Network. `https://cran.r-project.org/package=Boom`, accessed 27.5.2018.

Scott, Steven L. – Varian, Hal R. (2013) *Predicting the present with Bayesian structural time series.* `https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41335.pdf`, accessed 6.2.2018. Draft version of Scott and Varian (2014) used for sourcing graphs with colour.

Scott, Steven L. – Varian, Hal R. (2014) Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, vol. 5 (1/2), 4–23.

Scott, Steven L. – Varian, Hal R. (2015) Bayesian variable selection for nowcasting economic time series. *Economic Analysis of the Digital Economy.* National Bureau of Economic Research. University of Chicago Press, 119–135.

Sivia, Devinderjit – Skilling, John (2006) *Data analysis. A Bayesian tutorial.* 2nd ed. Oxford: Oxford University Press.

Stigler, Stephen M. (2016) *The seven pillars of statistical wisdom.* Cambridge, Massachusetts: Harvard University Press.

U.S. Bureau of the Census (2018) *Retail sales: retail (excluding food services) [mrtssm44000uss].* Federal Reserve Bank of St. Louis. `https://fred.stlouisfed.org/series/MRTSSM44000USS`, accessed 3.5.2018.

U.S. Employment and Training Administration (2018) *Initial claims [icnsa].* Federal Reserve Bank of St. Louis. `https://fred.stlouisfed.org/series/ICNSA`, accessed 1.4.2018.

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