Assessing the Use of Google Trends Search Query Data to Forecast Number of Nonresident Hotel Registrations in Puerto Rico

Roberto Rivera*

Abstract
Recently, studies have used search query volume (SQV) data to forecast a given process of interest. However, Google Trends SQV data comes from a periodic sample of queries. As a result, Google Trends data is different every week. We propose a Dynamic Linear Model that treats SQV data as a representation of an unobservable process. We apply our model to forecast the number of hotel nonresident registrations in Puerto Rico using SQV data downloaded in 11 different occasions. The model provides better inference on the association between the number of hotel nonresident registrations and SQV than using Google Trends data retrieved only on one occasion. However, compared to simpler models we only find evidence of better performance when making forecasts on a horizon of over 6 months.

Key Words: combining multiple time series, dynamic linear model, hotel registrations, search query volume data, Google Trends, forecasting models

1. Introduction

In recent years, there has been an interest in exploiting search query data available through sources such as Google Trends (www.google.com/trends) to model temporal processes. Choi and Varian (2009a,b) used search query data to model tourism demand, auto sales, home sales, and initial unemployment claims. Ginsberg et al. (2009) relied on Google search queries to model influenza activity in the U.S. Studies have also suggested search query based tools to model consumer behavior (Goel et al. 2010), dengue (Gluskin et al. 2014) and more. It is not exactly known how the search query volume algorithm by Google generates its results. Moreover, the time series of search query volume generated by the algorithm changes every week.

Puerto Rico has been going through an economic recession since 2006. Leaders on the island have been attempting to find ways to boost the economy. Although hotel registrations from July to November showed an increase of about 10% from fiscal year 2012 to 2013 (Junta de Planificación de Puerto Rico 2013a), over the long term the contribution of the hotel industry to Gross Domestic Product has stayed relatively constant (Ruiz 2012). With opportunities in many sectors of the economy dwindling, the government has been taking steps to improve the tourism sector. To accomplish this, efficient planning is crucial. Statistical inference can be used to forecast the number of hotel registrations by nonresidents, a proxy of tourism demand.

This is the first study to treat each weekly Google Trends output as a source of data of an unobservable process. We use this data to draw inference on the lagged association between the number of hotel nonresident registrations (NHNR) in Puerto Rico and search query volume (SQV). The performance of our Dynamic Linear Model in forecasting NHNR is compared to alternative models.

*Business department, University of Puerto Rico, Mayaguez, Email: roberto.rivera30@upr.edu

2. Data

Number of hotel nonresident registrations from January 2004 to September 2012 was provided by the Puerto Rico Tourism Company. Hotels and luxury hotels are required to provide registration data while short term stays and guest houses can provide it if they wish to do so. Although NHNR does not exactly measure the number of tourists that come to the island, it intuitively serves as a good proxy. A publicly available tool called Google Trends provides an index of relative volume of search queries based on a percentage of Google web searches. The data quantifies the standardized volume of searches for a given query and aggregates them, typically over 7 days. We emphasize the use of the word 'standardized' here, meaning Google Trends search query volume relative to the total number of searches done on Google over time, instead of absolute search volume. Thus, if the rate of absolute search query volume increase is smaller than the total search query volume, relative search query volume may decrease. Therefore the standardized SQV obtained are dependent on the region, category/subcategory, queries, and time frame selected. To fit the model, the SQV Google Trends data, provided on a 7 day scale, was converted to monthly data.

2.1 Challenges in using Google SQV data

As appealing as the availability of the search query data is, care must be taken. Butler (2013) found that Google Flu trends, a search query based tool, was not been performing as well as when it was introduced in 2009, sometimes estimating twice as many actual influenza cases. More recently, Lazer et al. (2014) showed that from August 21, 2011 to September 1, 2013, Google Flu Trends reported overly high flu prevalence 100 out of 108 weeks. Screening the search query data one finds that, for a fixed period of time of interest, fixed search queries and a fixed region, Google query output will differ over the time series. For example, if one is to obtain today data from Google Trends from 2004 to 2014 for "puerto rico hotels" performed in the United States, one would obtain a time series of results. However, if one would extract output under the same parameters next week, the time series has different entries. Every week the output will be different. This is different than data revision of economic data where only the most recent data changes. In the case of Google Trends, data at all time points change routinely. The issue is partly due to how the SQV data is provided through Google Trends. According to the help page of Google Trends, the companies algorithm analyzes a percentage of Google web searches to determine the amount of searches for the terms entered compared to the total number of Google searches done during the same time period. The statement implies that SQV is based on a sample of Google searches, but it doesn't specify the sample size or how samples are chosen. Another possible explanation is the fact that Google constantly changes its search algorithm (Lazer et al. 2014). Among recent changes, the use of social networking data and predicting misspellings to determine search results for users. Another aspect is that Google constantly changes its algorithm to determine search query volume. For example, nowadays Google Trends tool provides suggestions while keying in search queries, avoiding the possibility of misspellings and ambiguity in some terms. The challenges presented here are not to say that the search query data is not useful. One way of seeing it, is that the search query data provided by Google, is an observed version of the true search query process.

2.2 Choice of Google Trends Parameters to obtain SQV data

Search query volume data acquired from Google trends is a function of a series of settings the user determines (e.g. region or location where searches were made, categories and subcategories of the search queries, the search type, etc.). We treated each of these Google Trends optional settings as parameters. The region to obtain the search volume data was chosen to be the United States. Including other countries would likely blur the association between Google SQV and NHNR for the following reasons. First, most nonresident tourists come from the United States. According to Puerto Rico's Tourism Company, for the 2011 fiscal year 92.6% of visitors surveyed came from the United States (Junta de Planificación de Puerto Rico 2013b). Of visitors arriving from the U.S. 44.1% came from the East coast, most from New York and Florida. Secondly, Google's search market share overall is large, but it may vary considerably by country. Although no official numbers of search market share exists, estimates from several companies indicate that Google's market share is lower than local alternatives in some countries (e.g. South Korea, China, Russia, and Japan). Lastly, exploration of queries related to Puerto Rico travel from countries other than the U.S. often produced little search volume data, sometimes no data at all. Table 1 summarizes the settings we used while using Google Trends. Using feedback from experts at the tourism company and preliminary analysis, it was determined that the 9 queries shown where the best alternatives to forecast NHNR without exceeding the 30 word limit that Google Trends permits. Only Web search type volume was used from the Travel category. This Google Trends Travel category contains subcategories, but the search volume data for our queries of interest is spotty within these subcategories, so no subcategories were selected.

Search Volume Parameter	Parameter Setting
Queries	puerto rico hotels, puerto rico flights, san juan hotels, puerto rico resorts, puerto rico vacations, puerto rico vacation, puerto rico tourism, puerto rico travel, and puerto rico hotel deals
Region	United States
Search time frame	January 2004 - January 2014
Search type	Web Search
Category	Travel
Subcategory	NONE

Table 1: Search Volume parameter settings used with Google Trends to gather data to construct models to forecast the number of hotel nonresident registrations in Puerto Rico. Volume data gathered every Thursday from 10/2/14 to 12/11/14.

We used the results of these queries in the period from January 2004 to January 2014, but to fit the models we only used the time frame for which we have room registration data. Finally, SQV data was extracted in 11 consecutive Thursdays, from October 2 to December 11, 2014.

In the next section we discuss the models considered to study the capacity of query volume data to improve forecasts of NHNR.

3. Forecasting Models

We can express the NHNR data in a rather ambiguous form:

$$Y = g(\mu, S, \epsilon) \tag{1}$$

That is, the data is decomposed into a trend or an association with search traffic component (modeled through μ), seasonality component (S), and some irregular time dependence component (ϵ). $g(\cdot)$ determines the type of function of these components. We model each component in an additive way based on stochastic and deterministic approaches and, for one of the models, we let the data determine if the relationship among each component and the process should be linear or nonlinear.

3.1 Dynamic Linear Model

The Dynamic Linear Model (DLM) is a flexible way to intuitively capture how processes evolve in time. In fact, traditional time series models such as ARIMA and others can be viewed as special cases of the DLM. Yet the dynamic linear model can also incorporate nonstationarity, time-varying parameters, multivariate time series, data from multiple sources, irregular temporal observations, and missing data among other things. Shumway and Stoffer (2011); Chatfield (2003) provide nice introductions to DLMs while Durbin and Koopman (2012); Brockwell and Davis (2009) cover more advance theory on the subject. DLMs have been widely used to model environmental data (Cressie and Wikle (2011), Huerta et al. (2004)), and economic or financial data Shumway and Stoffer (2011). But DLM models have received much less attention in other business applications, and although it has been applied to model tourism data (Athanasopoulos and Hyndman 2008; du Preez and Witt 2003), they have not been applied to Google Trends data as done in this study. Let $\boldsymbol{Y}_t = (Y_{1,t}, Y_{2,t}, ..., Y_{m,t})'$ represent observations of m time series at time t. Hence each $\boldsymbol{Y}_t$ is a $m \times 1$ vector. Furthermore, let $\boldsymbol{X}_t = (X_{1,t}, ..., X_{q,t})'$ be the true q processes of interest, and $\boldsymbol{S}_t = (S_{1,t}, ..., S_{1,t-k}, ..., S_{q,t}, ..., S_{q,t-k})'$ represents the seasonal component of period k for each of the q processes in the model. We express DLM with the following equations:

$$\boldsymbol{Y}_t = \boldsymbol{F}\boldsymbol{X}_t + \boldsymbol{H}\boldsymbol{S}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{V}) \tag{2}$$

$$\boldsymbol{X}_t = \boldsymbol{G}^{(x)}\boldsymbol{X}_{t-1} + \boldsymbol{C}\boldsymbol{S}_t + \boldsymbol{\omega}_t^{(x)}, \qquad \boldsymbol{\omega}_t^{(x)} \sim N(\boldsymbol{0}, \boldsymbol{W}_t^{(x)}) \tag{3}$$

$$\boldsymbol{S}_t = \boldsymbol{G}^{(s)}\boldsymbol{S}_{t-1} + \boldsymbol{\omega}_t^{(s)}, \qquad \boldsymbol{\omega}_t^{(s)} \sim N(\boldsymbol{0}, \boldsymbol{W}^{(s)}) \tag{4}$$

Equation (2) is known as the observation or measurement equation, where $\boldsymbol{\nu_t}$ corresponds to Normally distributed measurement error with mean zero and covariance $\boldsymbol{V}$. $\boldsymbol{X}_t$ is referred to as the state or system vector (West and Harrison 1997) and contains all the parameters that relate to the trend of the temporal processes of interest. The set of equations imply that the state vector of interest $\boldsymbol{X}_t$ cannot be observed directly. $\boldsymbol{F}$ is a $m \times q$ matrix that may depend on parameters that need to be estimated. $\boldsymbol{H}$, $\boldsymbol{C}$ are matrices with dimension and entries depending on whether the seasonal component is modeled as a fixed effect or stochastically (see section 3.1.1 for details on our approach).

Equations (3) and (4) are known as state, system, or transition equations. These equations determine how $\boldsymbol{X}_t$ is generated from past values $\boldsymbol{X}_{t-1}$. $\boldsymbol{G}_t^{(s)}$ and $\boldsymbol{G}_t^{(x)}$ are referred to as the evolution matrices with dimensions $s \times s$ and $q \times q$ respectively, and $\boldsymbol{\omega}_t^{(s)}$, $\boldsymbol{\omega}_t^{(x)}$ are the evolution errors. As stated in Banerjee et al. (2004), usually the design problem at hand determines the form of $\boldsymbol{F}$ while modeling assumptions lead to how $\boldsymbol{G}^{(s)}, \boldsymbol{G}^{(x)}$ are represented. Specifically, dependence among $(Y_{1,t}, Y_{2,t}, ...Y_{m,t})'$ can be introduced into the model through $\boldsymbol{G}^{(s)}, \boldsymbol{G}^{(x)}$ and/or $\boldsymbol{W}_t^{(x)}, \boldsymbol{W}^{(s)}, \boldsymbol{V}$. Choosing the identity matrix as $\boldsymbol{G}^{(x)}$ results in a random walk representation for $\boldsymbol{X}_t$ for all t. More generally $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{G}^{(s)}, \boldsymbol{G}^{(x)}$ may be time dependent sequence of matrices, an extension that we do not pursue here.

3.1.1 Dynamic Linear Model for NHNR

To adapt the DLM to our case let $\boldsymbol{Y}_t = (Y_{1,t}, Y_{2,t}, ..., Y_{(a+1),t})'$ where $Y_{1,t}$ is the recorded number of hotel nonresident registrations for time t and $Y_{2,t}, ..., Y_{(a+1),t}$ are query volume data retrieved from Google Trends from their algorithm runs $1, ..., a$ for time t. Hence each $\boldsymbol{Y}_t$ is a $(a+1) \times 1$ vector. Of prime importance to us is the true process $\boldsymbol{X}_t = (X_{1,t}, X_{2,t})'$ where $X_{1,t}$ is the true NHNR for time t and $X_{2,t}$ is the true SQV data for time t. Assuming a seasonal component of period $k = 12$, which we model as a fixed factor, our DLM consists of the following equations in matrix form:

$$\begin{pmatrix} Y_{1,t} \\ \vdots \\ Y_{(a+1),t} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} + \begin{pmatrix} \nu_{1,t} \\ \vdots \\ \nu_{(a+1),t} \end{pmatrix} \tag{5}$$

where the observation errors are assumed to be independent from the state vector for all t, and no correlation is assumed between the observation errors $\nu_{1,t}$ and $\nu_{j,t}, j = 2, ..., (a+1)$. Conversely, correlation between $\nu_{j,t}, j = 2, ..., (a+1)$ are possible. However, not constraining the correlation between all $\nu_{j,t}, j = 2, ..., (a+1)$ requires the estimation of too many[1] off-diagonal parameters in $\boldsymbol{V}$. Therefore we assume $\boldsymbol{V} = diag(\sigma_1^{2(y)}, \sigma_2^{2(y)} I)$ where I is an identity matrix. The other equation looks as follows,

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \boldsymbol{C}\boldsymbol{S}_t + \begin{pmatrix} \omega_{1,t}^{(x)} \\ \omega_{2,t}^{(x)} \end{pmatrix} \tag{6}$$

where variances $\boldsymbol{W}_t^{(x)} = diag(\sigma_1^{2(x)}, \sigma_{t,2}^{2(x)})$, and the errors $\{\boldsymbol{\nu}_t\}, \{\boldsymbol{\omega}_t^{(x)}\}$ are uncorrelated. Note that (5) and (6) imply that the Seasonal component for each temporal process is modeled as fixed with $\boldsymbol{H} = \boldsymbol{0}$, $\boldsymbol{S}_t$ indicating the month at time t, $\boldsymbol{C}$ is a 2×12 matrix of parameters, and no stochastic component ($\boldsymbol{W}^{(s)} = diag(\sigma_1^{2(s)}, \sigma_2^{2(s)}) = 0$). The β parameter is linked to the linear association between NHNR at time t and SQV at time $t-1$. $\beta \neq 0$ would indicate that there is a linear association between $X_{2,t-1}$ and $X_{1,t}$ while $\beta = 0$ would indicate no linear association between these processes and hence, no practical use of SQV in forecasting NHNR. $\beta \neq 0$ implies that the true SQV is a leading indicator of NHNR. Leading indicators are useful in forecasting processes of interest, since they don't have to be forecasted themselves for short lead times. In this case, however, $X_{2,t-1}$ is not directly observed making its usefulness as a leading indicator less clear. This DLM allows

[1] In our data $a = 11$, hence not constraining the correlation between all $\nu_{j,t}, j = 2, ..., 12$ requires the estimation of $(121 - 11)/2 = 55$ off-diagonal parameters in $\boldsymbol{V}$. Moreover, a fixed covariance among all 10 search query output errors did not improve the model.

us to account for the information from multiple Google Trends algorithm runs to determine if there is a linear association between $X_{2,t-1}$ and $X_{1,t}$. We hypothesized that incorporating data from multiple search volume algorithms had an impact in determining the type of association between the true NHNR and the true SQV. We tested this hypothesis by comparing the inference on β from the DLM expressed above and a DLM using only the most recent search query algorithm data.

3.1.2 Estimation of parameters and Kalman recursions

Parameters in equations (5), (6) must be estimated. Direct Maximum likelihood estimation methods (Brockwell and Davis 2009), Expected Maximization (EM), and Bayesian methods (West and Harrison 1997) are some alternatives. In this work we used the EM algorithm described in (Holmes 2012) and implemented through the R package MARSS (Holmes et al. 2012). Confidence intervals for β were based on asymptotic Normality and an estimated Hessian matrix. Predictions $\hat{X}_{i,t'}$ and $\hat{Y}_{i,t'}$ at times t' for $i = 1, 2$ were obtained using Kalman Recursions (Brockwell and Davis 2009) through the DLM presented in this paper.

3.2 Other forecasting models

A Seasonal Autoregressive Integrated Moving Average (SARIMA), Holt-Winter, and a type of nonparametric additive model were also fit. Shumway and Stoffer (2011) discusses the SARIMA model while Chatfield (2003) briefly explains Holt-Winter models. SARIMA and Holt-Winter models have been used in the past to model tourism arrivals and they tend to perform well (du Preez and Witt 2003; Lim and McAleer 2001). For a review of recent tourism demand modeling approaches see Song and Li (2008). Our nonparametric additive model is a bit less traditional and we briefly describe it in what follows.

Additive models are a form of nonparametric model that decomposes g in (1) into separate unknown smooth functions of the covariates (Paciorek 2007; Hastie and Tibshirani 1986),

$$Y_t = \sum_{j=1}^{p} g_j(z_{tj}) + \epsilon_t$$

where z_{tj} is the t^{th} observation of covariate j. Specifically, every function $g_j(\cdot)$ is conveyed as a linear combination of basis functions such as splines, wavelets, or polynomials (Hastie et al. 2009). Each function can be fit using a penalized least squares criterion (Ruppert et al. 2003). Splines, break the explanatory variable range into mutually exclusive regions and expresses the fit of each $g_j(\cdot)$ in terms of low order piecewise polynomials. When the covariate is numeric, a cubic spline is often used to estimate $g_j(\cdot)$. A cubic spline is a curve constructed by sections of cubic polynomials (with two continuous derivatives) and these sections are joined through knots. At the knots the function value, as well as the first and second derivatives of the piecewise polynomials match. Cubic splines are frequently used because of their good approximation properties. Conventionally, each datum determines a knot, but to simplify computations a low rank spline method can be used. In this work we fit the semiparametric model,

$$Y_{1,t} = \alpha + \beta\tilde{Y}_{2,t-1} + g_1(t) + g_2(month) + \epsilon_t$$

where *month* is month of year, $\tilde{Y}_{2,t}$ is the average search volume over the 10 algorithm outputs, g_1 is a thin plate regression spline, and g_2 is a cyclic cubic regression spline (Wood 2006). g_1 captures the overall trend in time while g_2 captures the seasonality in the data. For theoretical relationships of dynamic linear models with ARIMA models, Holt-Winter and spline based methods see Durbin and Koopman (2012). All models were fit using R (R Core Team 2012).

4. Results

Over the broader time period from January 2000 to December 2012, peak number of hotel nonresident registrations occurred in 2012 (1,575,131), while 2010 and 2011 had an increase slightly above 5% from the previous year. However, from 2005 to 2009, a yearly decrease in NHNR occurred. The lowest registrations since 2003 occurred in 2009, likely related to the financial crisis in the U.S. From January 2004 until September 2012, a potential subtle nonlinear trend was detected on monthly NHNR. As expected, the hotel room registration data displayed a strong seasonal pattern (upper left panel Figure 1). Highest NHNR occurred around the dry season months with a peak in March while lowest NHNR occurred in the wet months with September providing the lowest occupancy (right panel Figure 1). In fact, seasonality dominates the time series, suggesting that the Holt-Winter's model is a viable option to generate forecasts of NHNR. The seasonality did not appear to vary widely on a year to year basis. Furthermore, there was moderate variability in the data when seasonality was accounted for (see upper right panel of Figure 1), therefore it is uncertain how an Additive Model will perform. Based on the periodical peaks seen in the sample autocorrelation (ACF) and the high correlation at small lags (see left panel Figure 2), plus the quick decay observed in the partial autocorrelation (PACF) plots after lag 12 and the high correlation at small lags seen here as well (right panel Figure 2), a SARIMA $(1, 0, 1) \times (1, 0, 0)_{12}$ appeared adequate. Further analysis using Akaike Information Criteria (AIC) supported choosing this number of parameters (Akaike 1973).

Turning to the search query volume data, the lower left panel of Figure 1 shows the time series obtained October 9, October 23 and December 11 of 2014. Although each time series was similar, variability among algorithm dates are visible. December 11 output had more pronounced seasonal peaks than the other Google Trends output, especially later in the time series. On the other hand October 9 and October 23 output displayed lower seasonal bottoms than December 11 output. Time series were prewhitened as suggested in Bisgaard and Kulahci (2011) to inspect cross correlation. As we can see from Figure 1 (lower right panel), there appears to be a significant one month lag association between the NHNR and SQV time series. However, the lag-1 cross correlation was estimated to be 0.32, which makes the benefit of using the chosen SQV data to forecast NHNR questionable. Moreover, when omitting seasonality neither of the two temporal processes displayed major changes in average value over time nor a strong increasing or decreasing trend.

4.1 DLM results to determine association between $X_{1,t}$ and $X_{2,t-1}$

We fitted two DLMs as presented in sections 3.1.1 and 3.1.2 to draw inference on β. The multivariate time series $\boldsymbol{Y}_t$ was demeaned before constructing the DLM models. For one model, DLM_1, we used Google Trends generated time series of search query volumes for $a_1 = 11$ weeks. The second DLM model, DLM_2, used only

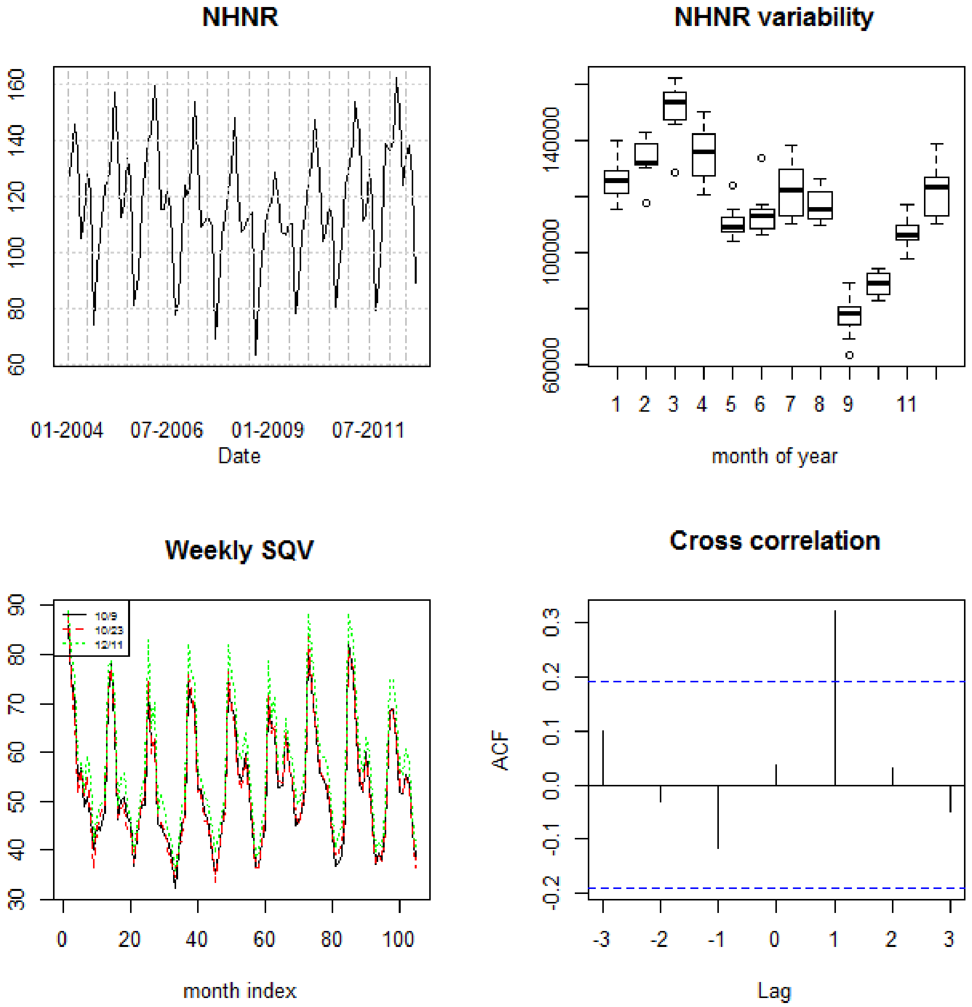

Figure 1: Upper left panel displays the Time series plot of number of hotel non-resident registrations (in thousands). Upper right panel shows boxplots of NHNR summaries by month of year. Output from 3 separate search query volume output can be seen in the lower left panel. Cross correlation based on prewhitened NHNR and prewhitened SQV (averaged over the 11 algorithmic outputs) are presented in the lower right panel.

the most recent Google Trends search query volume data $a_2 = 1$. Table 2 shows the resulting estimates of β, 95% confidence intervals based on asymptotic Normality and forecast accuracy measures mean absolute error (MAE), and mean absolute percentage error (MAPE) using one step ahead forecasts. We see that the estimate of β through DLM_2 was only about 1% smaller than through DLM_1. However, the β confidence interval based on DLM_1 did not include zero while the one based on DLM_2 did. Furthermore, DLM_1 had smaller length than the confidence interval based on DLM_2. By using output of 10 Google trends algorithm runs we can better infer about the linear association between $X_{1,t}$ and $X_{2,t-1}$. No difference was detected on the inference drawn from both models regarding $\sigma_1^{2(y)}$ and $\sigma_1^{2(x)}$. But, the results on $\sigma_2^{2(y)}$ and $\sigma_2^{2(x)}$ from DLM_1 imply that the search query volume is

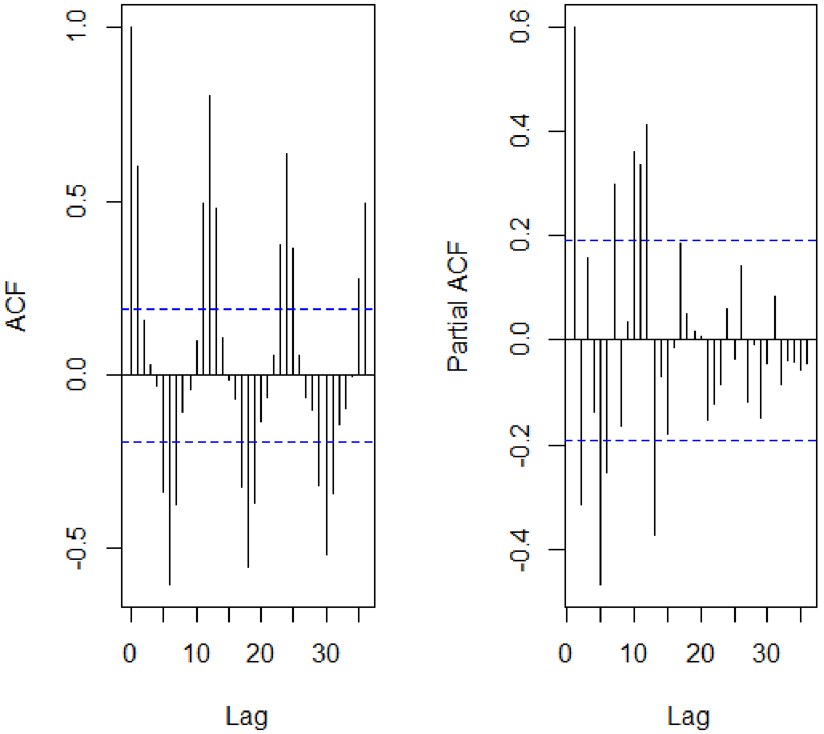

Figure 2: ACF and PACF plots for NHNR. Plots suggest a $(1,0,1) \times (1,0,0)_{12}$ SARIMA model

on average evolving in time while the results of DLM_2 put this in doubt with a confidence interval lower bound closer to zero.

	Parameter estimates					Forecast accuracy	
Model	β	$\sigma_1^{2(y)}$	$\sigma_2^{2(y)}$	$\sigma_1^{2(x)}$	$\sigma_2^{2(x)}$	MAE	$MAPE$
DLM_1	104.56	1.25×10^7	1.63	2.89×10^6	13.66	3560.78	3.18
	(2.6, 206.52)	(8.25×10^6, 1.76×10^7)	(1.50, 1.78)	(8.70×10^5, 6.10×10^6)	(10.11, 17.74)		
DLM_2	104.72	1.26×10^7	5.04	2.87×10^6	4.68	3599.80	3.21
	(-13.03, 222.47)	(8.26×10^6, 1.78×10^7)	(2.65, 8.19)	(8.08×10^5, 6.21×10^6)	(2.03, 8.44)		

Table 2: Comparison of the inference on parameters and forecast accuracy using DLM_1 and DLM_2. 95% confidence intervals (in parenthesis) were based on asymptotic Normality. The 95 % confidence interval for DLM_2 implied no linear association between $X_{1,t}$ and $X_{2,t-1}$ while the one for DLM_1 implied a statistically significant association.

The inference on β supports the preliminary argument made in the previous section, suggesting that although a statistically significant linear association exists between $X_{1,t}$ and $X_{2,t-1}$, this association appears to be weak. The effect of the choice of DLM was less important in terms of one step ahead forecast accuracy with only a slight decrease in MAE and MAPE when DLM_1 was used. In the next section we compare the performance of our DLM_1 with the models outlined in section 3.2.

4.2 Forecast accuracy

The most common methods to determine forecasting accuracy are functions of forecasting error. MAE, MAPE, and root mean square prediction error (RMSE) were calculated in sample, and out of sample for a time horizon of 6 months, and 7-12 months ahead. Given the small amount of out of sample data, the interpretation of these errors should be taken lightly and statistical inference on significance of difference in forecasting errors is unreliable and not presented here. Summaries of the errors are presented in Table 3. A dynamic linear model without using SQV, DLM_0, was also fit for this comparison. In general, it appears that SQV improves forecasts for a horizon of over 6 months, but DLM_0 performs better for the shorter horizon. Based on the in sample results, SARIMA had the worst fit to the data. Out of sample errors indicate that the semiparametric AM performed worst in terms of forecasting and that SARIMA had a competitive forecasting performance with other alternatives. Generally, these metrics suggested HW and DLM_1 were the best alternatives for short term forecasts. For horizons over 6 months, DLM_1 performed best followed by SARIMA and HW. It is unclear if the prediction error differences are statistically significant. As we can see from Figure 3, overall the forecast of all the models captured the general pattern in $X_{1,t}$. All forecasts underestimated the March 2013 NHNR (which turned out to be higher than in any other March) and overestimated the September 2013 NHNR.

Forecast accuracy									
	MAE			$MAPE$			$RMSE$		
Model	In	Out-6	Out-12	In	Out-6	Out-12	In	Out-6	Out-12
DLM_1	3560.78	7024.38	4160.76	3.18	5.01	3.76	4570.37	8111.41	4760.00
DLM_0	3633.70	6485.74	4490.72	3.26	4.56	4.13	4653.87	7751.16	5282.41
$SARIMA$	4709.43	7491.45	4972.53	4.14	5.50	4.50	6021.21	9451.35	5829.13
HW	4220.69	5901.93	4756.51	3.80	4.14	4.59	5422.68	7330.08	6364.35
AM	4326.67	7867.42	7618.90	3.98	5.84	7.56	5372.06	9217.15	11425.63

Table 3: Forecast accuracy comparison of models $DLM_1, DLM_0, SARIMA, HW$, and AM. 'In' column shows in sample errors, 'Out-6' errors up to 6 months ahead and 'Out-12' forecast errors for horizons of 7-12 months ahead.

Figure 4 presents the last few search query data observations (based on Google Trends algorithm run 11) with DLM_1 forecasts up to 12 months ahead. The data not used to construct the model is also included. We see that the model tended to overestimate the monthly search query volume for the first few months. Over the first 6 forecast months the MAPE when forecasting SQV was found to be 9.67 and for the forecasts 7-12 months ahead the MAPE was 3.74. The prediction errors over the first 6 forecast months are markedly higher than those for NHNR. Since the DLM_1 forecasts of $X_{1,t}$ depend on the forecasts of $X_{2,t-1}$ a poor performance in forecasting the latter process will hinder its accuracy in forecasting the former (Ashley 1983). These arguments explain the performance comparison of DLM_0 and DLM_1. Models with autoregressive features and with a growth component were also considered for $X_{2,t-1}$ but they did not improve on the results seen here.

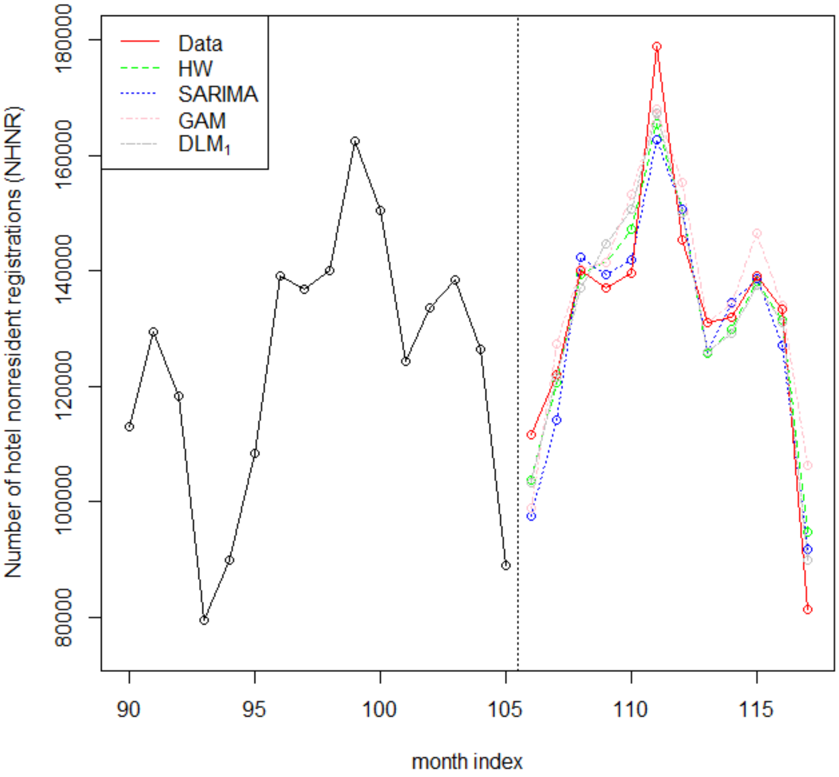

Figure 3: Last few NHNR observations with forecasts up to 12 months ahead from all the models. Data not used to construct the model is also included.

5. Conclusions

The aim of this work was to construct an adequate model to forecast the number of hotel nonresident registrations in Puerto Rico. The possibility of using search query volume data was considered. As far as we know, this is the first paper to account for the uncertainty of the Google SQV data. We showed that our proposed DLM allows to conduct more precise inference on the lagged linear association of the two temporal processes than downloading Google Trends output only once. The evidence showed a statistically significant linear association between $X_{1,t}$ and $X_{2,t-1}$. However, this association is weak to moderate and DLM forecasts results were mixed when compared to the simpler Holt-Winter and SARIMA models. Two explanations are given for the forecasting performance of our DLM. First, the rather weak linear association between $X_{1,t}$ and $X_{2,t-1}$, indicated by the traits of the corresponding time series and the resulting inference as explained above. du Preez and Witt (2003) obtain similar findings where univariate models outperformed multivariate ones due to the absence of strong cross correlation between the processes. Secondly, the performance of the DLM in forecasting $X_{2,t}$ was not good enough to compensate for the weak linear association between the processes. The findings of this research do not mean that overall Google Trends data is not useful, but that it might not be useful to forecast NHNR in Puerto Rico. However, we acknowledge

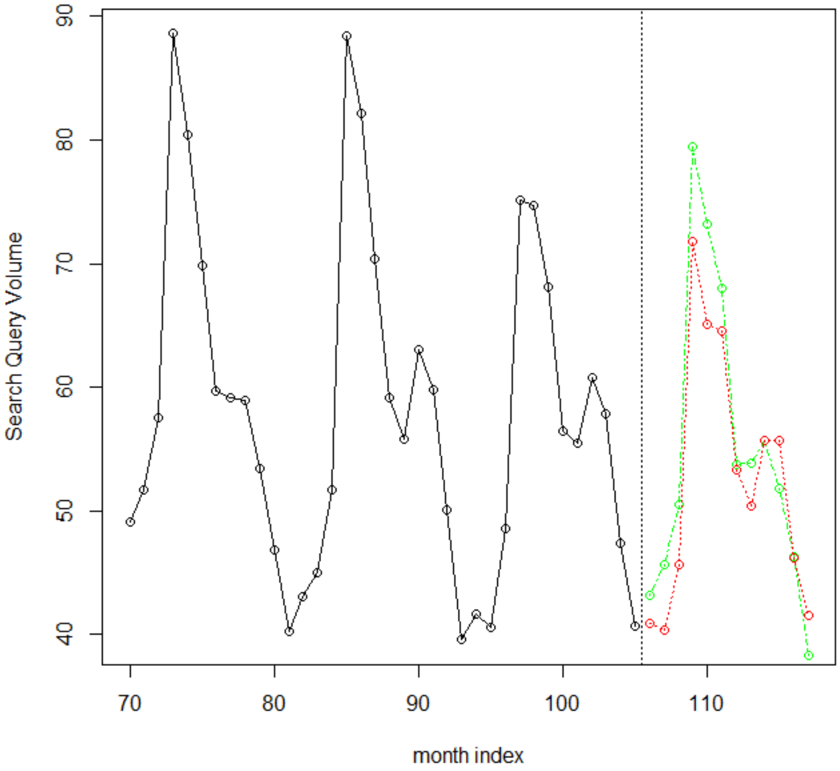

Figure 4: Last few search query data observations (based on Google Trends algorithm run 11) with DLM_1 forecasts up to 12 months ahead (dot dash line). Data not used to construct the model is also included (dotted line).

that our selection of SQV data was mostly heuristic. Further research is needed using more objective alternatives to choose which and how many search queries should be included within the limits that Google Trends allows. Moreover, the association between NHNR and SQV may be stronger at a weekly level, since some visitors may schedule their stay a few weeks before making their trip instead of a month before hand. More research is needed to see if a dynamic model incorporating a latent process, and mixed frequency time series data would help improve forecasts. Preliminary analysis indicates that the SQV data at a weekly level was noisier than its aggregated monthly counterpart, leading to higher prediction errors when forecasting SQV. At the very least, a stochastic seasonal component would be needed. SQV data retrieved from Google Trends may improve forecasts of processes, especially in situations when the main time series of interest and the SQV data display strong growth and when the search query volume data can be forecasted well. Google does not provide much detail on how they obtain their SQV data and why the data available through Google Trends changes routinely. The mechanism producing the data helps determine the right modeling approach (e.g determining if there's a need to adjust for bias). More transparency from Google would improve the chances of exploiting the promising tool of search query volume data.

Care must be taken when analyzing the forecasting accuracy of models. One must realize that the forecast accuracy measures are statistics, and different samples may result in different comparative results of these measures. Although inference based on forecast accuracy measures has been developed, simulations suggests that these hypothesis testing methods require a substantial amount of out of sample data, at least 40 observations in length to be useful (Ashley 2003). Also, although Kalman recursion allows for the estimate of prediction error covariance recursively, this estimate is dependent on assumptions taken about the covariance of the observation and system error. Typically, in practice the covariances parameters are unknown and must be estimated. A fully Bayesian perspective allows to measure the uncertainty involved in the estimation of these covariance parameters.

References

Akaike, H. (1973), "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, 60, 255–265.

Ashley, R. (1983), "On the Usefulness of Macroeconomic Forecasts as Inputs to Forecasting Models," *Journal of Forecasting*, 2, 211–223.

— (2003), "Statistically significant forecasting improvements: how much out-of-sample data is likely necessary?" *International Journal of Forecasting*, 19, 229–239.

Athanasopoulos, G. and Hyndman, R. J. (2008), "Modelling and forecasting Australian domestic tourism," *Tourism Management*, 29, 19 – 31.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall.

Bisgaard, S. and Kulahci, M. (2011), *Time Series Analysis and Forecasting by Example*, Wiley.

Brockwell, P. J. and Davis, R. A. (2009), *Time Series: Theory and Methods*, Springer-Verlag, 2nd ed.

Butler, D. (2013), "When Google got flu wrong," *Nature*, 494, 155–156.

Chatfield, C. (2003), *The Analysis of Time Series An Introduction.*, Chapman and Hall, sixth ed.

Choi, H. and Varian, H. (2009a), "Predicting the present with Google trends," Tech. rep., Google.

— (2009b), "Predicting initial claims for unemployment benefits," Tech. rep., Google.

Cressie, N. and Wikle, C. K. (2011), *Statistics for spatiotemporal data*, Wiley.

du Preez, J. and Witt, S. F. (2003), "Univariate versus multivariate time series forecasting: an application to international tourism demand," *International Journal of Forecasting*, 19, 435–451.

Durbin, J. and Koopman, S. J. (2012), *Time Series Analysis by State Space Methods.*, Oxford University Press, 2nd ed.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data," *Nature*, 457, 1012–1014.

Gluskin, R. T., Johansson, M. A., Santillana, M., and Brownstein, J. S. (2014), "Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends," *PLoS Neglected Tropical Diseases*, 8, e2713.

Goel, S., Hofman, J. M., Lahaihe, S., Pennock, D. M., and Watts, D. J. (2010), "Predicting consumer behavior with Web search," *Proceedings of the National Academy of Sciences*.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–318.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2nd ed.

Holmes, E. (2012), "Derivation of the EM algorithm for constrained and unconstrained marss models," Tech. rep., Northwest Fisheries Science Center, Mathematical Biology Program.

Holmes, E., Ward, E. J., and Wills, K. (2012), "MARSS: Multivariate autoregressive state-space models for analyzing time-series data," *The R Journal*, 4, 30.

Huerta, G., Sansó, B., and Stroud, J. R. (2004), "A spatiotemporal model for Mexico City ozone levels," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53, 231–248.

Junta de Planificación de Puerto Rico (2013a), "La Economía de Puerto Rico en el año fiscal 2012 y perspectivas para los años fiscales 2013 a 2014," Tech. rep., Junta de Planificación de Puerto Rico.

— (2013b), "Perfil de los Visitantes año fiscal 2011," Tech. rep., Junta de Planificación de Puerto Rico.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014), "The parable of Google Flu: traps in big data analysis," *Science*, 343, 1203–1205.

Lim, C. and McAleer, M. (2001), "Forecasting tourist arrivals," *Annals of Tourism Research*, 28, 965–977.

Paciorek, C. (2007), "Computational techniques for spatial logistic regression with large data sets," *Computational Statistics and Data Analysis*, 51, 3631–3653.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ruiz, A. L. (2012), "The Economic Impact of the Tourist Activity and Hotel Industry in the economy of Puerto Rico: An Analysis in Input-Output Framework." *Gran Tour: Revista de Investigaciones Turísticas*, 6, 8–43.

Ruppert, D., Wand, M. P., and Carroll, R. (2003), *Semiparametric Regression*, Cambridge University Press.

Shumway, R. H. and Stoffer, D. S. (2011), *Time Series Analysis and Its Applications with R Examples.*, Springer, 3rd ed.

Song, H. and Li, G. (2008), "Tourism demand modelling and forecasting?A review of recent research," *Tourism Management*, 29, 203 – 220.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag.

Wood, S. (2006), *Generalized additive models - An introduction with R*, Chapman and Hall.