Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Forecasting tourism demand through search queries and machine learning[1]

Rendell E. de Kort,
Central Bank of Aruba

[1] This paper was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Forecasting tourism demand through search queries and machine learning[1]

Rendell E. de Kort[2]

Abstract

This paper utilizes different machine learning techniques for tourism demand forecasting. Considering the magnitude of tourism in terms of economic contribution to Small Island Developing States (SIDS), policy making could benefit greatly from accurate tourism demand forecasting. This paper pursues a novel approach of identifying relevant search query features through google correlate and applying machine learning techniques to estimate individual source market series prior to aggregation. The prediction performance of several machine learning methods is assessed when applied to monthly tourist arrivals from individual source countries to Aruba from 1994 to 2016. The results indicate that machine learning techniques in combination with novel internet datasets sets pose great potential for achieving accurate tourism demand forecasts.

Keywords: Forecast combination, machine learning, feature selection, tourism demand forecasting, random forest, search data.

JEL classification: C22, C40, C52, C63

[1] The views expressed here are those of the author and do not necessarily reflect those of the Centrale Bank van Aruba.

[2] (R.e.dekort@cbaruba.org) Economist, Research Department, Centrale Bank van Aruba.

1. Introduction

A key challenge in many tourism destinations is the accurate forecasting of inbound tourism to support destination management decisions and to guide macroeconomic policy. The importance of the tourism industry is particularly evident in the case of Aruba, as it ranked second among tourism destinations in terms of relative contribution of travel and tourism to GDP in 2016 and where jobs in the tourism industry accounted for an estimated 89.3 percent of total employment (World Travel and Tourism Council, 2017).

However, by its very nature, tourism forecasting remains a very tricky endeavour. The sector is so unpredictable that even a small disturbance in the environment of the host country may bring down the level of demand significantly. Be it predictions about changes in the economic scenario leading to sudden inflation or deflation, any expected occurrences of hostile activities like war or terrorism, any warned natural disasters like earthquakes or floods, any likely incidences of cultural hostility or any kind of threat to public health owing to environmental imbalance or spread of some contagious diseases; all such factors have massive impact on the demand in tourism, making it almost impossible to forecast demand (O'Mahony et al., 2008).

Yet, given the significant impact of tourism on the wider Aruban economy, accurate forecasting of tourism demand is a fundamental input for key decisions on investments, as well as for gauging the conjectural situation and planning for demand flow.

Unfortunately, despite the consensus on the need to develop more accurate forecasts and the recognition of their corresponding benefits, there is no one model that stands out in terms of forecasting accuracy (Claveria et al, 2013). With recent advancements in Internet search technology, a new field has emerged (Google Econometrics), which utilizes time series data on Internet activity by obtaining correlations between keyword searches and macro-economic variables, including unemployment, tourism and consumer demand. Spurred by recent computational advancements, including refinements to the capacity to both efficiently process large volumes of data and run computationally intensive algorithms, there has been an increasing interest in machine learning techniques, including Artificial Neural Networks (ANN) and Random Forests (RF).

In the case of Aruba, tourism demand analysis faces several challenges in terms of, e.g., data availability, erratic factors and a dynamic economy that is inherently vulnerable. Destinations may be inherently vulnerable because they are open to both internal and external human and natural factors, and may have different capabilities to cope with the changes and disturbances originating from these factors (Ridderstaat, 2015). In an effort to counter some of these challenges, this paper conducts a forecasting exercise for tourism demand to Aruba by leveraging the availability of internet search data in combination with recent advances machine learning techniques.

The remainder of the paper is structured as follows. In section 2 the relevant literature is discussed while in section 3 I describe the main methodological frameworks utilized. The results are presented in section 4 and to finalize section 5 presents some concluding remarks.

2. Literature review

Search queries reflect how people show interest and attention on specific topics on the internet and has caught the attention of researchers as a potential useful source of information to model real world phenomenon (Mohebbi et al, 2011). Google provides two data sources that are useful in this context, namely Google correlate and Google Trends. While economic data is often reported with a lag of months or quarters, Google query data is available in real time. This means that queries are contemporaneously correlated with an economic time series, which may be helpful for economic 'nowcasting' (Stephens-Davidowitz and Varian, 2015). Furthermore, existing studies have demonstrated that these data can predict future trends (Choi and Varian, 2012). In the field of tourism, this development has not gone unnoticed, as the predictive power of internet searches has been explored to predict the number of visitors (Saidi et al, 2010; Li, 2016; Yang et al, 2014).

Given a temporal pattern of interest, Google Correlate provides an online, automated method for query selection which determines which queries best mimic the data (Mohebi et al., 2011). More specifically, when time series are uploaded, Google Correlate computes the Pearson Correlation Coefficient (r) between the time series of interest and the frequency time series for every query in the google database. Correlation coefficients range from r=-1.0 to r=+1.0. The queries that Google Correlate shows are the ones with the highest correlation coefficient (i.e. nearest to r=1.0) (Mohebbi, M. et al, 2011). Tourism demand modelling and forecasting studies have focused predominantly on tourist arrivals as proxy for tourism demand (Song and Li, 2008). However, the literature has presented at least three classes of tourism models, namely, those explaining the tourist expenditure, tourist arrivals and length of stay. The most accepted measure of tourism demand is tourism expenditure (Ahmed, 2013). For this study, tourism receipts are utilized since it is available and provides a closer proxy to what tourists contribute to the economy in monetary terms. The literature suggest that tourism demand very often exhibit patterns in term of seasonal, cyclic and trend components (Cankurt and Subasi, 2015). This is a challenge to traditional forecasting techniques to which machine learning could potentially aid. Also, real-time macroeconomic data are typically incomplete for today and the immediate past ('ragged edge') and subject to revision. To enable more timely forecasts, the 'ragged edge' issue can be framed as a standard "nowcasting" problem and addressed in similar fashion to the nowcasting framework of the Centrale Bank van Aruba, as outlined in Zult and Schreuder (2011).

In terms of techniques, compared to econometric models, machine learning based approaches count on several significant advantages, particularly when modelling large data sets. Machine-learning techniques are gaining ground among econometricians, and are particularly well suited to the nowcasting problem. Traditionally, econometrics and machine learning have focused on different types of problems, and have developed separately. Econometrics has generally focused on explanation, with particular attention to issues of causality, and a premium placed on models that are easy to interpret. A "good" model in this framework is mostly assessed on the basis of statistical significance and in-sample goodness-of-fit. Machine learning, on the other hand, has focused more on prediction, with emphasis instead on a model's accuracy rather than its interpretability. A "good" machine-learning model, then, is often determined by looking at its likely out-of-sample success, based on bootstrap-style simulation techniques (Tiffen, 2016).

Another interesting insight that has emerged from the machine learning literature is that averaging over many small models tends to give better out-of-sample prediction than choosing a single model (Varian, 2013). Furthermore, there has been an increasing interest in Artificial Neural Networks (ANN) due to controversial issues related to how to model the seasonal and trend components in time series and the limitations of linear methods (Claveria et al, 2013). In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression (Li, 2016). Machine learning models are also deemed superior in recognizing and learning the seasonal patterns without removing them from the raw data (Cankurt and Subasi, 2015).

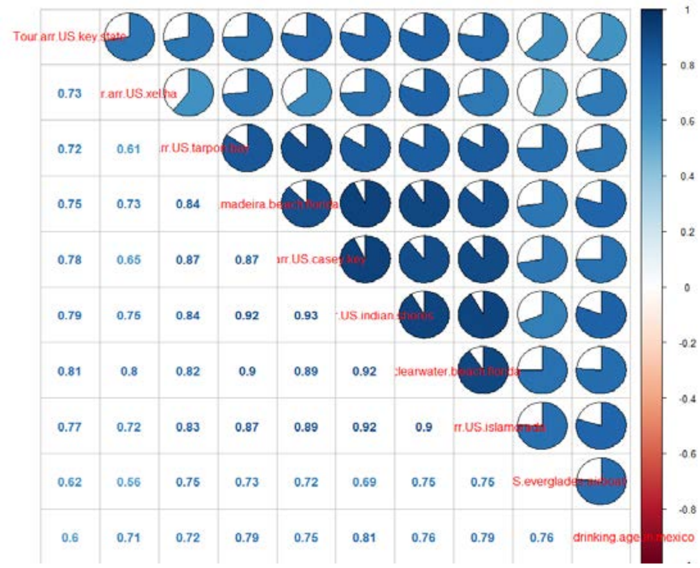
3. Methodology

As a proxy for the dependent variable representing tourism demand, quarterly tourism receipts were collected from the Centrale Bank van Aruba from 2004 to 2016. Tourism arrivals and nights for the 5 largest source markets are collected on a monthly basis and passed through google correlate to identify search terms that have a similar pattern of activity as our dependent variables ("features"). Google correlate surfaces search queries whose temporal patterns are most highly correlated (R2) with our target pattern. Google correlate employs a novel approximate nearest neighbour (ANN) algorithm over millions of candidate queries in an online search trees to produce results. The top 5 source countries combined are found to account for about 90 percent of arrivals/nights.

In total 100 features are collected (see Table 1). The fact that most of the features identified by google correlate are related to tourism provides initial face validity of their inclusion.

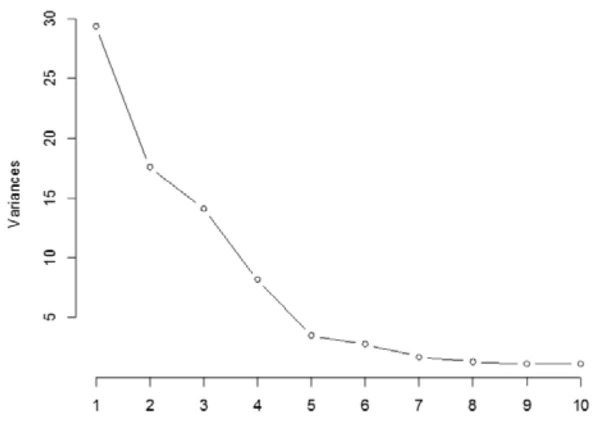
We utilize Google trends to download the features. In practice two ways to achieve this were considered: by (tediously) downloading CSV files from the Google Trends website or by scripting a connection to Google trends data through packages like "gtrendsR" in the R statistical software. Once the data was obtained, unsurprisingly, many of the collected variables illustrated strong co-movement. Figure 1 provides a correlogram of the first 10 features.

Figure 1: Correlogram (First ten features)



As Figure 1 illustrates, the 100 collected Google search term series exhibit a high degree of co-movement. The bulk of their dynamics can therefore be captured by relatively few common factors, effectively reducing the dimensions of the full dataset to a more manageable set (5) of key drivers (see Figure 2). The assumption being that the 5 principle components represent a concise and sufficient summary of underlying processes that drive tourism demand. As is evident by Figure 2, the marginal improvements in captured variance diminishes greatly after the 5th principal component. In terms of approach, the reduction through PCA closely resembles the methodology adopted by Zult and Schreuder (2011).

Figure 2: Correlogram Principle Components



The 'ragged edge' issue of incomplete real-time macroeconomic data is particularly apparent in Aruba, were the dependent variable of interest (tourism receipts) can have a lag of up to 6 months in comparison to real-time Google data. Therefore, to fully take advantage of the monthly frequency and timely availability of the collected features, the dependent variable (tourism receipts) is disaggregated using the "Chow Lin" with 'sum' disaggregation method which converts the series from a quarterly to monthly frequency (see: Sax and Steiner, 2013) using the following equation:

$$REC_t = \propto + \beta_1 Arrivals_t + \beta_2 Nights_t + \beta_3 Time_t + \beta_4 D1 + \beta_5 D3 + e_t,$$

Where the dependent variable tourism receipt is a function of total tourist arrivals, total nights, a time variable, and 2 seasonal dummies.

Temporal disaggregation — Table 1

Variables	Coefficient	Std. Error	T value	Prob
(Intercept)	1.27.0e+02	0.2423	5.243	<0.001
Arrivals	-2.543e-03	7.817e-04	-3.253	0.002
Nights	3.928e-04	9.608e-05	4.088	<0.001
Time	5.503e+01	6.386e-02	8.658	<0.001
D1	3.299e+01	3.805e+00	8.670	<0.001
D3	-1.325e+01	3.085e+00	-4.294	<0.001

Chow-Lin Min RSS Ecotrim disaggregation with 'sum' conversion.

In general, learning algorithms benefit from standardization of the data set. The intention is to counteract the effects of different features having different scales (which then causes models to assign incorrect weights). The data is therefore normalized between 0 and 1 by:

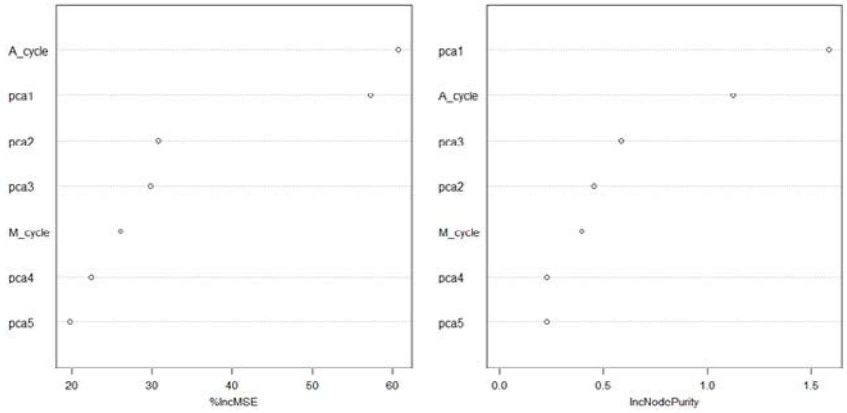
$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

To implement machine learning algorithms, the prediction is framed as a supervised learning problem where we have to infer from historical data the possibly nonlinear dependence between the input and the output (future value). To run the machine learning algorithms, the dataset is split between a training set (January 2004 – December 2014) and a test set (January 2015 – December 2016). The forecast period is defined to cover 12 month beyond the test set (January 2017 – December 2017). More specifically, 3 machine learning techniques are implemented, namely: random forest including Google data, neural network auto regression, and a neural network including Google data.

Random Forest (RF)

At core, these methods are based on the notion of a decision tree, which aims to deliver a structured set of yes/no questions that can quickly sort through a wide set of features, and produce an accurate prediction of a particular outcome. Decision trees are computationally efficient, and work well for problems where there are important nonlinearities. The RF algorithm seeks to improve the model's predictive ability by growing numerous (unpruned) trees and combining the result. This method produces surprisingly good out-of-sample results, particularly with highly nonlinear data. In fact, Random Forests have been accredited as the most successful general-purpose algorithm in modern times (Varian, 2013). A more detailed methodological discussion on how RF works in the context of time series forecasting is provided by Tiffen (2016). In constructing the RF, the 5 Google based principle components are utilized along with two additional time variables to account for annual and monthly cyclical behaviour (Figure 3).

Figure 3: Random Forest



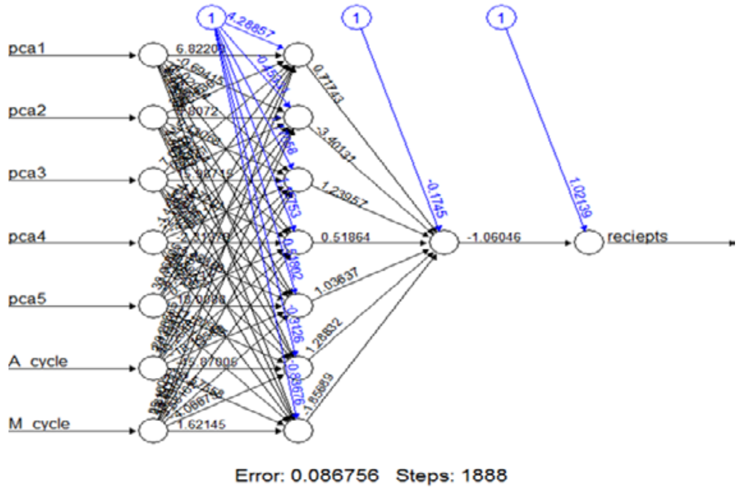
Neural Network Autoregression (NNA)

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors. A neural network consists of an input layer, an output layer, and usually one or more hidden layers. Each of these layers contains nodes, and these nodes are connected to nodes at adjacent layer(s). In the neural network autoregression, lagged values of the time series are used as inputs to a neural network (similar to a linear autoregressive model). We consider a feed-forward network with one hidden layer. The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a “learning algorithm” that minimises a “cost function” such as MSE (Hyndman and Athanaspoloulos, 2013).

Neural Network (using Google data)

Consistent with the input in the previously mentioned RF calculation, the 5 Google based principle components are supplemented with two time variables to account for annual and monthly cyclical behaviour. We consider a feed-forward network with one hidden layer. Figure 4 provides a visual representation of the neural network and the inter-relationship between the different layers.

Figure 4: Neural Network (using Google data)



4. Results

In this section we evaluate the forecasting accuracy of the three machine learning techniques (Random Forest, Neural Network Auto Regression and Neural Network including Google data) by examining out-of-sample predictions of tourism receipts in Aruba. The collected data was divided in training, validation and test sets to assess the performance of the algorithms on unseen data. The forecasting performances are compared in terms of their relative performance for the test set (January 2015 – December 2016). The results of our forecasting competition are shown in Table 2.

Forecast model accuracy

Table 2

	ME	RMSE	MAE	MPE	MAPE	ACF1	Thiel's U
Random Forest	0.093	0.137	0.097	11.571	12.404	0.562	1.018
Neural Network AR	0.037	0.051	0.042	5.023	5.972	0.352	0.405
Neural Network (Google)	-0.037	0.072	0.063	-7.334	10.164	0.493	0.675

When comparing forecasting performance, the various measures are consistent in contending that the prediction error is substantially less for the Neural Network AR model, followed by the Neural Network using the Google variables.

Annex 2 provides a visual example of a Neural network model fitted based on the training dataset, tested for accuracy using the test set and forecasted for 12 months ahead.

5. Conclusion

In terms of interpretability, it should be noted that both neural networks and random forest resemble black boxes: explaining their outcome is much more difficult than explaining the outcome of simpler models (such as a linear models) due to their complexity. Nevertheless, these models have the advantage of providing fairly accurate estimates and despite their computational complexity, improvements in computing technology enable relatively quick execution of machine learning algorithms. This paper provided an example where the combination of near real-time Google search information along with machine learning techniques provides forecasters with a new set of tools to model complex relationships such as tourism demand, but which could easily be transferred to similar macro-economic variables within other domains.

6. References

Ahmed, Y. (2013). Analytical review of tourism demand studies from 1960 to 2014. International Journal of Science and Research (IJSR).

Breiman, L. (2001). Statistical Modeling: The two cultures. Statistical Science 2001, Vol. 16, No. 3, 199-231

Cankurt, S. and Subasi, A. (2015). Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components. Balkan Journal of Electrical & Computer Engineering, 2015, Vol.3, No.1.

Claveria, O. et al (2013). Tourism demand forecasting with different neural network models. Research Institute of Applied Economics. Working paper 2013/21.

Croes, R. and Vanegas, M. (2005). An econometric study of tourist arrivals in Aruba and its implications. Tourism Management. December 2005.

Hassink, W., de Kort, R. and Ridderstaat, J. (2015) "De economische consequenties van de verdwijning van Natalee Holloway", Me Judice, 30 mei 2015.

Hyndman, R.J. and Athanasopoulos, G. (2013) Forecasting: principles and practice. OTexts: Melbourne, Australia. http://otexts.org/fpp/. Accessed on 9/14/2017.

Law, R. and Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. Tourism Management 20 (1999) 89-97.

Mohebbi, M. et al (2011). Google Correlate Whitepaper. Draft date: June 9, 2011.

O'Mahony, B., Lee, C., Bergin-Seers, S., Galloway, G. & McMurray, A. (2008). Seasonality in the Tourism Industry: Impacts and Strategies.

Ridderstaat, J.R. (2015). Studies on Determinants of Tourism Demand: Dynamics in a Small Island Destination. The Case of Aruba

Sax, C. and Steiner, P. (2013). Temporal Disaggregation of Time Series. The R journal Vol. 5/2, December 2013.

Shmueli, G. and Lichtendahl, K. (2016). Practical Time Sries Forecasting with R: A hands-On Guide [2nd Edition].

Song, H. and Li, G. (2008). Tourism Demand Modelling and Forecasting. A review or Recent Research.

Stephens-Davidowitz, S. and Varian, H. (2015). A hands-on guide to Google data. Draft date: March 7, 2015.

Tiffen, A. (2016). Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon. IMF Working Paper. WP/16/56

World Travel and Tourism Council (2017). Economic Impact 2017 Aruba.

Varian, H. (2013). Big Data: New Tricks for Econometrics.

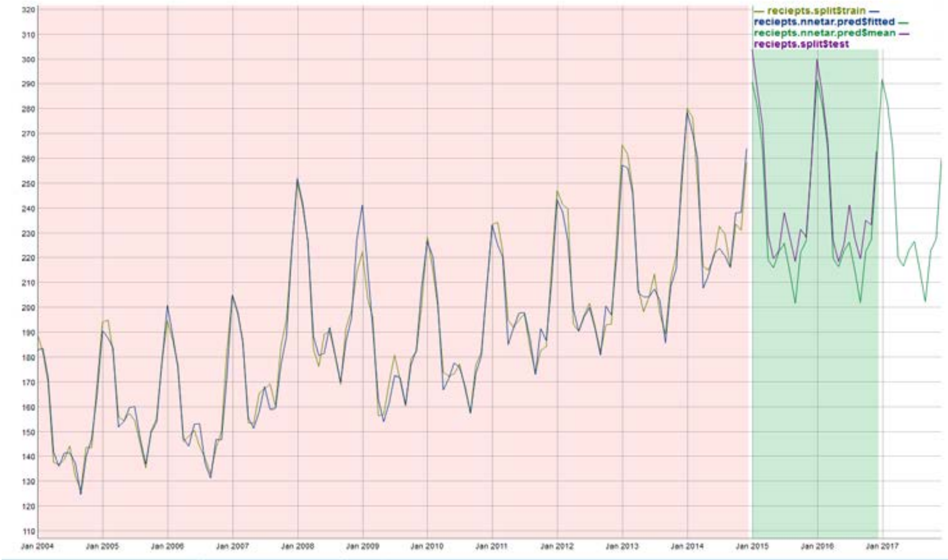
Zult, D. and Schreuder, G. (2011). Monthly Nowcast of Aruban Year-on-Year Growth in GDP. Statistics Netherlands. February 2011.

Appendix 1: google variable selection

		Google correlate predictor	Correlation			Google correlate predictor	Correlation
1	Tourism arrivals United States	madeira beach florida	0.7325	51	Tourism nights United States	disney florida	0.7703
2	Tourism arrivals United States	everglades airboat	0.725	52	Tourism nights United States	arenal costa rica	0.7634
3	Tourism arrivals United States	casey key	0.719	53	Tourism nights United States	old san juan	0.7623
4	Tourism arrivals United States	drinking age in mexico	0.7094	54	Tourism nights United States	pine key	0.7518
5	Tourism arrivals United States	clearwater beach florida	0.7062	55	Tourism nights United States	marathon florida	0.7458
6	Tourism arrivals United States	tarpon bay	0.7061	56	Tourism nights United States	everglades airboat	0.7433
7	Tourism arrivals United States	islamorada	0.7039	57	Tourism nights United States	lauderdale by the sea	0.7412
8	Tourism arrivals United States	key state	0.7025	58	Tourism nights United States	jaco costa rica	0.7392
9	Tourism arrivals United States	xel ha	0.7018	59	Tourism nights United States	marco island	0.7389
10	Tourism arrivals United States	indian shores	0.7016	60	Tourism nights United States	ferry to key west	0.7377
11	Tourism arrivals Venezuela	blusas	0.8946	61	Tourism nights Venezuela	bow target	0.8814
12	Tourism arrivals Venezuela	el emergente	0.8896	62	Tourism nights Venezuela	kid shoes	0.8631
13	Tourism arrivals Venezuela	oficinas zoom	0.8892	63	Tourism nights Venezuela	youth football gloves	0.8601
14	Tourism arrivals Venezuela	outfit	0.8891	64	Tourism nights Venezuela	snake boots	0.8591
15	Tourism arrivals Venezuela	pantalon	0.8888	65	Tourism nights Venezuela	command hooks	0.8532
16	Tourism arrivals Venezuela	zapatos reebok	0.8877	66	Tourism nights Venezuela	crossbow target	0.8524
17	Tourism arrivals Venezuela	blusas de	0.886	67	Tourism nights Venezuela	pencil holder	0.8522
18	Tourism arrivals Venezuela	chores	0.8854	68	Tourism nights Venezuela	kid shoe	0.8519
19	Tourism arrivals Venezuela	zapatos timberland	0.8852	69	Tourism nights Venezuela	boys shoes	0.8508
20	Tourism arrivals Venezuela	cabellos	0.8838	70	Tourism nights Venezuela	under armour youth	0.8506
21	Tourism arrivals Colombia	coomotor	0.8334	71	Tourism nights Colombia	ensaladas	0.8025
22	Tourism arrivals Colombia	terminal	0.8132	72	Tourism nights Colombia	boyacense	0.7994
23	Tourism arrivals Colombia	a prima	0.8111	73	Tourism nights Colombia	cinco pa las doce	0.7993
24	Tourism arrivals Colombia	flota	0.8084	74	Tourism nights Colombia	grinch	0.7986
25	Tourism arrivals Colombia	brasilia	0.8025	75	Tourism nights Colombia	feliz aÃ±o	0.7974
26	Tourism arrivals Colombia	copetran	0.7979	76	Tourism nights Colombia	tamales	0.7962
27	Tourism arrivals Colombia	comotor	0.7948	77	Tourism nights Colombia	mensajes de fin de aÃ±o	0.7951
28	Tourism arrivals Colombia	prima a	0.7924	78	Tourism nights Colombia	inocentadas	0.7949
29	Tourism arrivals Colombia	ruta bogota	0.7915	79	Tourism nights Colombia	aÃ±o viejo	0.7947
30	Tourism arrivals Colombia	la prima	0.7815	80	Tourism nights Colombia	feliz navidad	0.7934
31	Tourism Arrivals Netherands	route 4	0.6797	81	Tourism nights Netherlands	friese ballonfeesten	0.804
32	Tourism Arrivals Netherands	etape du tour	0.6767	82	Tourism nights Netherlands	paardenmarkt voorschoten	0.8012
33	Tourism Arrivals Netherands	truckstar	0.6751	83	Tourism nights Netherlands	kermis tilburg	0.7965
34	Tourism Arrivals Netherands	cross	0.6689	84	Tourism nights Netherlands	tilburgse kermis	0.7925
35	Tourism Arrivals Netherands	wedren nijmegen	0.6659	85	Tourism nights Netherlands	parade utrecht	0.7894
36	Tourism Arrivals Netherands	ardennen last minute	0.6651	86	Tourism nights Netherlands	tilburg kermis	0.7879
37	Tourism Arrivals Netherands	buenas noches	0.6638	87	Tourism nights Netherlands	brielle blues	0.7868
38	Tourism Arrivals Netherands	laatste minuut	0.6618	88	Tourism nights Netherlands	bierhal	0.7861
39	Tourism Arrivals Netherands	bernard hinault	0.661	89	Tourism nights Netherlands	acht van chaam	0.7856
40	Tourism Arrivals Netherands	de kans	0.6602	90	Tourism nights Netherlands	roze maandag	0.7854
41	Tourism Arrivals Canada	palm springs weather	0.9094	91	Tourism nights Canada	mont video	0.9164
42	Tourism Arrivals Canada	springs weather	0.9036	92	Tourism nights Canada	palm springs weather	0.9137
43	Tourism Arrivals Canada	mont video	0.8933	93	Tourism nights Canada	lift tickets	0.9129
44	Tourism Arrivals Canada	ski resort weather	0.8847	94	Tourism nights Canada	night skiing	0.904
45	Tourism Arrivals Canada	ncaab	0.8797	95	Tourism nights Canada	snow report	0.9035
46	Tourism Arrivals Canada	stomach flu	0.8789	96	Tourism nights Canada	springs weather	0.8983
47	Tourism Arrivals Canada	lauderdale weather	0.8781	97	Tourism nights Canada	ski resort weather	0.8964
48	Tourism Arrivals Canada	snow report	0.8759	98	Tourism nights Canada	grand fond	0.8963
49	Tourism Arrivals Canada	surfaceuse	0.8748	99	Tourism nights Canada	mont grand fond	0.8956
50	Tourism Arrivals Canada	fort lauderdale weather	0.8748	100	Tourism nights Canada	rabais ski	0.8947

Data Source: Google Correlate (http://correlate.googlelabs.com)

Appendix 2: Forecast example



Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Forecasting tourism demand through search queries and machine learning[1]

Rendell E. de Kort,

Central Bank of Aruba

[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

CENTRALE BANK VAN ARUBA

Forecasting Tourism demand through search queries and machine learning

Rendell E. de Kort

IFC – Bank Indonesia Satellite Seminar on "Big Data", Bali, Indonesia, 21 March 2017

1. Background

Statistical modeling: The two cultures

- Prediction
- Information

Y ← **Nature** ← X

The data modeling culture

Y ← **Linear regression / Logistic regression / Cox model** ← X

The algorithmic modeling culture

Y ← **Unknown** ← X

Y ↔ Decision trees, Neural nets ↔ X

Source: Breiman, L. (2001). Statistical Modeling: The two cultures. Statistical Science 2001, Vol. 16, No. 3, 199-231

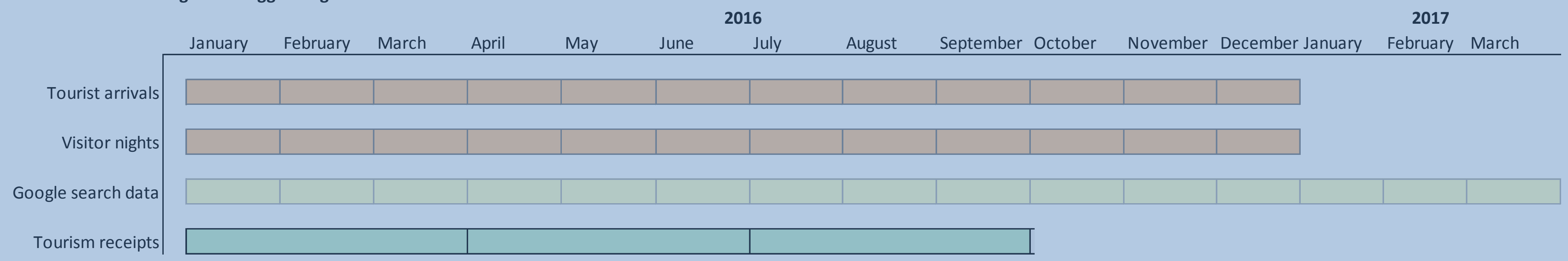
2. Considerations

- Traditional statistical models require eliminating the effects of seasonality prior to forecasting.
- Literature points to the ability of machine learning models to recognize and learn seasonal patterns without removing them from the raw data.
- Traditional statistical forecasting methods are mostly linear models while the literature indicates machine learning techniques cope well with possible nonlinearities.

2. Considerations

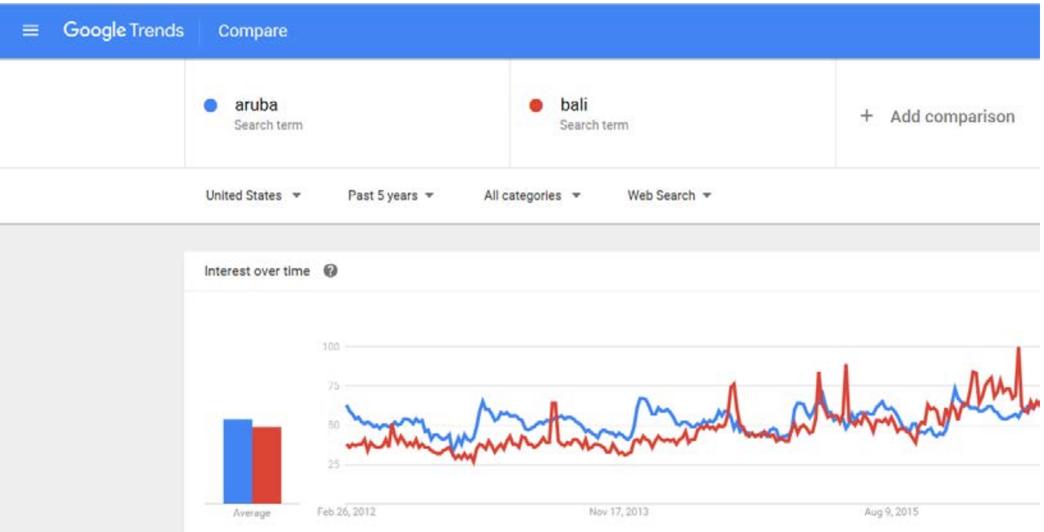
- Real-time macroeconomic data are typically incomplete for today and the immediate past (‘ragged edge’) and subject to revision.
- To enable more timely forecasts the issue is initially framed as a standard “*nowcasting*” problem.

Figure 1: Ragged edge



3. Preprocess

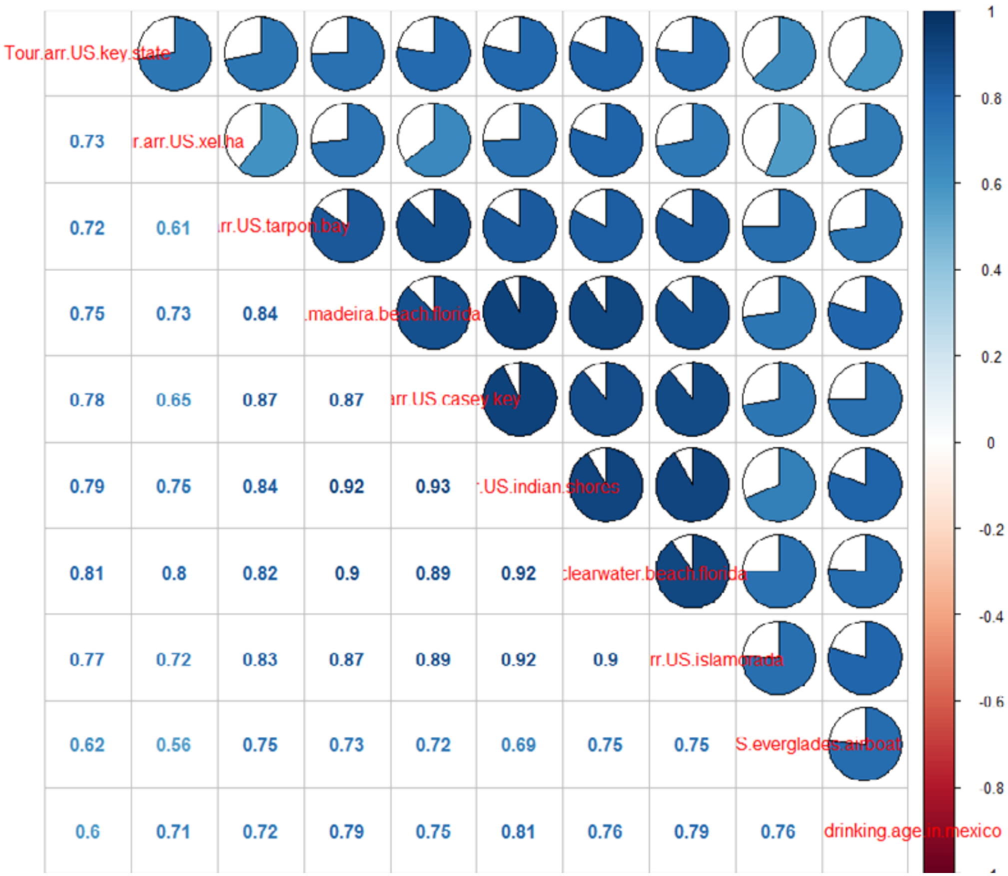
- Google trends enables easy download of Google search query time series.
- Download can take place by:
  - (tediously) downloading CSV files from the Google Trends website.
  - Using packages like “gtrendsR” to connect with your Google account and downloading Trends data directly into R with a simple script.

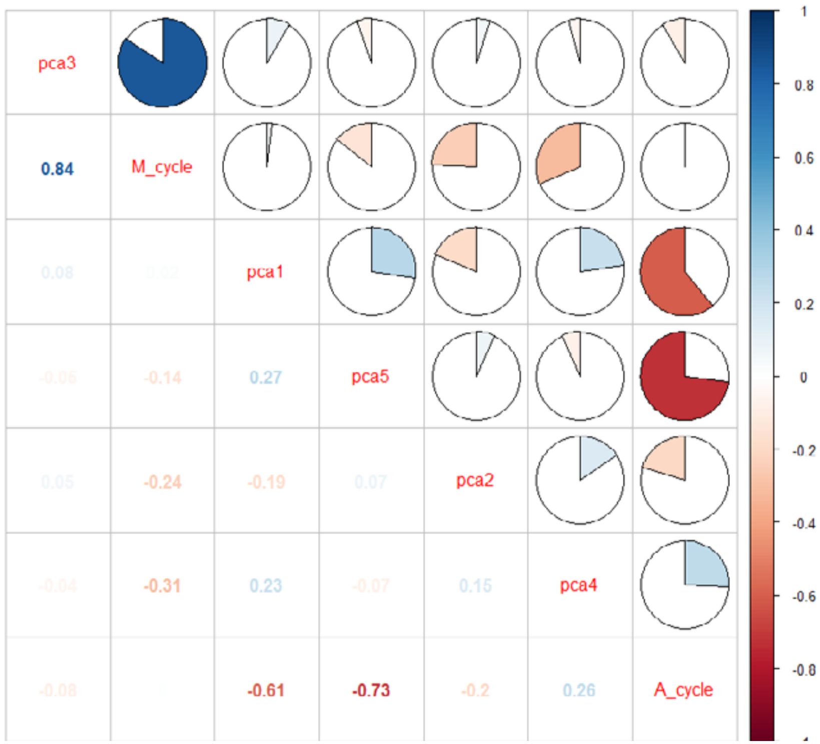


Google correlate predictor	Correlation
madeira beach florida	0.7325
everglades airboat	0.725
casey key	0.719
drinking age in mexico	0.7094
clearwater beach florida	0.7062
tarpon bay	0.7061
islamorada	0.7039
key state	0.7025
xel ha	0.7018
indian shores	0.7016

Data Source: Google Correlate (http://www.google.com/trends/correlate)

3. Preprocess

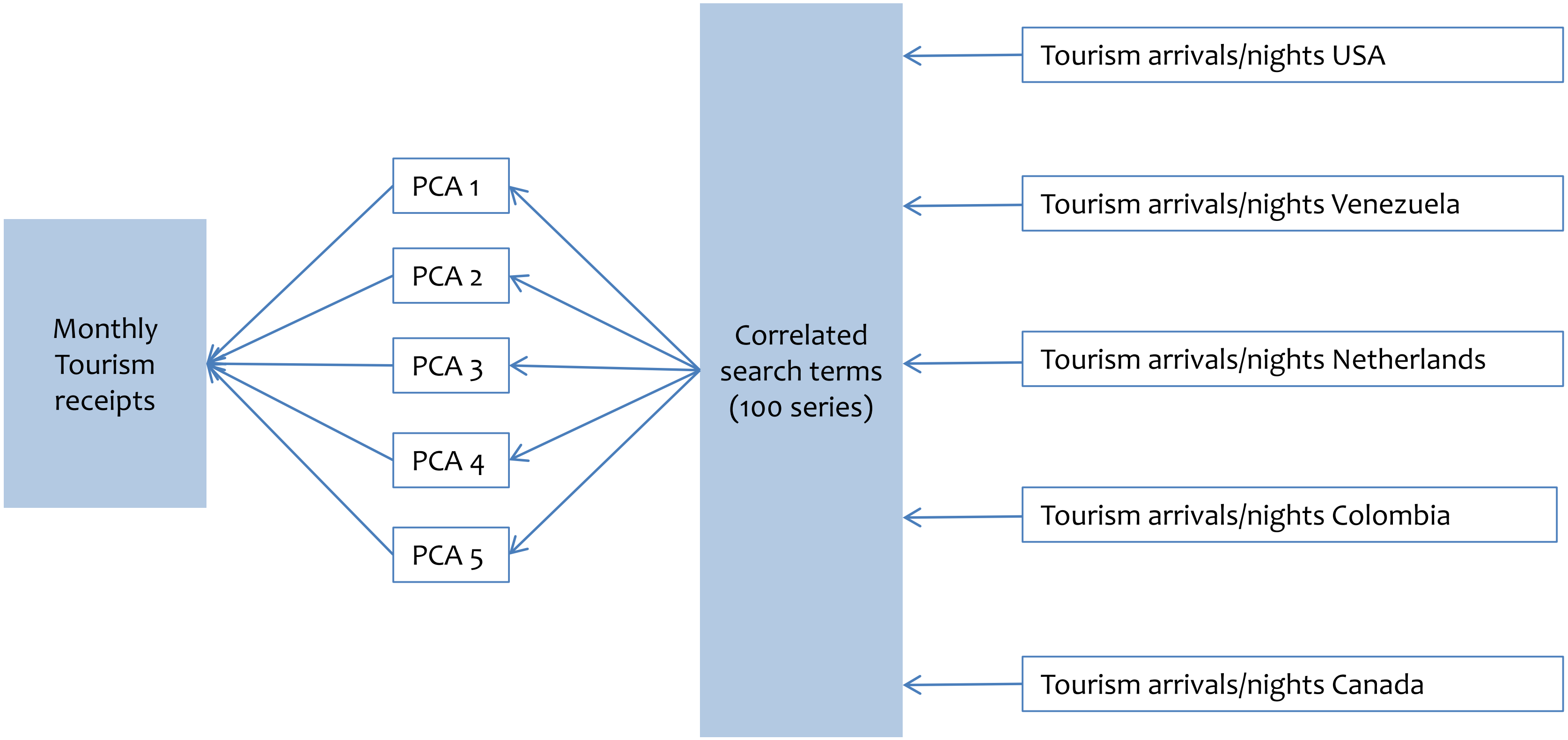




* Important to be mindful of extracting information from a large number of correlated proxies (100 in our example).

3. Preprocess

- The dependent variable "tourism receipts" is measured on a quarterly basis. To take full advantage of features collected on a monthly basis, we're disaggregating the series.



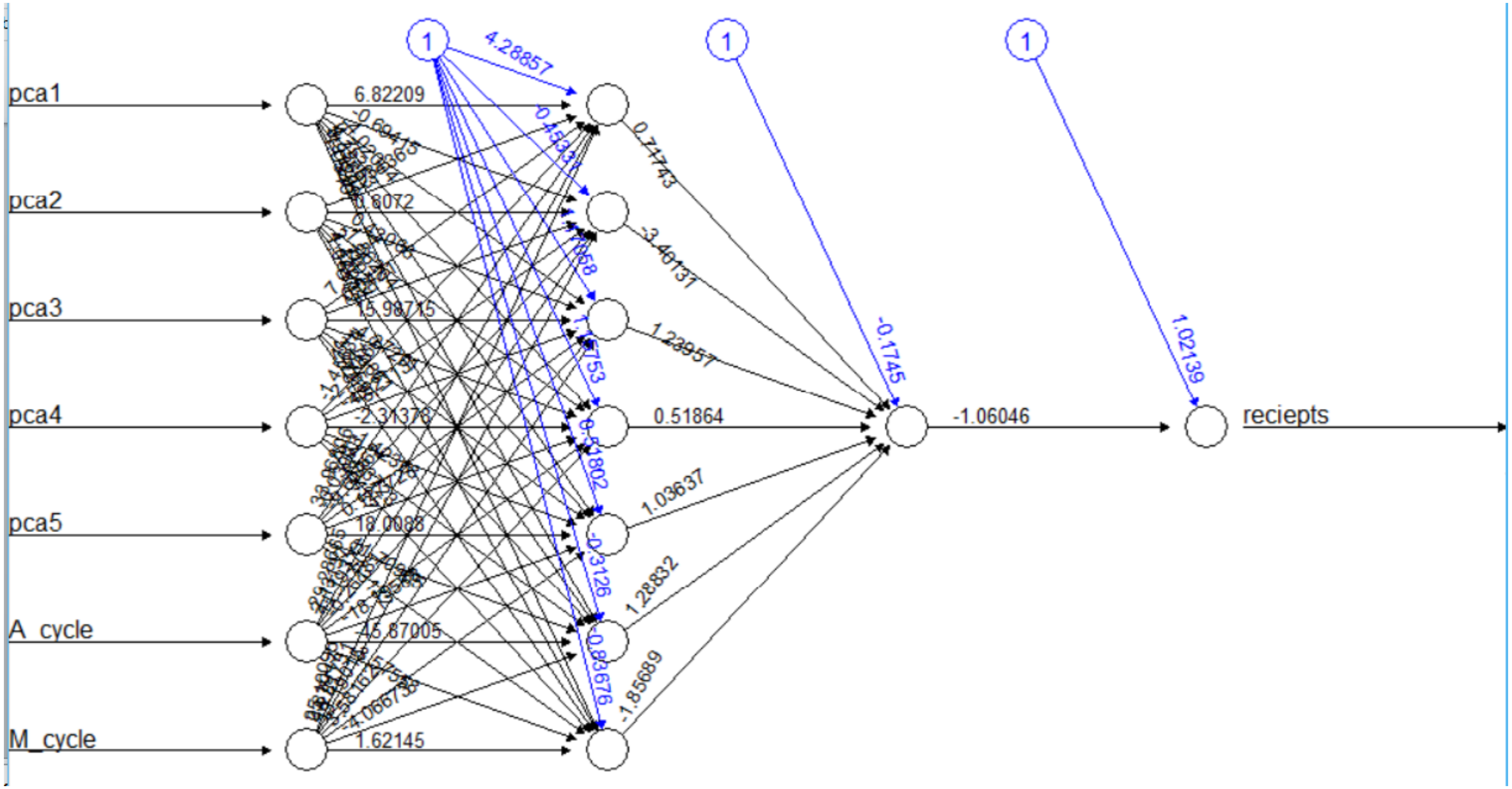
4. Machine learning estimations

Random Forest

- At core, these methods are based on the notion of a decision tree, which aims to deliver a structured set of yes/no questions that can quickly sort through a wide set of features, and produce an accurate prediction of a particular outcome.
- Decision trees are computationally efficient, and work well for problems where there are important nonlinearities.
- The RF algorithm seeks to improve the model's predictive ability by growing numerous (unpruned) trees and combining the result.

4. Machine learning estimations

Neural Network (NN)



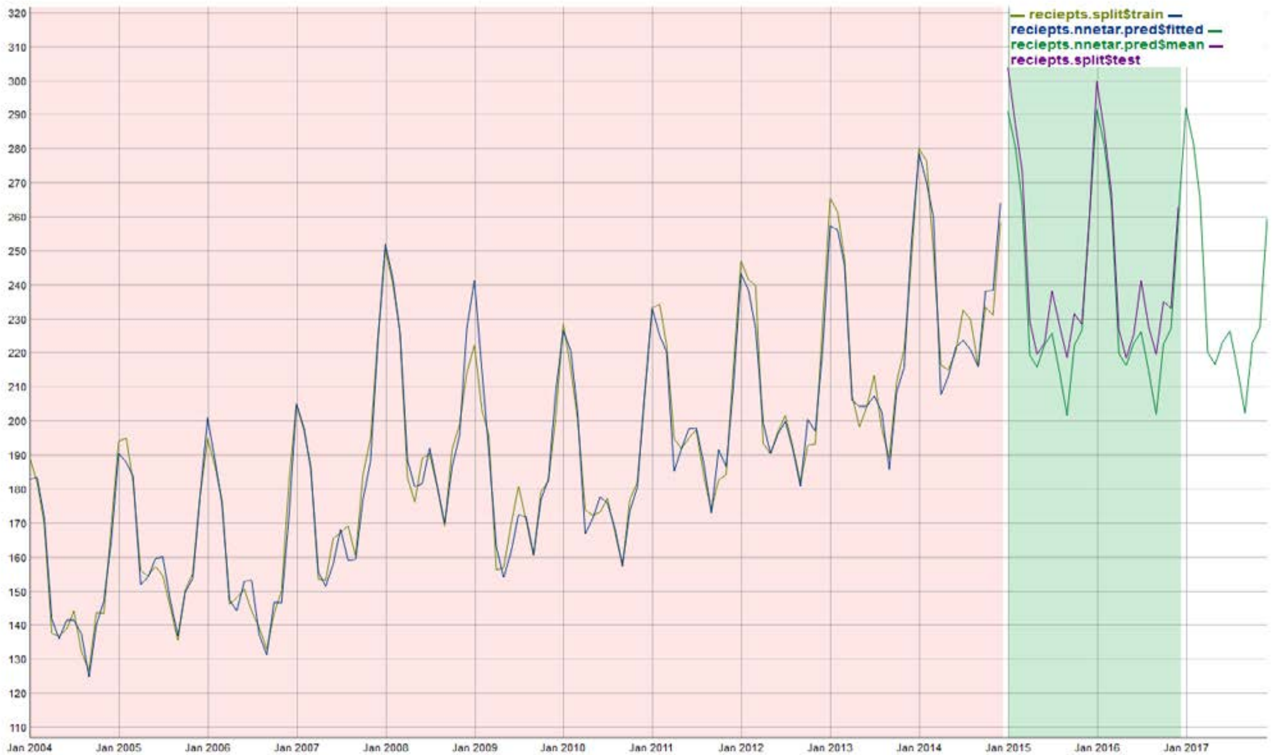
4. Machine learning estimations

Neural network autoregression

- With time series data, lagged values of the time series can be used as inputs to a neural network (similar to a linear autoregressive model).
- we consider a feed-forward network with one hidden layer
- Using the “nnetar” function in R

4. Machine learning estimations

Example



	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Random Forest	0.093	0.137	0.097	11.571	12.404	0.562	1.018
Nueral Network AR	0.037	0.051	0.042	5.023	5.972	0.352	0.405
Nueral Network (Google)	-0.037	0.072	0.063	-7.334	10.164	0.493	0.675

5. Concluding remarks

- Machine learning models provide “good” out-of-sample success.
- Takes advantage of additional search information.
- Tradeoff between interpretability of the model and forecasting performance (predictive not descriptive).
- Benefit: ease of processing once the script is in place.
- Opportunities exist to include the google data in a expanded framework to forecast economic growth.
- The R script available on GitHub: https://github.com/rendell

THANK YOU

Centrale Bank van Aruba

TERIMA KASIH