SYM-OP-IS 2010
XXXVII Simpozijum o operacionim istraživanjima
MINISTARSTVO ODBRANE REPUBLIKE SRBIJE
VOJSKA SRBIJE

CASE STUDY: GOOGLE DATA AS EXPLANATORY VARIABLE FOR SHORT TERM FORECAST OF AIR TRAFFIC

NIKOLA IVANOV
Faculty of Transport and Traffic Engineering, Belgrade , n.ivanov@sf.bg.ac.rs

DAVID MARSH
EUROCONTROL, Brussels, david.marsh@eurocontrol.int

Abstract: *This paper presents selected results from the study undertaken by EUROCONTROL and supported by Faculty of Transport and Traffic Engineering, Air Transport Department. Study objective was to find robust short term leading indicators for changes in air traffic growth to improve EUROCONTROL's Short Term Forecast. One of the methods used in the study to get leading information was Flight Data time series modeling with Google Data as explanatory variable for two month ahead forecast. It was shown that some information on trend brakes in air traffic growth might be obtained with data provided by Google Insights for Search, but information were not sufficient for operational use.*

Keywords: *explanatory variable, leading indicators, Google data, search term, interest*

1. INTRODUCTION

Air traffic has been increasing for the last 40 years with number of flights doubled every 20 years – the tendency that is likely to be continued. Quality information on the traffic trends even in short time horizon, coming year or two, is essential to most congested airports, as well as for the other segments of air transport system which are facing capacity challenges even today.

Eurocontrol is European Organization for the Safety of Air Navigation and counts 38 member states covering almost whole European Sky. One of many Eurocontrol Agency's departments is Statistics and Forecast Service-STATFOR, which objective is to provide statistics and forecasts on air traffic in Europe and to monitor and analyze the evolution of the Air Transport Industry.

STATFOR's Short Term Forecasts (STF) are good at capturing recent trends month by month and projecting these into the immediate future - up to two years ahead using time series modeling. But STATFOR lacks in information that help to indentify changes in traffic trends for shorter time frame, the coming weeks and months.

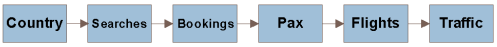
Information on traffic trend changes in the coming months could be obtained with leading indicators. In practice, an indicator is anything that can be used to predict future financial, economic or other trends. Leading indicators are indicators which change before the observed phenomenon changes.

One of available data sources which could provide leading indication of changes in air traffic in the next months is Google Insights for Search™ Service, and is the only one used in the study.

To get such information from Google dataset an effort to use this data as explanatory variable (as one of the methods used) for modelling and forecasting was made.

2. PROJECT IDEA AND ASSUMPTIONS

To become an air passenger one has to book an airline ticket. The airline tickets are usually booked a certain time before the flights. But before that, a search process has to be accomplished to find information for the flights. More searches for flights should lead to more bookings and as a result, more passengers. As a consequence, airlines will increase number of flights producing more traffic (Picture 1).



Picture 1: *Project idea scheme*

The idea lies in a logical assumption that number of searches for flights (representing air transport demand) are correlated with air traffic volume, i.e. more searches produce more flights[1] hence traffic. This correlation is lagged because people search for flights in advance thus allowing us to get leading indication of traffic trend changes. Changes in number of searches for flights could be a leading indicator for changes in air traffic.

Related to air traffic one remark should be noted. Air traffic for one country consists of Arrivals, Departures, Internals and Overflights. Overflights are not likely to be indicated because of the proposed way of indicating: flights over one country don't depend on the number of

[1] With certain assumptions.

searches in that country. Further in the text 'flights' refers to Arrivals, Departures and Internals only.

In short, major assumptions made in this study are listed below:

- For flights search purpose people mostly use internet,
- On the internet people generally use search engines to find information about flights,
- Most frequently used search engine is Google.

Google Insights for Search™ (GIfS from now on) is Google service developed to track a particular search term's popularity across the Web and geographic regions of the world in time. It was used in this study as a tool to measure people's search activity.

To find flights in the proposed manner one may type different search terms in Google search bar, like 'flights' or 'lufthansa'. For given set of quarry filters (country, time, etc.) GIfS provides for download INTEREST for the given term. Interest is computed out of the number of searches for the given term and filter parameters, normalized and scaled after (Picture 2).

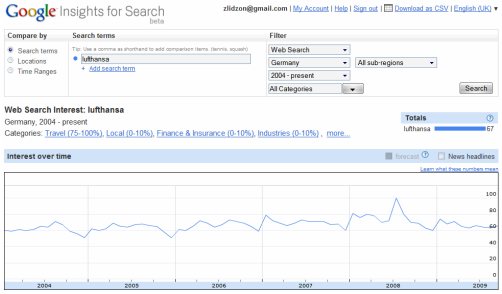
From variety of possible search terms used to find information about flights only four were chosen to represent air transport demand for each country. Changes in interest for the search term, possible leading indicator, should reflect changes in the number of flights for the chosen country and period. The intention, therefore, was put on four different search terms:

- **National carrier**. More than one half of all flights are made by traditional airlines and probably preferred carrier in a country is the flag carrier. For example, „air france" for France or „finnair" for Finland. (AIRLINE series)
- **Low Cost Carrier**. Lately, key drivers for growth in aviation industry were Low Cost airlines. For instance one may use, „air berlin" for Germany or "easyjet" for UK. (MARKET series)
- **Flights**. This is a sort of general term and might be used for finding flights information regardless of market segment[2]. As „flüge" for Austria or „voli" for Italy. (FLIGTHS series)
- **Travel**. Although more general, it could also reflect air transport demand. Such as „voyage" for France or „viajes" for Spain. (TRAVEL series)

3. DATA SOURCES AND METHODOLOGY

Two data sources were used in this study:

- **Google Insights for Search** service provided interests for the four search terms, regarded as possible leading indicators. Interests data were downloaded during the first week of September 2009 to form the interests dataset base for period January 2004 – August 2009.
- **PRISME** (Pan-European Repository of Information Supporting the Management of EATM) database was used to obtain information on flights for each country and to determine airline's market share.



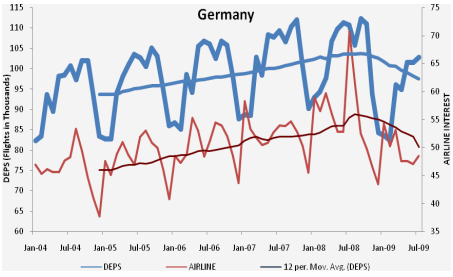
Picture 2: *GIfS output for the search term 'lufthansa'*

Dataset table for each country (Table 1) consists of monthly data „MONTHS", flights for that country „DEPS", and interests „AIRLINE", „FLIGHTS", „TRAVEL" and „MARKET", representing interests in time (from January 2004 to August 2009).

Table 1: *Part of the data set table for Germany; Example*

Observ	*TIME*	*FLIGHS*	*EXPLANATORY VARIABLE DATA*			
	MONTH	DEPS	AIRLINE	FLIGHTS	TRAVEL	MARKET
1	JAN04	82322	46.2	70.2	73.7	46.7
2	FEB04	83391	44.6	63.8	62.7	45.6
3	MAR04	93668	46.2	67.1	59.9	50.7
4	APR04	89543	45.9	68.6	57.7	49.2
5	MAY04	98190	46.0	72.0	61.3	51.3
6	JUN04	98465	48.1	81.7	73.1	55.6
7	JUL04	100720	49.0	95.0	94.3	66.5

Picture 3 shows cross plot (Germany, Jan04-Sep09) for DEPS (bold series) and AIRLINE (thinner series) along with trend obtained with moving average of order 12.



Picture 3: *Cross series plot*

Based on the project assumptions and data preprocessing we decided to use this methodology:

- Countries for which GIfS service is available were divided in two groups 'Mature' and 'Immature',
- based on economy, internet usage, etc.
- As the initial stage, an attempt was made to try to indentify 'trend brakes'[3] in DEPS series in advance

[2] For word translation in appropriate language, Google Translation™ was used.

[3] Significant changes in air traffic, e.g. decline due to recession in October and November 2008 shown in Picture 3.

using EXPLANATORY VARIABLE DATA series.
- To do so, time series modeling with explanatory variables was applied. This method allowed additional Google data quality check.

STATFOR operational request was to have simple way to obtain leading indication (automatic if possible). Leading indicator used should be reliable and stable in time. To avoid coincidence, misleading conclusions and to fulfill other operational requests decision was made to use only one, same leading indicator, i.e. search term, for the whole time period available, for all the countries.

3. DATA ANALYSIS

For time series modeling and forecasting, STATFOR uses SAS® Forecast Studio® 1.4 and SAS® Enterprise Guide® 4.1 edition. Forecast Studio is used for generating potential models, after that manually adjusted and improved in Enterprise Guide and supplemented with additional explanatory variables if considered necessary.

It was impractical (though theoreticly possible) to use GifS data with existing STF models for many reasons, so first decision was to create new models. Intention was not to model each time series manually, but automatically, because of the number of time series used (5 for each country). For automatic forecasting of large numbers of time series, only the most robust models should be used.

New models for each country were generated, where dependent variable was „DEPS" with explanatory variables „AIRLINE", „FLIGHTS", „TRAVEL" and „MARKET". There can be several causal factors that might or might not influence the dependent time series. The multivariate time series diagnostics determine which of the causal factors significantly influence the dependent time series. These diagnostics include cross-correlation analysis and transfer function analysis.

To explain cross-correlation, consider two real valued functions *f* and *g* that differ only by a shift along the x-axis. One can calculate the cross-correlation to figure out how much *g* must be shifted along the x-axis to make it 'identical' to *f*. The formula essentially slides the *g* function along the x-axis, calculating the integral of their product for each possible amount of sliding; when the functions match, the value of function is maximized.

General Transfer Function model is

$$Y_t = \alpha + \sum_{i=1}^{N} \sum_{j=0}^{\infty} \beta_{i,j} X_{i,t-j} + Z_t,$$

where X and Z are independent ARIMA[4] time series, *N* is number of variables, *j* is lag, α and β are parameters. This model allows Y to depend on current and past values of Xs. For simplicity, look at the model with one variable,

$$Y_t = \alpha + \sum_{j=0}^{\infty} \beta_j X_{t-j} + Z_t.$$

If $\beta_0 = \beta_1 = 0$, and for example, $\beta_2 \neq 0$, then Y respond two periods later to movements in X. X is now called leading indicator for Y because its movements allow Y movements prediction two periods ahead.

We decided to use only two month ahead forecast with Google data for several reasons (among else, this period is enough to get a leading indication for STF). Two month ahead forecasts were produced for three time periods; October and November 08 representing huge traffic decline and April and May 09 as potential 'out-turn month' showing slight recovery. Third period used was the last two month of available data July and August 09, as period with 'stable' traffic situation. For the initial results these three periods were practical and sufficient.

Out of sample test on forecast region data was used to measure model performance. This means that actual data, which exists, is compared against model forecast values. For model estimation out of sample data is not used, just the preceding data. Actual data for flights for two forecasted months were used to compare different models forecasts while preceding data was used to fit models.

Criterion used for model comparison Mean Absolute Percentage Error (MAPE) is frequently used measure of accuracy in a fitted time series value, specifically trending. MAPE is calculated as

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where A_t is actual value, F_t is fitted or forecasted value as in this case, and n is the number of points. This makes it a percentage error so one can compare the error of fitted time series that differ in level. It was applied on holdout sample (two months).

SAS® Forecast Studio® 1.4 generated several models for each country: one or two without explanatory variables (mostly Seasonal Winters and ARIMA models for dependant variable) and at least one model with explanatory variables. Some models with explanatory variables had more than one variable (for example, AIRLINE, TRAVEL and FLIGHTS) with different lags. Surprisingly though, in one case SAS Forecast Studio failed to create model with explanatory variables. For these 3 periods, MAPE was calculated for both, models without explanatory variables - MAPE and models with explanatory variables - MAPE VAR, Table 2. Results are shown for 'Mature market' only.

In practice, MAPE values below 10 are consider excellent, so the results achieved even with automatic modeling were good. The results in terms of numbers were not that important: models with and without variables might be improved by adjustments, STATFOR uses all the data available (from 1990) for modeling and

[4] AutoRegressive Integrated Moving Average

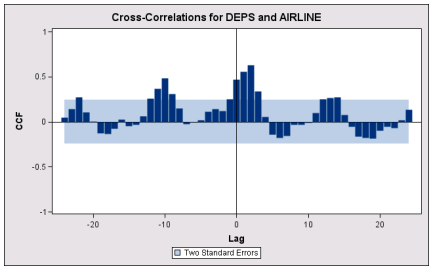
Google data is available as from 2004, etc. Besides, in almost every case forecasted DEPS values fell within the 80% confidence intervals (STATFOR practice), except for Oct & Nov 08 period.

Table 2: *MAPE for MATURE MRKET*

SAS Forecast	Oct&Nov 08		Apr&May 08		July&Aug 09	
MATURE MARKET	MAPE	MAPE VAR	MAPE	MAPE VAR	MAPE	MAPE VAR
Austria	3.27	5.22	2.97	**1.13**	1.85	4.93
Belgium/Lux	5.05	6.88	0.51	2.05	1.04	1.46
France	5.88	**1.55**	0.96	2.10	0.75	**0.36**
Germany	5.87	**5.29**	0.57	1.23	1.50	2.48
Ireland	3.79	**2.39**	7.87	**5.88**	1.05	**0.69**
Netherlands	5.84	**4.73**	0.89	1.78	1.35	1.85
Switzerland	4.93	6.26	0.73	3.79	0.85	3.61
UK	8.22	**6.74**	2.31	2.56	0.74	**0.51**
Denmark	7.60	**4.74**	4.78	**4.48**	1.53	2.67
Finland	3.42	**2.38**	7.76	8.97	2.53	9.37
Norway	3.47	no model	4.93	7.29	0.18	2.22
Sweden	8.91	**6.76**	8.03	17.99	3.46	**2.89**
Greece	3.12	3.67	5.34	**3.33**	1.80	**1.14**
Italy	6.83	9.81	3.22	5.19	0.81	1.66
Portugal	6.41	**4.41**	1.89	4.64	5.38	7.66
Spain	7.02	**4.45**	1.56	**0.44**	0.74	3.72

Interesting is that for period Oct & Nov 08, where most of the countries in 'Mature group' have seen huge decline in air traffic, automatically generated models with explanatory variables provided better results than models without (in 10 of 16 cases, shaded cells). This trend brake in DEPS series might be 'explained' with trend brake and significant change in some of the explanatory variables' components used in model. Other two, rather stable periods showed no improvement when using Google data as explanatory variable. Also is evident that models' outputs were better for these 'stable' periods compared to Oct&Nov 08 when DEPS series brakes down. This is just a confirmation that time series modeling and forecasting are performing well with relatively 'stable' series in time and needs 'extra information' when brakes occur.

By looking at cross-correlation tables and cross-series plots, project assumption that most of the people book two months ahead was in a way confirmed (for majority of countries and search terms) as shown in Picture 4.



Picture 4: *Cross Correlation Function Plot*

Cross correlation function peaks at Lag 2 (AIRLINE leads DEPS for 2 months). This might indicate that 2 months after searches had been done flights occurred, and no matter the search term used, the results were more or less the same.

t was expected for these terms to have a seasonal pattern for interest (in general, flight series has the same seasonal pattern) and this conclusion should not be 'take as given', but is promising one for further research.

5. CONCLUDING REMARKS

Trying to use Google Data as explanatory variable for two month ahead forecast in order to get leading indication for STF, as initial method used in the study, was beneficial for a number of reasons.

First, it was shown that separating European countries in two groups was rational because of poor Google data quality (to represent air transport demand for 'Immature' countries) and hence model results. Further methods were applied to 'Mature Market' only, which considerably decreased work load.

Cross-correlation plots and functions showed, for majority of countries and search terms, that most of the people search for the flights and perhaps book the tickets 2 months in advance. Although it is something more or less practice is familiar with this could be sort of 'evidence'.

Though we get an impression that in case of huge changes in air traffic, it is possible to use Google data, i.e. people searches for specific search terms, to improve short term forecast, a lot of operational requests were not satisfied. The most important were that it was not possible to use only one search term as explanatory variable to get leading information and vast time required to adjust and set model for forecasting, etc. So we decided not to use this method to get a leading indication.

Nevertheless, time series modeling with Google data and results obtained gave as an idea for new methods so enhancing our methodology. Speaking of data provided by Google Insights for Search service, it went beyond all our expectation in terms of quality for this study.

6. REFERENCES

[1] EUROCONTROL *Short Term Forecast, Septemner 2009. IFR Flight Movements 2009-2010.* EUROCONTROL/STATFOR/Doc 361, Brussels, September 2009

[2] SAS® High Forecasting Performance 2.3: *User's Guide*, SAS Institute Inc., SAS Campus Drive, Cary, North Carolina, 1st printing, December 2007

[3] Dickey, A, Brocklebank, J, *SAS® for Forecasting Time Series* second edition, SAS Institute Inc., SAS Campus Drive, Cary, North Carolina, April 2003

[4] Tashman, L, *Out of sample test of forecasting accuracy: an analysis and review,* International Journal of Forecasting, 16(4) (2000) 437-50

[5] Cohen, J, Garman, S and Gorr, W, *Empirical calibration of time series monitoring methods using receiveroperating characteristic curves*, International Journal of Forecasting 25(2009) 484–97