Reading tea leaves in the tourism industry

A Case Study in the Gulf Oil Spill

Hyunyoung Choi[1], Paul Liu[2]

Google Inc.

March 23, 2011

Abstract

There has been significant interest from the travel industry in using search data to predict hotel bookings and other travel-related expenditures in advance. When we compared Google Trends data with a reference travel dataset from Smith Travel Research, Inc, we find that searches for travel take place on Google typically a few weeks to about a month before the actual travel. We then used time series techniques to forecast lodging demand in the Gulf region following the Gulf oil spill and estimated the impact of the oil spill. We found that demand in the non-Gulf region rose while demand in the Gulf region decreased. The findings were consistent at the state level and metro level.

[1]email: hyunyoung@google.com
[2]email: ipaul@google.com

1 Introduction

An earlier study by Choi and Varian (2009) showed that Google Trends could help predict economic indexes. In their paper, they used panel data to show that tourism traffic prediction could be improved by using search trends and other factors such as the foreign exchange rate. In this paper, we extend the methodology and propose a methodology to estimate the economic impact. We also examine whether or not the methodology can be applied to more granular geographical areas.

The US travel industry is a roughly $700 billion industry, generating 7.4 million jobs. The Gulf oil spill disaster in April 2010 hit a travel industry already hard-hit by the deep recession beginning in 2007. While by early 2010 some parts of travel industry in the United States had started showing signs of a modest recovery, it appeared likely that the oil spill would result in travelers (especially vacation travelers) shifting their plans to other destinations not affected by the oil spill. Can we see this shift in the search trends data? As many consumers spend significant time researching their travel plans online, this seems likely, though far from certain.

In this case, we had at least three points of references, which unfortunately, gave rather disparate answers. Oxford Economics recently estimated that the impact of the oil spill on travel could be as great as $22.7 billion and could last up to 3 years. This was based on an extrapolation from studies of recent natural and manmade disasters, including previous oil spills, hurricanes, SARS and H1N1. MasterCard's US lodging sales data showed a spending increase of 3.6% year-over-year in 2010 compared to a 6.2% decline in 2009, highlighting the difficulty of making year-over-year comparisons given the state of the economy.[iii] Finally, a consumer survey conducted by AAA in May 2010 found that while 1 in 8 respondents had changed plans, none had attributed the change to the oil spill, and of the remainder who had not changed plans, less than a quarter said that they were 'likely' or 'very likely' to change their plans subsequently.[iv]

We used data from Smith Travel Research, Inc. on actual reported hotel rooms sold (excepting complimentary rooms) in each metro area in the Gulf region and adjacent areas. To examine the effects of the oil spill, we presumed a differential impact between areas along the Gulf coast of Florida, where the oil spill have been perceived to have an impact and areas along the Atlantic coast of Florida, where the oil spill was probably not perceived to have any impact. Unlike in the earlier application examined by Choi and Varian (2009), granularity of these data (daily)

[iii] Data Source: MasterCard Spending Pulse

[iv] Data Source: http://www.aaasouth.com/documents/OilSpillSurvey.pdf

and timeliness are not issues. But in this case the delay between the research phase of planning the travel and the actual travel itself may be significant. So while "impact" can ostensibly be measured from the actual data, the real interest here lay in the ability of search data to anticipate the impact and possibly allow for mitigation (e.g., in the form of increased advertising), especially if more fine-grained geo-targeting is successful.

Search data have been used in various contexts, including capital markets (Vlastakis and Markellos (2010), Da et al (2010)), entertainment (Goel et al (2010), Choi and Shideler (2010), Choi and Slotwiner (2009)), labor markets (Askitas and Zimmermann (2009), Suhoy(2009), D'Amuri and Marcucci (2010)), real estate markets (Wu and Brynjolfsson (2009), Bradhan et al (2010)), healthcare (Polgreen et al (2008), Brownstein et al (2009), Corley et al (2009), Hulth et al (2009), Pelat et al. (2009), Wilson and Brownstein (2009), Valdivia and Monge-Corella (2010)) and economic indices (Ettredge et al (2005), Schmidt and Vosen (2009), Huang and Della (2009)).

Section 2 discusses the Google Trends data, Section 3 discusses the oil spill impact analysis using Google Trends, Section 4 demonstrates how to build a forecast model at the metro level and Section 5 concludes.

2 Background

Google Trends[v] provides an index of the volume of Google queries by geographic location and category measured against total query volume. Queries can be filtered by query string and query category. The number represents the normalized query shares by each filter.

For a given query of interest, Google Trends uses a 'broad match algorithm such that any query string containing the query of interest as a substring is included in the statistics. For example, the Google Trends data for the query string 'florida' includes the contributions of 'disney florida,' 'universal florida,' 'orlando florida,' 'florida disney world' and so on.

A given query string can be assigned to a category based on the context of the session in which the query string appears. A session here represents a series of completely anonymized search queries performed within a relatively short time period and helps in classifying ambiguous words. For example, the query 'florida' may be classified as `Local`, `Society`, `Travel`, or something else depending on the context. Within the `Travel` category, Google uses eight subcategories – `Adventure`

[v] `http://www.google.com/insights/search/`

Travel, Bus & Rail, Hotels & Accommodations, Air Travel, Car Rental & Taxi Services, Cruises & Charters, Attractions & Activities, Vacation Destinations. Travel planning involves a number of steps – booking the flight (Air Travel), reserving a hotel (Hotels & Accommodations), booking a rental car (Car Rental & Taxi Services), and then searching for things to do (Attractions & Activities). These correspond to given subcategories in the parentheses, which we rely on to identify travel-related search trends.

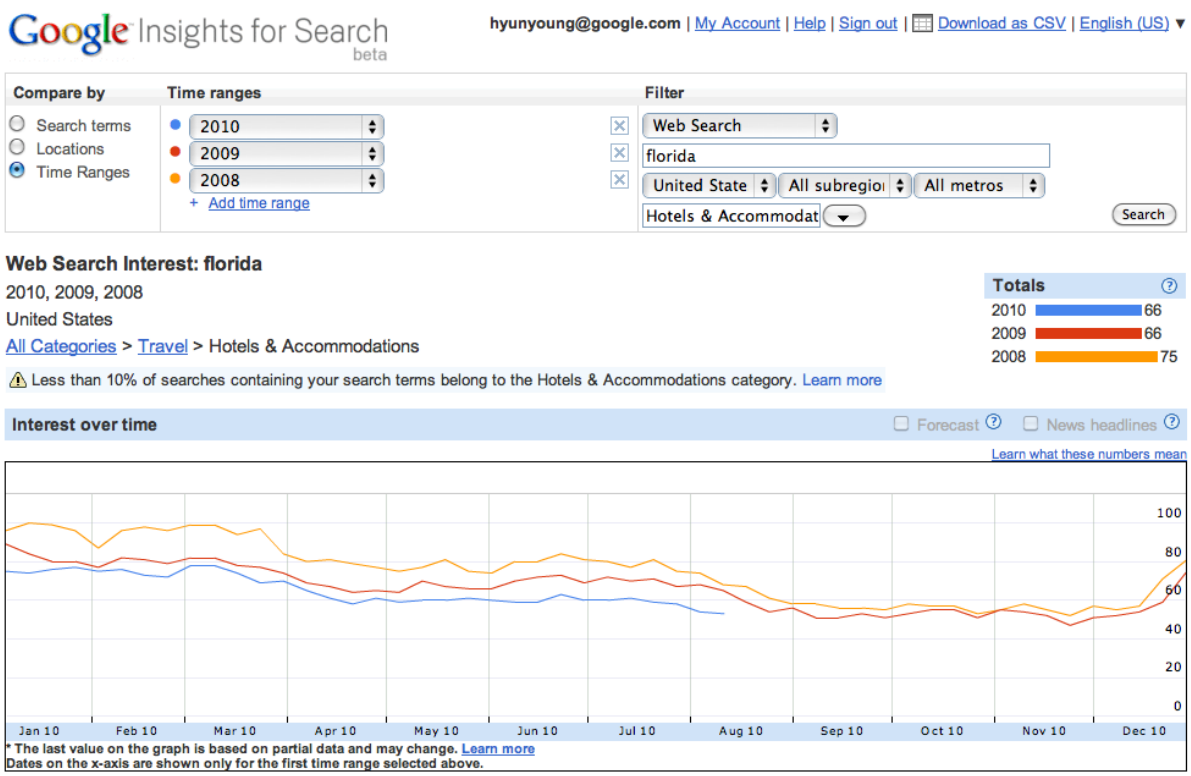
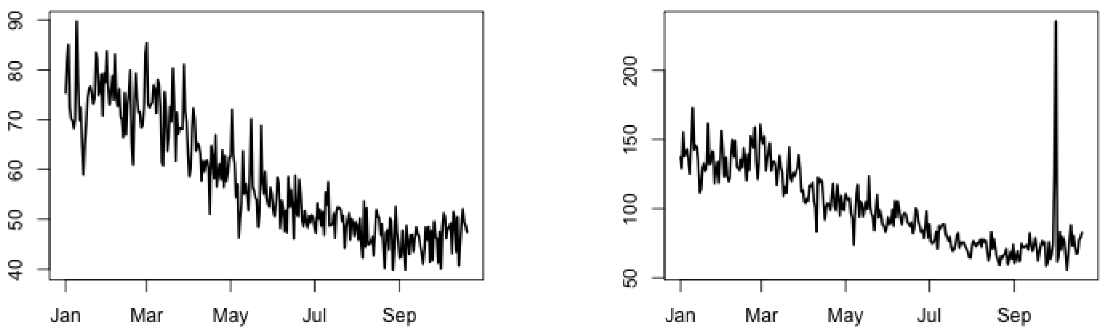


Figure 1: Google Search Insights

Figure 1 shows searches for 'florida' within the content of Hotels & Accommodation from 2008 to 2010. The totals above the graph are the average value of the search index by year. On average, the search volume at 2009 was down by 9 units. Through mid-August, the search interest on 'florida' had been lower than the equivalent period in 2009, even prior to the oil spill in April.

Because consumers often search for flights, hotels, car rentals, and attractions at the same time, the trends for these subcategories are highly correlated to each other. Figure 2 shows search trends from the two categories, `Hotels & Accommodation` and `Attractions & Activities`. The correlation between these two time series is 0.8376. Note the unusual search pattern on 9/30/2010 and 10/1/2010, involving news on the 'Jules Undersea Lodge.' Excluding these dates, the correlation rises to 0.9070. We use the search for 'florida' aggregated across these subcategories to measure the travel intent to Florida.



(a) Hotels & Accommodations

(b) Attractions & Activities

Figure 2: Google Trends Comparison with search keywords 'florida' at 2010

3 Gulf Oil Spill Impact Study

An explosion on BP's Deepwater Horizon oil rig in the Gulf of Mexico on April 10, 2010 resulted in an oil spill that released an estimated 4.9 million barrels of crude oil into the Gulf before it was capped on July 15, 2010. Oil began washing up on the shores of Mississippi, Alabama, Louisiana, and parts of the west coast of Florida in early June, resulting in beach closures and fishing bans (Figure 3) .

How can we measure the impact of the oil spill? Because Gulf destinations actually represents a wide variety of destinations, it was helpful to narrow the analysis to comparable geographical areas. It appeared that Florida could serve as a good test case, as the coastal destinations within Florida were arguably more homogenous than other coastal destinations in Louisiana, Alabama,

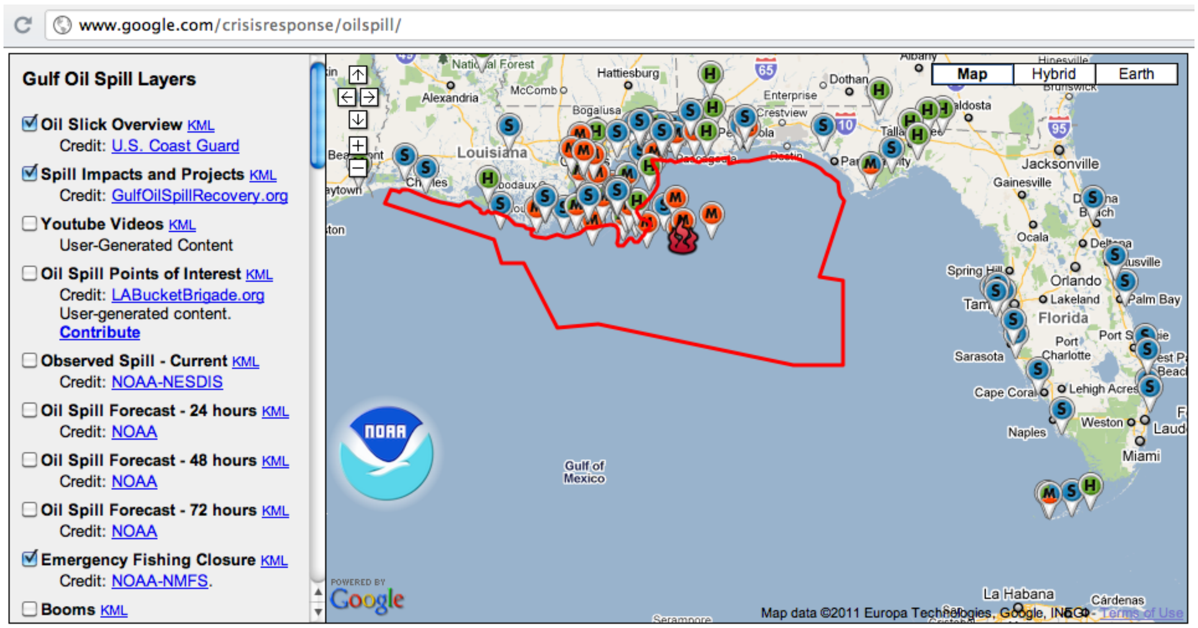


Figure 3: Google Crisis Response - Gulf of Mexico Oil Spill

and Mississippi. We chose to compare lodging impacts along the west (Gulf) coast of Florida, which was probably impacted by the spill, to lodging impacts along the east (Atlantic) coast of Florida, which may have benefited from travelers going there instead. Data for the Gulf coast consist of 11 cities including Tampa and Sarasota, while data for the Atlantic coast consist of 8 cities including Miami.

Can Google Trends help us understand the demand? Recall that when the oil spill occurred, the US was coming out of the deepest recession in many decades. Thus it was important to take this into account so as not to confound the impact of the recession with any impact from the oil spill. We controlled for the effects of the recession on Florida overall by using Search Trends travel queries targeted for Florida. From Figure 1, one can readily see search query share declining as the recession deepened, and more recently showing signs of recovery as time went on. Thus the indications are that using Search Trends travel queries targeted for Florida is likely to do a reasonable job of controlling for the state of the economy.[vi]

The specific methodology we used to estimate the effect of the oil spill included the following

[vi]Recall that the presumption here is that travelers going to Florida will choose one coast or the other. Based on conversations with travel industry experts, this does not seem unreasonable. However, if travelers instead choose destinations outside Florida as a result of the oil spill, our Search Trends data may be potentially affected as well.

steps:

1. Collect weekly lodging and search data.
2. Build a forecasting model with the data before the oil spill event.
3. Make forecast and set it as the baseline.
4. Compare the baseline with the actual with the forecast. If the actual demand is lower/higher than the forecast, we know that the visitors changed their vacation destination.

Figure 4(a) depicts the weekly lodging demand along the Gulf and Atlantic coasts of Florida. They are highly correlated with a correlation of 0.9. Figure 4(b) shows lodging demand compared to Google search volume along the Atlantic coast of Florida. The demand exhibits autocorrelation and the seasonality (Figure 5(a)) and the search trends lead the demand by 4 weeks except for the summer vacation period from June to August (Figure 5(b)). The correlation between search trends and demand in the summer vacation is not as apparent due to the scaling in the graph, but actually it is positive and quite similar to that found during the rest of the year ($\rho = 0.55$ during the summer vacation season and $\rho = 0.57$ during the remainder of the year). In any case, a seasonal AR model (e.g., with a 52 week lag) should control for the autocorrelation.

$$\text{Model 1} \quad : \quad (1-\Phi_1 B^{52})(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)(Y_t-\beta \text{Google}_t) = \mu + e_t \tag{1}$$

$$\text{Model 2} \quad : \quad (1-\Phi_1 B^{52})(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)Y_t = \mu + e_t \tag{2}$$

where $e_t \sim N(0, \sigma^2)$ and B is the backshift operator. In Model 1, Y_t represents the log of the demand at time t and Google_t is the Google search index. The first part of the model explains the seasonality, the second part of the model explains the autocorrelation, and the last part of the model explains the correlation between lodging demand and Google search volume. Weekly data from January 2005 to the week of oil spill was used to train the model; data from the following week until the end of June 2010 was used to measure the oil spill impact. Table 1 displays the estimates and standard errors for Model 1, using maximum likelihood estimation. The model fits very well with R^2 close to 1.

$$\text{PE}_t = Y_t - \hat{Y}_t = \log(y_t) - \log(\hat{y}_t) \tag{3}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} |\text{PE}_t|$$

Table 1: Estimates and Standard Error from Model (1)

		ϕ_1^i	ϕ_2^i	ϕ_3^i	ϕ_4^i	Φ_{52}^i	β^i	μ^i
Non-Gulf (i = 0)	Est	0.24	0.2	0.24	0.07	0.82	13.66	48.54
	s.e.	0.06	0.06	0.06	0.06	0.03	0.09	12.11
Gulf (i = 1)	Est	0.21	0.25	0.18	0.21	0.87	12.92	20.30
	s.e.	0.06	0.06	0.06	0.06	0.03	0.11	12.71

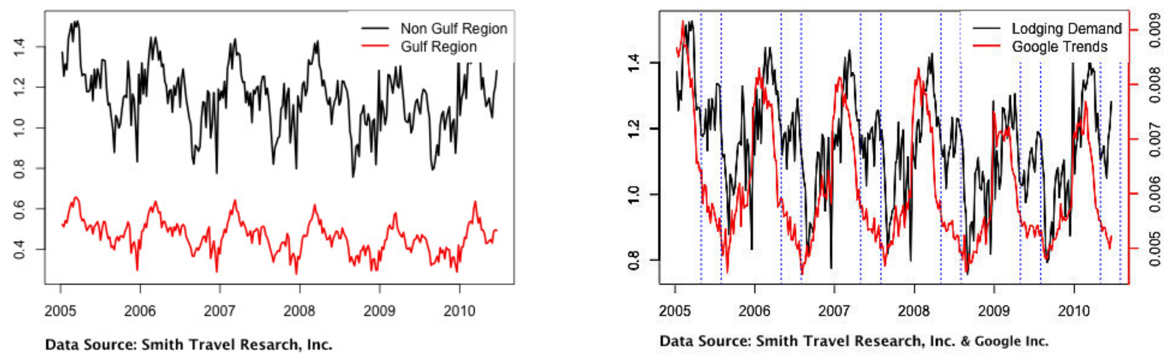
The MAE (mean absolute error) measures how closely the forecasted values line up with actual lodging demand. It is defined in Equation (3). The MAEs for the nine weeks following the oil spill were 5.78% and 6.40% for the Atlantic and Gulf coasts, respectively. For the nine weeks following the oil spill, we can see the effect of the spill by looking at how actual lodging demand varied relative to the forecast taken prior to the spill. On average, lodging demand along the Gulf coast of Florida were 4.25% less than predicted, while lodging demand along the Atlantic coast of Florida was 4.89% greater than predicted.

Without Google Trends (Model 2), the MAEs for the nine weeks following the oil spill were 10.58% and 3.87% for the Atlantic and Gulf coasts, respectively. On average, lodging demand along the Atlantic coast was less than predicted by 8.07% and demand along the Gulf coast was greater than predicted by 2.15%, which suggests the counterintuitive result of no impact on lodging demand from the Gulf oil spill. But as we know, this model lacks the Google Trends variable.

Thus our findings are consistent with our prior expectation that there could have been some shift in travel plans from parts of the Florida coast affected by the oil spill to parts of the Florida coast not affected. Note, however, that since Florida was arguably less impacted by the Gulf oil spill than the other affected states of Louisiana, Mississippi, and Alabama, these impact estimates may be something closer to a lower bound. And note that this is only an estimate for the period immediately following the spill; we leave estimates of long-run impact and recovery for future analysis.

4 Metro level data analysis

Usually, traffic data get noisier when going from state level data to finer geographic gradations such as metro level. However, the same forecasting methodology can still be applied. With the



(a) Demand($1 million) by Region

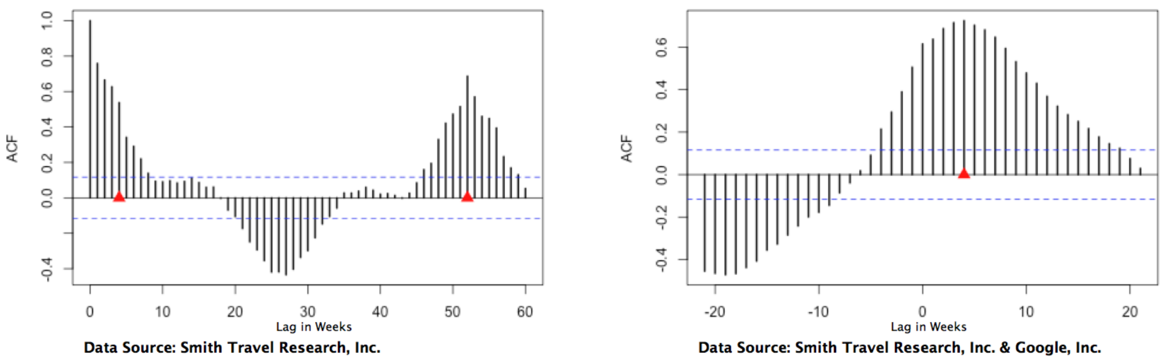
(b) Demand($1 million) vs. Google Trends

Figure 4: Lodging Demand & Google Search Volume on 'florida'

same logic as Section 3, we used Google searches for a particular metro name filtered to the travel category to proxy for travel interest. For example, a search on 'Orlando' under travel related categories is likely to be part of a user's trip planning session. In some cases cities with the same name can be difficult to disambiguate, though this is not always the case. For example, while 'Orlando, FL' is one of most popular tourist destination, 'Orlando, OK' is a small city with an estimated population of less than 300. In this case, it seems reasonable to think that searches for 'Orlando' by and large will approximate travel interest to 'Orlando, FL.' However, searches for 'Naples' may not be as clear because nine different cities in US share that name, and while Naples, Florida, is a popular tourist destination, there is some possibility that the search was intended to target one of the other cities. In this case, the obvious solution, to add a state identifier such as in 'orlando fl', but searches including the state identifier typically have much lower volumes, often not enough to pass the privacy threshold set by Google.

With this in mind, we focus on three cities where disambiguation is not an issue: Miami, Tampa, and Sarasota.[vii] Figure 6 depicts the demand and search trends from the three metros. Miami and Tampa, besides being two of the top vacation destinations in Florida, also have substantial metropolitan populations of 4 million and 3.4 million respectively. Sarasota had a

[vii]We recognize that there is a Miami in Ohio but other cities with these names are clearly dominated by the ones in Florida.



(a) Autocorrelation of Demand

(b) Cross correlation

Figure 5: Autocorrelation & Cross Correlation

population of about 50,000 in 2009 and is clearly a smaller vacation destination, but results for this city show that relative lodging demand does not detract from the analysis. With the same logic and testing methodology we used in Section 3, we found that search queries targeting these three cities lead demand by four weeks.

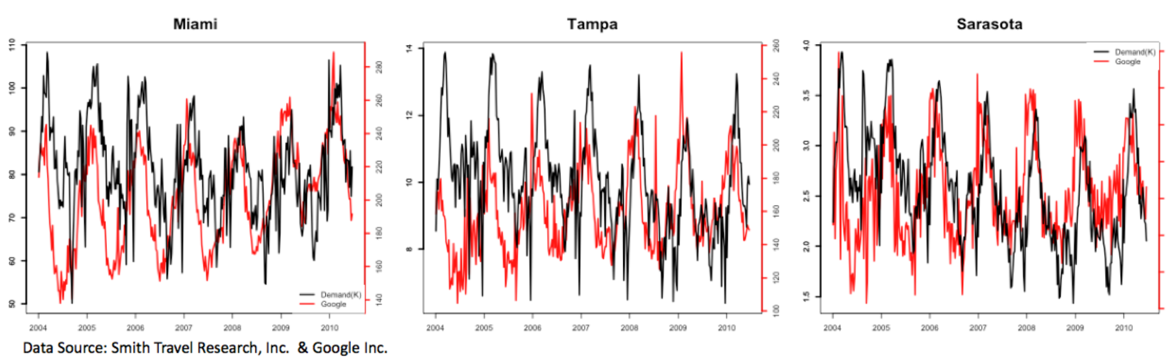


Figure 6: Demand vs. Search Trends at Selected Florida Metros

Table 2 summarizes the estimates after fitting Model 1 to each metro data. Lodging demand exhibits high autocorrelation and strong seasonality and search trends data are positively correlated to the lodging demand. In each case, R^2 is close to 1 and MAE with one-week ahead prediction for the 20 weeks from February is less than 5% for these three metros. Obviously the prediction is better for the bigger metros because the data gets more regular as demand and

Table 2: Metro level Estimates and Standard Error from Model (1)

		ϕ_1	ϕ_2	ϕ_3	ϕ_4	Φ_{52}	β	μ	MAE
Miami	Est	0.47	0.13	-0.02	0.14	0.62	0.003	12.09	3.69%
	s.e.	0.06	0.06	0.06	0.05	0.05	0.001	0.09	
Tampa	Est	0.26	0.17	0.10	0.23	0.80	0.010	11.98	4.82%
	s.e.	0.05	0.05	0.05	0.05	0.03	0.006	0.07	
Saratoga	Est	0.55	0.12	0.15	0.03	0.70	0.013	10.46	4.89%
	s.e.	0.03	0.05	0.03	0.07	0.03	0.011	0.07	

search volume increases, but the method performs well even for a small city like Saratoga as long as the search is targeted to that particular city.

5 Conclusion

In this paper we found that once again Google Trends is a useful predictor of subsequent economic data, this time leading realized lodging data by four weeks. Previous speculation that Google Trends data could be used to identify turning points has been confirmed here in the case study of the Gulf oil spill where we find that parts of Florida were negatively impacted while other parts of Florida appear to have benefited. We then showed that the methodology applied at more macro levels such as the state-level could be successfully applied to finer geographical areas such as metro areas, suggesting that a more generalized tool could prove very insightful to the travel industry.

References

[1] AAA(2010) "*Consumer Pulse: Intent to Visit Florida*" `http://www.aaasouth.com/documents/OilSpillSurvey.pdf`

[2] Askitas, N. & Zimmermann, K. (2009) "Google Econometrics and Unemployment Forecasting" *Institute for the Study of Labor*, No. 4201.

[3] Bardhan, A., Hicks, D., Kroll, C. & Yu, T. (2010) "The Past as Prologue? Business Cycles and Forecasting since the 1960s" *Fisher Center Working Papers –*

[4] Brownstein, J., Freifeld, C., and Madoff, L. (2009) "Digital Disease Detection ? Harnessing the Web for Public Health Surveillance" *The New England Journal of Medicine*

[5] Choi, H. & Varian, H.(2009) "Predicting the present with google trends" *Google Technical Report*, http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

[6] Choi, H. & Varian, H. (2009) "Predicting initial claims for unemployment benefits" *Google Technical Report*, http://research.google.com/archive/papers/initialclaimsUS.pdf

[7] Choi, H. & Shideler, M. (2010) "Predicting Game Sales with Google Trends" *Google Working Paper*

[8] Choi, H. & Slotwiner, D.(2009) Predicting TV Rating with Google Trends, *Google Working Paper*, Google, 2009

[9] Corley, C., Mikler, A., Singh, K. & Cook, D. (2009) "Monitoring Influenza Trends through Mining Social Media" *Proceedings of the 2009 International*

[10] Da, Z., Engelberg, J. & Gao, P. "In Search of Attention" *AFA 2010 Atlanta Meetings Paper*

[11] D'Amuri, F. and Marcucci, J. (2010) "'Google It!' Forecasting the US Unemployment Rate with A Google Job Search Index"

[12] Ettredge, M., Gerdes, J. & Karuga, G. (2005) "Using web-based search data to predict macroeconomic statistics" *Communications of the ACM*, Vol 48, No 11.

[13] Goel, S., Hofman, J., Lahaie, S., Pennock, D. & Watts, D. (2010) "The Predictive Power of Search" *Proceedings of the National Academy of Sciences*

[14] Huang, H. & Della Penna, N. (2009) "Constructing Consumer Sentiment Index for U.S. Using Google Searches" *University of Alberta Working Papers Series*, No 2009-26.

[15] Hulth, A., Rydevik, G. & Linde, A. (2009) "Web Queries as a Source for Syndromic Surveillance" *PLoS ONE*, Vol 4

[16] Oxford Economics (2010) "*Potential Impact of the Gulf Oil Spill on Tourism*" http://www.ustravel.org/sites/default/files/page/2009/11/Roadmap-Oxford_2pager.pdf

[17] Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A. & Valleron A. (2009) "More diseases tracked by using Google Trends." *Emerging Infectious Diseases*, 15:1327?8

[18] Polgreen, P., Chen, Y., Pennock, D. & Nelson, F. (2008) "Using internet searches for influenza surveillance" *Clinical Infectious Diseases*, 47:1443-8

[19] Schmidt, T. & Vosen, S. (2009) "Forecasting Private Consumption: Survey-based Indicators vs. Google Trends" *Ruhr Economic Papers Series*, No 0155.

[20] Suhoy, T. (2009) "Query Indices and a 2008 Downturn: Israeli Data" *Bank of Israel Discussion Paper*, No. 2009.06

[21] Valdivia, A. & Monge-Corella, S. (2010) "Diseases tracked by using Google Trends, Spain" *Emerging Infectious Diseases*

[22] Vlastakis, N. & Markellos, R. (2010) "Information Demand and Stock Market Volatility" *Athens University*

[23] Wu, L. & Brynjolfsson, E.(2009) "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities" *ICIS 2009 Proceedings*, Paper 147.

Acknowledgement

The authors also wish to acknowledge Smith Travel Research, Inc("STR") which provided data that have been used in the analyses presented here. STR do not accept responsibility for any inferences or conclusions derived from the data analysis presented in this paper.

Based in Hendersonville, Tennessee and founded in 1985, Smith Travel Research, Inc. provides clients – including hotel operators, developers, financiers, analysts and suppliers to the hotel industry – access to hotel research with regular and custom reports covering North America, Mexico and Caribbean. STR provides a single source of global hotel data covering daily and monthly performance data, forecasts, annual profitability, pipeline and census information. STR founded the STR family of companies and is proudly associated with STR Global, RRC Associates, STR Analytics, and HotelNewsNow.com.