Incorporating Google Trends Data Into Sales Forecasting

TONYA BOONE, RAM GANESHAN, AND ROBERT L. HICKS

PREVIEW *Forecasters are learning that Internet search data can be valuable additions to their models. In this new study, Tonya, Ram, and Robert show specifically how to take search-based data from Google Trends and build them into an individual firm's sales forecasting model. Their case study shows the potential for improved accuracy.*

INTRODUCTION

Sales and Operations Planning (S&OP) managers have traditionally relied on historical sales data to forecast demand. These forecasts serve as the basis for planning supply-chain activities—sourcing, making, and distributing to the customers. They are not perfect, however, and forecast errors are a source of considerable risk to revenues. For example, overestimating sales leads to excessive markdowns, and underestimating sales can result in lost revenues.

The last decade has seen the widespread use of digital technologies—websites and Internet-enabled objects—that collect an enormous amount of data about products, processes, and customers. In addition to traditional sales data, S&OP managers now have access to data both inside and outside the firm that can be used to improve forecasts and enable better efficiencies in consequent operations.

Examples of data available within the firm include digital clickstreams, sensors, tags, beacons, trackers, and other smart devices that collect pertinent data in real time. Significant data are also collected outside the firm through social-media chatter on the firm's products and services: news, blog and forum entries, and trend-spotting data, all of which are free and publicly available.

The challenge for today's managers is to integrate these data into S&OP processes—a challenge because the data are generated in real time (high velocity), in large volumes, and in myriad varieties. Based on our experience with an online retailer, our hope in this article is to provide one efficient way to integrate publicly available trend data into product forecasting. We illustrate the use of two tools that are right at the S&OP manager's fingertips: *Google Analytics* (the internal data generated by customer transactions) and *Google Trends* (a freely available tool from Google that quantifies trends on the Internet).

Google Analytics

Google Analytics is a service that tracks the traffic and e-commerce transactions on a website. S&OP managers, especially of online companies or those that have online divisions, can gain insight on who is visiting their site, how the visitors arrived there, what they browsed, and the percent of visitors that were "converted" into customers (i.e., purchased a product from the website).

Customer transactions by stock-keeping unit (SKU) can also be easily and automatically retrieved from Google Analytics in real time. The sales data in the ensuing models, for example, were automatically extracted from Google Analytics. While the basic service is free of charge, Google also offers a premium service for a fee. For more on Google Analytics, see Ram Ganeshan's article in the Spring 2014 issue of *Foresight*.

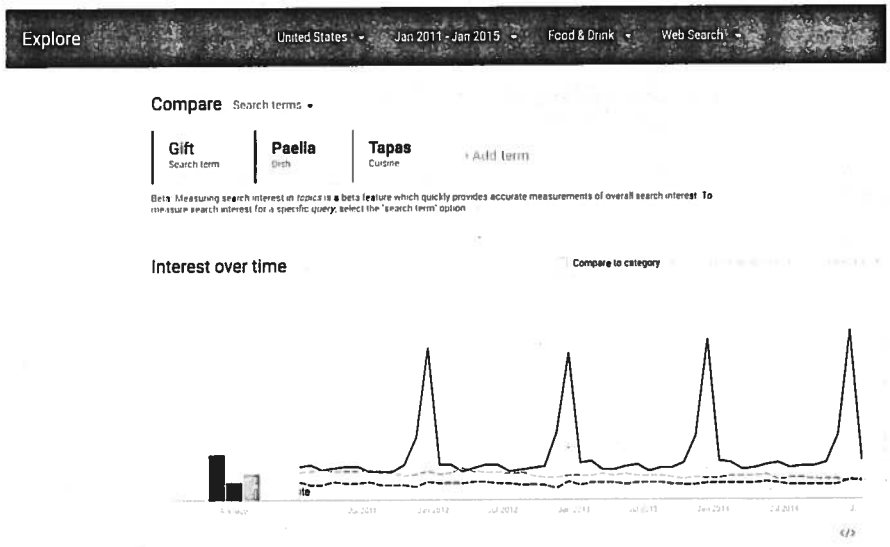
Google Trends

Google Trends is a publicly available service that provides an index of search queries

Key Points

- The last decade has seen the widespread use of digital technologies—websites and Internet-enabled objects—that collect an enormous amount of data about products, processes, and customers.
- The challenge for today's S&OP managers is to integrate these data into their processes—a challenge because the data are generated in real time (high velocity), in large volumes, and in myriad varieties.
- We illustrate the use of two tools that are readily available to S&OP managers for incorporating trend data into product forecasting: *Google Analytics* (the internal data generated by customer transactions) and *Google Trends* (a freely available tool from Google that quantifies trends on the Internet).
- Adding search terms to traditional forecasting models improves model fit.

Figure 1. Google Trends



by search term, category, and geographic location. For example, **Figure 1** shows the trend information for the terms "gift," "tapas," and "paella" in the category "Food and Drink" in the United States from 2004 to the present.

The index is normalized from 0 to 100: the higher the index, the higher the search intensity for the chosen search term. With the widespread use of the Internet to research purchase decisions, the premise is that trends for certain terms foreshadow sales of certain SKUs or product categories. In our example, the trends for "gift," "tapas," and "paella" could indicate the intent to purchase certain specific SKUs that are highly correlated with these search terms.

A CASE STUDY

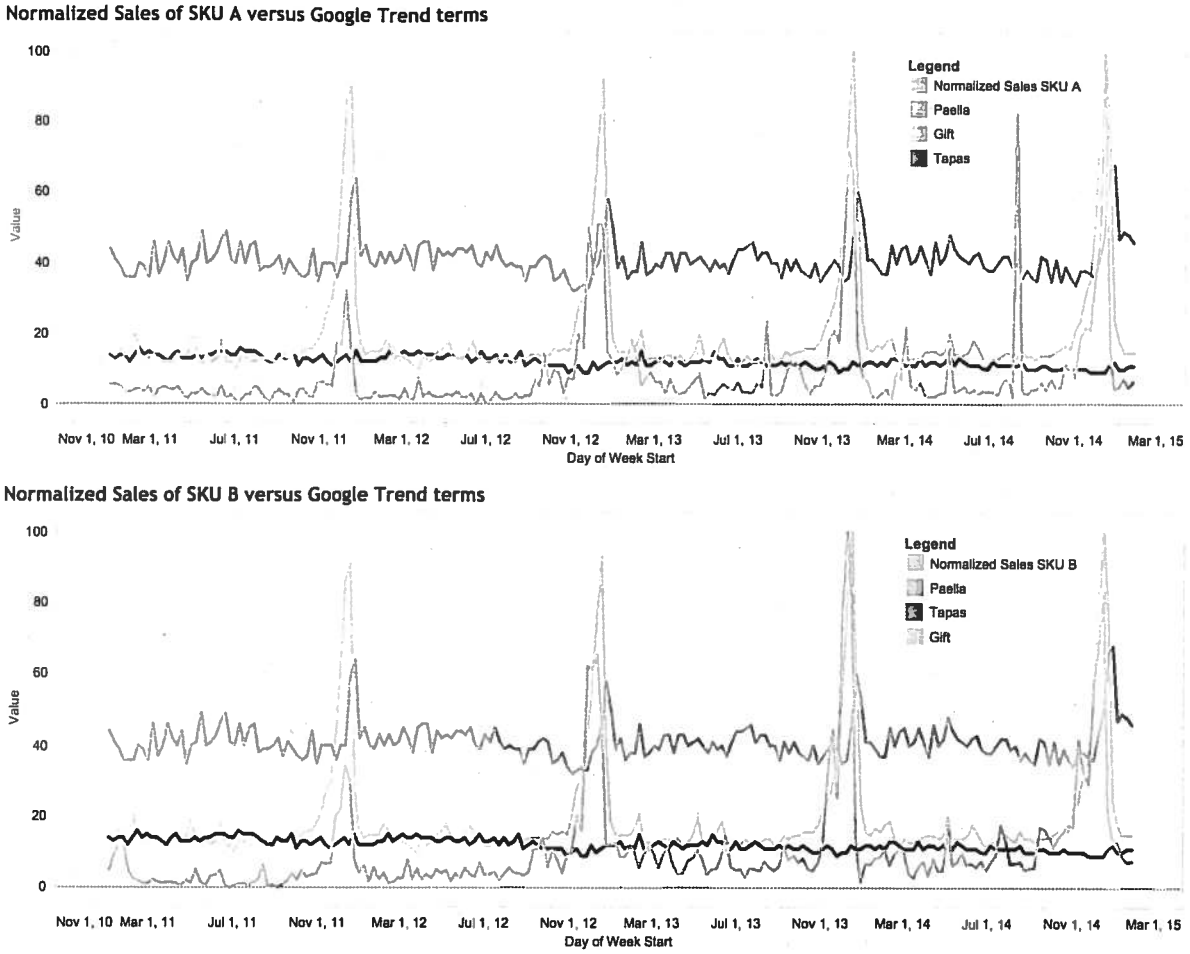
Researchers have shown that Google Trends data can be successfully used to predict social and economic trends. Hal Varian, Google's Chief Economist (Choi & Varian, 2009, 2012), shows how forecasts of macroeconomic trends in retail, automotive, housing, and travel can be improved by incorporating Google Trends terms into predictive models.

Google engineers working with the Centers for Disease Control (Ginsberg and colleagues, 2009) showed that flu outbreaks can be predicted early by tracking search data on flu-related topics.

In a recent article in this journal, Trosten Schmidt and Simeon Vosen (2013) wrote on how trend data can be used to forecast consumer consumption patterns. Their models show that incorporation of search terms reduced the root mean square error by 66% and improved out-of-sample forecast errors significantly. Their models, however, measured aggregate economic trends – not specific firm or product dynamics.

The key insight from these studies is that Google Trends data show promise for predicting macroeconomic activity early and accurately, enabling decision makers to react to such changes in economic and social activities in a more effective manner.

Figure 2. Tracking of SKU Sales and Search-Term Trends



S&OP managers are ultimately interested in improving SKU-level forecasts. Incorporating Google Trends data that is free, fine-grained, available in real time, and easy to integrate into the S&OP process can potentially yield significant benefits.

To illustrate, we've chosen two specialty-food SKUs—call them A and B—sold by a retailer specializing in food and cookware (we have left the SKUs unnamed to protect the identity of the retailer). These SKUs are often given as a gift, and are also popular as ingredients in Spanish-inspired cuisine, especially in appetizers ("tapas") or in rice ("paella").

Figure 2 shows the sales for SKUs A and B (normalized between 0 and 100) and the corresponding Google Trends information for "gift," "paella," and "tapas." Trends in SKU sales seem to track search-term trends, especially for the terms "gift" and "paella." The ensuing model illustrates a straightforward way to integrate these trends into the forecast.

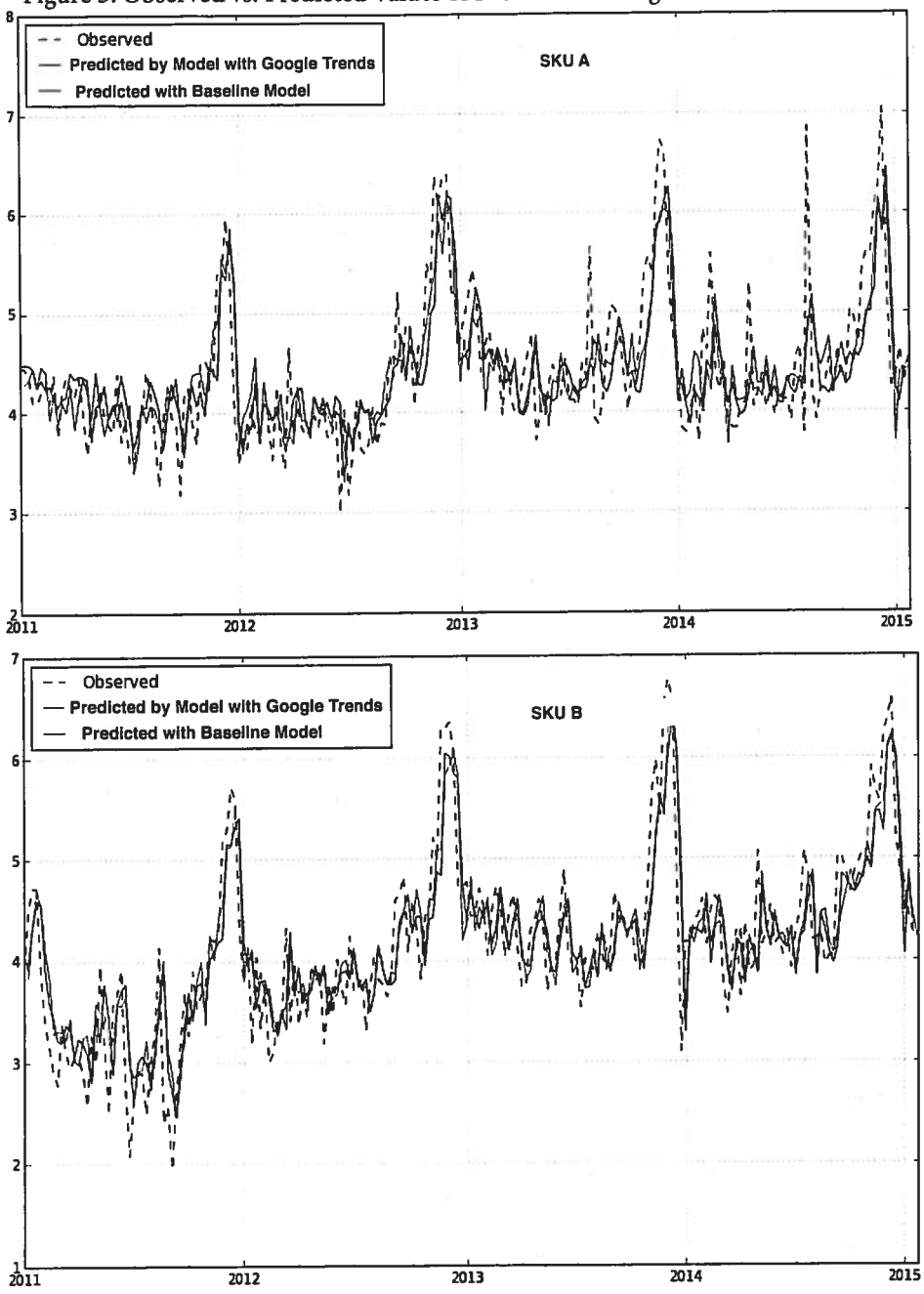
The Baseline Model

We model sales in any given week t (denoted as s_t) as a function of the price of the SKU in that week,), and the sales in each of the previous four weeks. The SKUs in question are promoted heavily during the Christmas holiday season, so we use a dummy variable that captures the peak for December sales. The sales and price variables are transformed into logs to estimate our baseline regression model:

$$LogS_t = K + \alpha log P_t + \sum_{i=1} \beta_i log S_{t-i} + \gamma D_{12} + \varepsilon_t$$

The intercept K, α, $\beta_{i\,i}$, and y are parameters to be estimated ε_t is the estimation error.

Figure 3. Observed vs. Predicted Values of Baseline vs. Google-Trends Models



Model with Google Trends

We incorporated Google Trends data by adding additional variables that correlate with search terms. The choice of which search terms to include is a combination of the manager's intuition and trial and error of potentially related searches.

In our example, SKUs A and B are traditionally bought as Christmas gifts or used as ingredients in tapas and paella. We decided to search on these terms as well as on the name of the SKU, since customers presumably are also looking specifically for this item. G^{gift}, for example, is the search intensity of the term "gift" in the corresponding week. G^{sku} is the search intensity associated with the SKU's name.

We did not find a significant lag between the search intensity and the purchase of the SKUs in question. For certain product categories – consumer electronics, for example – we would expect that a customer will search and research a product well in advance of a purchase. Depending on the specific SKU,

a lag between the search intensity and the sales can be built into the model.

Our regression model that includes our search terms can be represented as:

$$LogS_t = K + \alpha logPt + \sum_{i=1}^{4} \beta_i logS_{t-1} + \gamma D_{12} + \delta logG^{gift} + \theta logG^{paella} + \omega logG^{tapas} + \mu logG^{SKU} + \varepsilon_t$$

As in the baseline model, K is the intercept, and β_i, $\gamma, \delta, \theta, \omega$, and μ are model parameters that need to be estimated.

We implemented the estimation algorithm in Python, an object-oriented programming language (***http://www.python.org***) with the ability to connect, download, visualize, and analyze Google Analytics data from company servers as well as trend data from the Internet. If the S&OP manager wants to automate the forecast process using trend data, tools such as Python can prove to be very useful. Interested readers can contact the authors for the Python implementation of this regression model.

These models can also be implemented using commonly available statistical software packages. However, in most cases, the sales data from company servers and the trend data from the Internet must first be downloaded separately and then merged prior to the analysis.

Comparison of Model Results

For the baseline model, we find that weekly sales are strongly influenced by the same-week price and previous-week's sales. Lagged sales of two weeks and longer have a smaller impact for the chosen SKUs A and B, but could be significant when predicting sales of other SKUs.

When we included the search terms in the regression model, we find that the terms "gift" and "paella" are significant predictors, while the term "tapas" is less so.

Figure 3 shows the observed and predicted sales for both SKUs (and the corresponding residuals). For SKU A, the root mean squared error (RMSE) for the baseline model was 7.12; but when trend terms were added, the RMSE dropped to 6.72, a 5.6% decrease. For SKU B, the RMSE for the baseline model was 7.13 and the model with trends yielded a RMSE of 6.81, a 4.4% decrease.

It's clear, then, that adding search terms to the baseline model improved the fit of our model. However, we would need to perform *out-of-sample tests* of the model over many SKUs and multiple product categories before we can conclusively say that inclusion of relevant search terms translates into better forecasts. Still, potentially modest reductions in forecasting error can have a significant impact on the bottom line, especially when margins are low. The efficiency of supply chains that fulfill these SKUs can be increased with more accurate demand signals.

PRACTICAL CONSIDERATIONS

Our intent is not to say that Google Trends data always do a better job at predicting sales. Rather, it is that S&OP managers now have an additional resource that is free, fine-grained, and that shows promise for improving forecasts.

Extended Uses

S&OP managers can use Google Trends data effectively in other situations. When a new SKU is introduced, patterns for sales can be planned not merely on the basis of similar SKUs but by incorporating potential search terms that lead the customer to the product.

Trend data may also prove useful when planning promotions. SKU sales data can be correlated with search terms such as "free shipping" or "two-for-one" to get a better picture of how these factors affect sales.

In addition to marketing insights, trend patterns can lead to interesting statistical insights. In our example, customers likely purchased SKUs A and B when in fact they were originally looking for other things, like recipes for paella or when searching for a nonspecific gift. Such information can be used to run more narrowly and sharply targeted campaigns to attract customers.

Challenges

Incorporating trend data poses challenges. First, the manager needs to decide which search terms to include in the model. Some

search terms are obvious, but there may well be others with greater explanatory power. To find relevant and useful search terms that can be used across multiple SKUs in a product category, we frequently must resort to a trial-and-error approach, which is often counterintuitive and time consuming.

Traps

But there can be traps in the use of search terms, so the S&OP manager must exercise caution. An example comes from a recent article by Lazer and colleagues (2014) examining predictions from Google Flu Trends. Their abstract:

> *In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu-tracking system would have hoped.* Nature *reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports.* **http://www.sciencemag.org/content/343/6176/1203**

FSS Support

Automating the process requires integration of trend data into the typical transactional database. While this is not difficult, it may change the typical work-flow in the forecasting process, and might not be supported at all by the forecasting support system used by the company. If Microsoft Excel is used, for example, transaction and trend data need to be collated on a spreadsheet before models can be run and predictions made.

So there are limitations on how comprehensively the approach can be applied. It makes most sense to identify those SKUs and product categories that can benefit most by the inclusion of trend data. Refinements to the trend model can then be made by removing or adding search terms to stay ahead of current trends. These could be SKUs with high forecast errors, new SKUs, or SKUs that are promoted. The overall objective, of course, is to improve forecast accuracy, consequently achieving better business performance.

REFERENCES

Choi, H. & Varian, H. (2012). Predicting the Present with Google Trends, URL: **http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf**.

Choi, H. & Varian, H. (2009). Predicting the Present with Google Trends, Technical Report, Google. URL: **http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf**.

Ganeshan, R. (2014). Clickstream Analysis for Forecasting Online Behavior, *Foresight*, Issue 33 (Spring 2014), 15-19.

Ginsberg, J.M., Mohebbi, H., Patel, R.S., Brammer, L., Smolinski, M.S. & Brilliant, L. (2009). Detecting Influenza Epidemics Using Search Engine Query Data, *Nature*, 457, 1012-1014.

Lazer, D.M., Kennedy, R., King, G. & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis, *Science*, 343, 6176, 1203-1205.

Schmidt, T. & Vosen, S. (2013). Forecasting Consumer Purchases Using Google Trends, *Foresight*, Issue 30 (Summer 2013), 38-41.

Tonya Boone is Associate Professor of Operations and Information Technology at the Mason School of Business, the College of William and Mary. Currently, she is engaged in projects involving sustainable product design, service supply chain strategies, inter-organizational knowledge transfer, and diffusion of environmental innovations.

Tonya.Boone@mason.wm.edu

Ram Ganeshan is D. Hillsdon Ryan Professor of Business, Mason School of Business, College of William and Mary. In 2001, the Production & Operations Management Society (POMS) awarded him the prestigious Wickham Skinner Award for his research on how supply chains can be efficiently managed. He is currently exploring how big data can improve supply chain performance.

Ram.Ganeshan@mason.wm.edu

Robert L. Hicks is Professor of Economics at the College of William and Mary, where he is an affiliate in the Environmental Science and Policy Program and the Thomas Jefferson Program in Public Policy.

rob.hicks@wm.edu