GOOGLE SEARCH ENGINE TRAFFIC IN ECONOMIC PREDICTION:
A CASE STUDY USING US CONSUMER BANKRUPTCIES

by

Mike Haymes

An essay submitted to the Department of Economics
in partial fulfillment of the requirements
for the degree of Master of Arts

Queen's University
Kingston, Ontario, Canada
July 2010

Acknowledgements

Sincerest thanks to Ian Keay for his invaluable guidance and supervision.

Table of Contents

Introduction

Increasingly, consumers now rely on the Internet to aid and inform economic decisions. Recently, new data from the Google Search Engine has presented researchers with the possibility of tapping into this massive information network to reveal the underlying behavioural characteristics and trends of Internet users.

In summer 2008, Google introduced a beta version of Google Insights for Search, a new web application allowing access to data on the relative popularity of search queries entered into the Google Search Engine. Because many search terms are indicative of specific types of user behaviour, this new data source represents an immense and timely database with which to measure and predict consumer behaviour.

The Insights web application provides time-series data describing the relative popularity of any researcher-specified search term, available from January 2004 to the present and disaggregated across national and subnational regions.[1] Using this novel data source, researchers can now peer into an immense network of search terms, many of which can be viewed as proxies to actual consumer behaviour. By identifying key search terms associated with a particular behaviour, the data can provide timely on-demand indicators of both current and future economic conditions. In addition, the new Google data presents several key advantages over traditional measures of consumer behaviour.

[1] http://www.google.com/insights/search/

The primary goal of this paper is to further demonstrate the potential of Internet search data for measurement and prediction of economic behaviour, using consumer bankruptcy prediction as a case study. Recently, a small but burgeoning field of research has demonstrated the usefulness of this new search query data in analyzing current and future economic conditions. To my knowledge, this paper represents the first attempt to predict consumer bankruptcy rates using online search traffic.

It is my belief that the current research represents only a small sample of the potential applications of this unique and vast data set. To demonstrate this, I employ a simple prediction model to evaluate whether time-series data on the relative popularity of Google searches such as "how to declare bankruptcy" can be utilized as a leading indicator to improve predictions of actual consumer bankruptcy rates in a panel of 32 US states. To evaluate the robustness of the forecast estimates, the predictive power of this Google indicator is compared against several competing consumer survey indices, which I believe may also possess some predictive ability for consumer bankruptcies. These various prediction models are evaluated based on goodness-of-fit, as well as in- and out-of-sample prediction errors.

The intuition behind the predictive power of search queries is simple. Prior to declaring bankruptcy, most individuals would first need to undertake a certain amount of research into the process. A primary source for this research is the Internet, where

Google possesses a near monopoly on search engine traffic – the company processed over 80 percent of worldwide searches in 2009, more than 10 times its nearest competitor Yahoo! Inc.[2] The popularity of online searches of common terms used to research bankruptcy – search terms such as "how to declare bankruptcy" or "bankruptcy trustee", for example – can be seen as proxies to the number of individuals engaging in this preparatory research stage of bankruptcy filing. From this proxy, it is possible to create predictions for the total number of bankruptcy filings.

These Google predictions present several distinct advantages. Firstly, they are extremely timely. It is assumed that an individual's online research phase typically precedes an official court filing by a period of several weeks to several months. This gives the Google data a significant lead on both the actual filings and the official reporting of bankruptcy numbers, which is subject to an additional publication lag. In addition to policy and planning benefits, improved bankruptcy forecasts may also possess some potential to complement existing leading economic indicators, providing a clearer picture of current consumer sentiment and future prospects for the economy as a whole. The data also has value for the purposes of "nowcasting", that is, predicting the current number of bankruptcy filings using contemporaneous Google search data, which is available in advance of the official published filings from the Administrative Office of the United States Courts.

[2] Net Applications. "Search Engine Market Share," http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qptimeframe=M&qpsp=120&qpnp=12

Secondly, Google forecasts may also present a purer measure of consumer intentions. Whereas surveys measure an individual's stated intention to undertake an action, search traffic data measures the preparatory steps towards completing that action. In many cases, the action of searching for terms such as "how to declare bankruptcy" can be viewed as a preliminary step in the broader process of filing for bankruptcy. Compared to survey measures of stated future intentions, online search behaviour may present a more trustworthy measure of true intentions. As such, Google forecasts do not suffer from recall or response bias, or other common problems in survey design.

With the ubiquity of the Internet, search data presents an immense and cost-free sample of human behaviour in the developed world. Moreover, as internet-use becomes more universal across age groups and demographics, the search data will become increasingly representative of true population characteristics. The data is updated daily and is available for a nearly limitless number of search terms. It is available aggregated to a worldwide level, or disaggregated by geography to the level of a single subnational region such as a US state or metropolitan region.

Finally, and perhaps most importantly, this new Google search data allows measurement of types of human behaviour that were previously impractical or impossible to collect using traditional survey methods. Due to their rarity, consumer bankruptcies provide a particularly good example of this. On average between 2003 and 2009, approximately 0.036% of the US population declared bankruptcy in a given month.[3,4] As such, a survey asking consumers directly about their intention to declare bankruptcy would require an average random sample of several thousand individuals to identify even a single consumer bankruptcy. To produce state-level monthly bankruptcy estimates using survey methods would require an immense sample size to produce accurate forecasting. The cost of such sampling is extremely prohibitive. By contrast, Google search data represents a feasible and potentially superior proxy to true consumer intentions. Unlike survey-based methods, even extremely rare behaviour is measurable using the Google search data, which utilizes information from billions of searches conducted each month in the United States (Wu & Brynjolfsson, 2009). The data is also readily accessible, easy to generate and free of charge.

Consumer bankruptcy forecasting represents just one application of the possible consumer behaviours observable through online search queries. There is a large potential for this data to indirectly measure a wide variety of consumer behaviours that are presently infeasible to measure directly. Largely, the method is limited only by the creativity of researchers in identify search terms that are strongly associated with a particular consumer behaviour or sentiment.

[3] US Bankruptcy Courts. "Table F2 - Business And Nonbusiness Bankruptcy Cases Commenced, By Chapter Of The Bankruptcy Code".

[4] US Census Bureau. "National and State Population Estimates - Annual Population Estimates 2000 to 2009,"_Accessed June 2009 at http://www.census.gov/popest/states/NST-ann-est.html

Google Search Data in Prediction

Recently, a small body of literature has emerged that examines the use of Google search data in various prediction applications. Notable examples appear across many academic disciplines, such as business, finance, economics, marketing, and public health.

In one of the first and most innovative applications of the Google search data in prediction, Ginsberg et al. (2009) demonstrated that Google search data could effectively predict seasonal flu trends in advance of official reports. With unfettered access to raw Google search data containing 50 million online searches, the researchers were able to identify 45 search terms that displayed strong correlation with flu activity, as measured by official reporting from the Center for Disease Control. These terms included searches such as "cold/flu remedy" and "symptoms of influenza". Using a weighted index of the search volume of each term, the authors were able to successfully predict flu activity and turning points one to two weeks in advance of the official CDC reports.

In a technical report, Choi and Varian (2009a) evaluate the ability of Google data to predict automobile sales, retail sales, housing prices, and travel patterns using basic autoregressive prediction models. Using a simple AR prediction model augmented with weekly data from the Google Trends web application, the authors are able to demonstrate reductions in the mean absolute errors of out-of-sample “nowcasting” predictions available in advance of official figures. The magnitude of these reductions varies by application. For example, they are able to demonstrate a three percent reduction in out-of-sample nowcasting prediction errors for automotive sales, and a 15-18 percent reduction for a model predicting current retail sales of automotive parts. Rather than using specific search terms, the authors use data measuring the relative popularity of search “categories”. Search terms are automatically grouped into these categories using a Google computer algorithm. A drawback of this method is that the exact search terms in the categories are unknown to the researchers.

In another economic application, Della Penna & Huang (2009) utilize Google search data to measure consumer sentiment. The authors construct an index of consumer sentiment using Google search category data measuring searches relating to bankruptcy, office furniture, luxury goods, and energy costs. These searches are deemed to be indicative of consumer sentiment. The one-month-lagged value of the Google index is added to an autoregressive framework to predict the value of the Conference Board’s Consumer Confidence Index, as well as a separate AR model to predict the University of Michigan’s Index of Consumer Sentiment. The Google index is observed to be statistically significant in both AR models. However, the opposite is found not to be true – that is, the lagged values of the two consumer sentiment indices are not statistically significant when added to an AR model explaining the Google index. Thus the authors conclude that the Google index has some explanatory power for predicting consumer sentiment indices.

Additional recent literature has demonstrated the efficacy of Google search data in a wide variety of economic prediction exercises, including the prediction of unemployment rates (Askitas & Zimmermann, 2009, and Choi & Varian, 2009b), private consumption (Konstantin et al., 2009, and Schmidt & Vosen, 2009), housing prices and sales (Wu & Bynojolfsson, 2009), and home foreclosures (Webb, 2009).

Data Quality

Google Insights for Search is one of two interfaces for accessing search data from the Google Search Engine. The other, Google Trends, provides similar data but with slightly less flexibility in how the data is disaggregated. Both interfaces offer data on the relative popularity of user-specified search terms from January 2004 to the present. The data is available as daily, weekly, monthly series.[5] Google Insights for Search offers additional functionality over Google Trends, such as the ability to compare the relative popularity of up to five search terms in one region, or to compare the relative popularity of one search term in up to five regions.

[5] As of June 2010, daily data is only available for series that span a 3-month period or shorter.

For empirical applications, Google search data faces several key limitations. Transformations and restrictions are applied to the raw search data, likely to preserve the anonymity of Internet users. Google is not forthcoming with specific details of the algorithms applied to the raw data.[6]

Google's data transformation procedure involves two manipulations of the raw data. First, the raw search data is normalized to produce a measure of relative popularity of a particular search term. As Google describes on the Insights website, this normalization process divides the raw search data by a "common variable" to produce a relative measure of search popularity, so that search popularity can be compared across regions with differing levels of overall search traffic. Though Google is not explicit about this common variable, it is assumed that this measure represents either a measure of total Internet search activity for a particular region or, if sampling is used to construct the data, then this common variable represents the size of the sample of total searches examined in that region. In either case, the procedure produces a measure of a particular search term's share of overall search traffic in a particular region.

The second manipulation involves scaling the data to form an index of values between zero and 100. This index is created by scaling the data by the largest

[6] Privacy concerns for internet searches are not unfounded. In 2006, a class-action lawsuit was filed against AOL Inc. following the release of data on 20 million keyword searches performed over a three month period. Using the data, a *New York Times* reporter was able to uncover the identity of an individual AOL user from their search record.

observation value in any particular data series. Finally, the data availability is restricted so that users may only access data for terms that receive a minimum amount of search traffic. Again, this measure is likely designed to protect the anonymity of Internet users. The details of this minimum level of search traffic are not publicly available.

For the empiricist, these manipulations and restrictions severely degrade the information contained in the index, imposing several challenges. The data is first normalized then scaled, both times employing division by unknown denominators. For a particular search term, the normalization process destroys any information about the true number of searches performed. By contrast, the scaling process destroys any information about the relative share of searches performed compared to indices for other search terms. What remains is an index of relative popularity of a search term over time, measured on an unknown scale. As such, coefficient estimates from regression analysis will possess scant real-world interpretability. They can be interpreted in the context of the index itself, not in terms of actual online consumer behaviour.

The transformed index retains its ordinal ranking over time, allowing researchers to compare the share of total search traffic of a particular term across different points in time. However, due to the scaling procedure, two distinct series are not directly comparable, as the observations will have been scaled separately by the highest peak values in each series. During the normalization process, the data is also divided by a measure of total traffic to convert the series to a measure of a search term's share of total traffic. The frequency with which Google updates this denominator is not disclosed and could present a source of noise in the data. This is particularly problematic if the denominator is highly variable and is updated frequently, in which case the Google data would largely be driven by variation in overall search traffic rather than variation in the specified search terms. Alternatively, the denominator may be constant over time if the index is created from a fixed-size sample of total searches performed in a given time period.

Google's scaling procedure also presents concerns about loss of information due to rounding of observations. The scaling procedure divides by a highest peak value then rounds to fit observations to an integer scale of zero to 100. This rounding presents data quality concerns, particularly if the data's highest peak is a large outlier. In this case, the scaling will produce significant "flattening" of the series, rounding away much of the variability. The scaling also represents some problems for replicability of empirical results. These problems arise because data series examined over different time frames may possess different "highest peak" observations, and thus will not necessarily share the same scale. Combined with other data quality issues, there is the potential that small deviations in the data series timeframe may lead to significant differences in coefficient estimates or overall results.

Finally, the evolving role of the Internet in society also presents several challenges when dealing with the data. Total internet usage continues to rise, but there is no reason to expect that queries relating to specific economic behaviour such as bankruptcy will rise proportionally with overall Internet traffic. For example, if total Internet searches have risen proportionally faster than searches relating to bankruptcy, we would expect to see a downward pressure on the data series over time. The nature of searches themselves may also be changing over time. As aggregate computer literacy improves, we may expect to see more "advanced" Google search behaviour, including the use of Boolean operators and more exacting search strings. In addition, as the Internet and search technologies evolve information becomes easier and quicker to access, the cost associated with online researching falls. This declining cost may induce more casual researching of bankruptcy information, altering the characteristics of bankruptcy searchers. The share of bankruptcy searches that could be perceived as "preparatory" behaviour for bankruptcy would fall, and the relationship between bankruptcy searches and the bankruptcy rate would evolve dynamically over time.

Despite these limitations, existing research has demonstrated that the Google search data index still possesses strong potential for measuring and forecasting a wide range of consumer behaviours, many of which are infeasible to measure using alternate means. As of June 2010, the Google Insights for Search application still exists as a preliminary beta release. Ideally, future iterations of the application will address some of these data quality issues while still protecting users' rights to online privacy. It is a reasonable assumption that improvements to the data quality will lead to improved performance of search traffic data in forecasting and nowcasting applications.

Methods

The methodology of this paper is twofold. First, three competing state-level Google data indices are extracted from the Google Insights for Search web application. Each index measures the monthly relative popularity of a group of search terms believed to be indicative of preparatory behaviour towards declaring bankruptcy, such as the search phrase "how to declare bankruptcy". The relationship between each index and the bankruptcy rate is examined using intertemporal correlations and estimation methods.

Next, the index deemed the strongest candidate for prediction is used to create in-series and out-of-series bankruptcy rate predictions. To evaluate the robustness of these predictions, they are compared against predictions produced using the same methodology, only substituting survey-based consumer sentiment indices in place of the Google data. These survey-based indices are believed to also possess some predictive ability for consumer bankruptcies. The efficacy of the prediction models incorporating these competing indices is used as a baseline against which the Google prediction models are evaluated.

I. Identifying Predictive Search Terms

Each Google index is comprised of several search term strings suspected of possessing forecasting ability for consumer bankruptcies. The Boolean operator "OR" is used to extract a Google index representing the aggregate share of search traffic for multiple search strings. Many of the search strings used in the indexes represent various ways of searching for online information on how to declare bankruptcy. All three indices contain search terms that are suspected of evidencing preparatory behaviour for declaring bankruptcy, such as "how to declare bankruptcy" or "file bankruptcy". The indices differ in how broadly the search terms are defined. Index 1, for example, contains the most general search terms of the three indices, and includes the term "bankruptcy" itself. Index 2 contains the fairly general terms "chapter 7" and "chapter 13". Index 3 includes only search terms deemed specific to preparatory behaviour for declaring bankruptcy. This index includes only terms such as "how to declare bankruptcy" and "filing for bankruptcy", as well as terms such as "bankruptcy lawyer" and "bankruptcy trustee". While there are many reasons why an Internet user might search for the term "bankruptcy", searches for terms such as "how to declare bankruptcy" are assumed to be fairly indicative of a user's higher likelihood of declaring bankruptcy in the upcoming weeks or months. A complete list of search terms used to construct the indices is presented in Appendix 1.

The rationale for choosing three indices is twofold. First, the three indices allow us to test the *a priori* assumption that Index 3, the index most specific to preparatory behaviour for a bankruptcy filing, should possess the highest predictive power and the strongest leading correlations with actual bankruptcy filings. The second rationale is more pragmatic. To protect user privacy, Google imposes a minimum search volume requirement on any data extractions. Prime candidates for individual predictive search terms such as "how to declare bankruptcy" do not generate enough searches to meet this minimum data extraction requirement on their own. Rather, they must be aggregated with other, similar search terms in order to generate a sufficient amount of data for testing purposes. The indices are evaluated for current and leading correlation with the actual bankruptcy rate. In addition, a simple fixed-effects panel regression in which the Google index is used as the sole regressor to predict the state-level bankruptcy rate is examined. This fixed-effect regression is of the form

$$Y_{i,t} = \beta_0 + \beta_1 G_{i,t-k} + v_i + \varepsilon_{i,t} \qquad \textbf{Model (1.1)}$$

where: $Y_{i,t}$ is the consumer bankruptcy rate in state *i* in month *t*; $G_{i,t\text{-k}}$ is the state-specific value of the Google index, lagged by a k-month period; and v_i is the state-specific fixed effect. The model is estimated multiple times for each index, using k values from zero to six. The purpose of this simple panel regression is not to produce forecasts of bankruptcies, but rather to better identify the intertemporal pattern of association between the Google search data and the bankruptcy rate, and to interpret how this relates to an individual consumer's online search behaviour.

The strongest of the three indices is preserved for testing in the formal forecasting exercise described in the proceeding section.

II. Evaluating Consumer Bankruptcies Using Google Search Data

Data on monthly relative popularity of the search index was extracted for each of the indices across a panel of US states for the period January 2004 to March 2010. Despite the aggregation of search terms, 17 states still did not meet Google's minimum search volume privacy requirements for one or more of the indices. Thus, a complete data series for the timeframe could not be extracted and these states were excluded from the panel. Two states were also deemed as outliers in the Google data and were dropped from the sample, resulting in a final panel of 32 states. [7]

In October 2005, the US Bankruptcy Abuse Prevention and Consumer Protection Act came into effect, introducing a new means test that individuals must pass in order to declare chapter 7 bankruptcy. The effects of this policy change are evident in both

[7] Virginia and Iowa were deemed outliers due to abnormally weak positive correlations between the Google indices and the state bankruptcy rate for the post-policy subsample. The remaining 32 states in the panel displayed medium to strong positive current correlations between the two variables for this period, with a mean current correlation coefficient of 0.74 for index 3 (standard deviation of 0.16). The current correlations for Virginia and Iowa were 0.15 and 0.14 respectively, more than three standard deviations from the average of the correlation coefficients of the remaining 32 state panel. Why these two states display anomalous search data patterns relative to the other 32 states is unknown. Retaining these two states in the panel does not significantly alter the results presented in this paper.

the Google and actual consumer bankruptcy data. The relative popularity of bankruptcy searches and the actual bankruptcy filing rate increase significantly prior to the policy change, as individuals rush to file bankruptcy under the more lenient regime. Additionally, the bankruptcy rate and Google index plummet following the policy change, presumably due to the large number of individuals who expedited their filings to avoid the new regulations.

To accommodate the effects of this policy change in the data, the sample is split into pre- and post-policy subsamples, which are examined separately. This is done under the rationale that, by restricting the eligibility of individuals to declare bankruptcy, the policy change should fundamentally alter the relationship between the popularity of bankruptcy related searches and the number of individuals who file a petition for bankruptcy. This assumption of a structural break in the relationship was formally tested using a Chow test procedure. The hypothesis that the model parameters are unchanged by the policy is strongly rejected, supporting the decision to examine the subsamples separately. The data is split into a pre-policy subsample spanning January 2004 to August 2005 and a post-policy subsample spanning January 2006 to March 2010.[8]

[8] The policy change also led to large outliers in the bankruptcy data for several months preceding and following October 2005. To adjust for these outliers, a Cook's distance test is performed and outlying observations are dropped from the sample.

To test the value of the Google search data in predicting consumer bankruptcies, I employ a method similar to the one utilized by Choi and Varian (2009a) to estimate current consumer retail and automotive sales in advance of published sales data. The method involves first constructing a simple baseline fixed-effects panel forecasting model for state-level consumer bankruptcy rates. This baseline specification consists of a 32-state seasonal AR panel model of the form

$$lnY_{i,t} = \beta_0 + \beta_1 lnY_{i,t-1} + \beta_2 lnY_{i,t-12} + v_i + \varepsilon_{i,t} \qquad \textbf{Model (2.1)}$$

where $Y_{i,t}$ represents the bankruptcy rate at time *t* in state *i*, v_i represents the fixed state-specific effect for state *i*, and $\square_{i,t}$ is a random error. The state-level bankruptcy rate is calculated as the number of total consumer bankruptcy filings per 100,000 in state *i*. A simple straight-line method is used to generate monthly state-level population estimates from the US Census Bureau's annual state population estimates.[9]

Two additional Google models are then constructed by augmenting the original baseline model with Google search data. The first augmented model is a nowcasting model that uses the current month as well as previous months of Google search data to predict the current bankruptcy rate. The potential value of such a model lies in its ability to predict the current bankruptcy rate in advance of the official publication of figures from US bankruptcy courts, which are subject to a publication lag. The second

[9] US Census Bureau. National and State Population Estimates - Annual Population Estimates 2000 to 2009, Accessed June 2009 at http://www.census.gov/popest/states/NST-ann-est.html

augmented model is a true forecasting model, which uses Google search data from previous months to predict the current state bankruptcy rate. The augmented models take the form:

Nowcasting Model:

$$lnY_{i,t} = \beta_0 + \beta_1 lnY_{i,t-1} + \beta_2 lnY_{i,t-12} + \sum_{k=0}^{2} \gamma_k lnG_{i,t-k} + v_i + \varepsilon_{i,t} \qquad \textbf{Model (2.2)}$$

Forecasting Model:

$$lnY_{i,t} = \beta_0 + \beta_1 lnY_{i,t-1} + \beta_2 lnY_{i,t-12} + \sum_{k=1}^{2} \gamma_k lnG_{i,t-k} + v_i + \varepsilon_{i,t} \qquad \textbf{Model (2.3)}$$

Where $G_{i,t-k}$ represents the value of the Google search index or a competing predictor index in month *t-k.* The nowcasting model relies on data from the current month Google search index, in addition to the previous two months. The forecasting model uses information from the previous two months to predict the bankruptcy rate in month *t*. [10]

A fixed-effect model was chosen largely to account for the "highest peak" scaling that is applied by Google to the data. The result of this scaling is that the Google values are measured across different scales for different states. For example, an observation value of 70 in Arizona is fundamentally different from an observation of 70 in

[10] The choice of lags for the nowcasting and forecasting models were determined by examining a fixed-effects estimation of the bankruptcy rate using up to five-month-lagged values of the Google data as the sole regressors. Statistical significance of the coefficients was used to determine the choice of lags to apply to the Google data in Models (2.2) and (2.3)

California. I believe that using fixed-effects within-estimates may improve the comparability across states by utilizing log-differences from the series' means, since percentage deviation from mean should not be as affected by the scaling procedure. The model does not completely remedy the scaling-comparability issue, however, and this should be acknowledged as a potential confounder of the results. The poor comparability of the state-level Google data is also the primary rationale for using population-adjusted bankruptcy rates, as opposed to actual bankruptcy figures. A Levin-Lin-Chu unit root test is undertaken to test for stationarity of the panel bankruptcy rates for the post-policy subsample using lags suggested by the Akaike and Bayesian Information Criteria.[11] The null hypothesis that the panel series are non-stationary is strongly rejected using the BIC suggested lags (p-value ≈ 0); however, the null cannot be rejected using the more rigorous lag requirements suggested by the AIC (p-value= 0.1483).[12]

To evaluate the additional predictive value added by the Google search data, the augmented models are evaluated against the simple baseline forecasting model on the basis of goodness-of-fit, as well as a comparison on the basis of in- and out-of-sample

[11] The pre-policy subsample poorly suited for the Levin-Lin-Chu test due to the unbalanced nature of the panel and the relatively small number of series observations, particularly given the large number of lags required for the test. Instead, an additional Levin-Lin-Chu test was performed for the entire sample. The null that the series are non-stationary is rejected, regardless of whether the AIC- or BIC-specified lags are used.

[12] Because the BIC specification is considered more representative of the "true model" number of lag parameters, the assumption of stationarity of the bankruptcy rates is upheld throughout this paper (Burnham, 2004).

mean absolute prediction errors. A difference-of-means test is conducted to test the significance of the MAE reductions of the index models relative to the baseline case. The out-of-sample predictions are constructed for the final nine months of 2009, using the post-policy subsample data. For any given month, *t*, the forecasting predictions are constructed using information available in month *t-1*. By contrast, the nowcasting predictions utilize information available at the end of month *t* to predict the current bankruptcy rate in month *t*, in advance of the official published numbers.

Finally, additional comparison models are created by replacing the Google search data index in models (2.2) and (2.3) with alternative survey-based indices representing consumer sentiment and expectations. These indices are assumed to also possess some predictive power in consumer bankruptcy rates. The comparison models are evaluated against both the Google-augmented models and the baseline model. Like the Google-augmented models, these comparisons are evaluated on the basis of goodness-of-fit, and in- and out-of-sample prediction error.

Five survey-based indices are used for these comparisons, each using monthly data from the University of Michigan/Reuters Consumer Surveys. These indices include: the Index of Consumer Sentiment, an index measuring consumers' financial well-being based on the past year and future expectations; the Index of Current Economic Conditions, an index measuring a consumer's financial progress from the previous year and their current sentiment towards major purchases; and the Index of Consumer Expectations, an index measuring consumers' expectations about their future financial situation over the coming 12 months. Two additional indexes from the Consumer Surveys are also modelled, both derived from consumer responses to a particular survey question: "Looking ahead--do you think that a year from now you (and your family living there) will be better off financially, worse off, or just about the same as now?" The first index, labelled "Poor Expectations Index 1", measures the percentage of consumers who responded that they expect to be worse off financially in the coming year. The second index, labelled "Poor Expectations Index 2", is an index that measures the gap between the percentage of people who responded that they expect to be better off financially and the percentage who responded that they expect to be worse off.

The comparison indices serve to provide a second baseline against which to evaluate the robustness of the Google search data, and to avoid the possibility of creating a "straw man" comparison when evaluating the more heavily specified Google-augmented models against the less-specified baseline model. One potential drawback to this comparison is that, unlike the Google search data, the monthly survey-based indices are not available disaggregated by state. Rather, the data is only available disaggregated across five US geographical regions and has been mapped to the state level in the data panel. Though this geographic aggregation may result in some loss of predictive power at a state level, it also represents a “best approximation” to monthly state-level data available through these survey-based indices. As a result, any effect on the prediction efficacy of the two measures arises from a “real-world” advantage of the Google data – namely, its superior availability across various levels of geographic and temporal disaggregation.

As an extension, a single data series containing data for the entire United States from January 2004 to March 2010 is examined, utilizing the same baseline and Google-augmented nowcasting and forecasting model specifications discussed previously, though adapted for a non-panel model. Because the national data subsamples are small relative to the panel, the national data extension is presented primarily as an additional robustness test of the general method presented in the panel model. By comparing models using national-level Google data against models containing national-level consumer confidence index data, the extension also overcomes the geographic discrepancy discussed above for the panel model.

For this extension, the Google search data is constructed by extracting monthly data for Google Index 3 at a national level. In this case, the Google-augmented models are compared against three prominent survey-based indices from the US Conference Board: the Consumer Confidence Index, the Present Situation Index, and the Expectations Index. The purpose of this national-level extension is to further evaluate the robustness of the Google search data in predicting consumer bankruptcies, utilizing survey-based indicators for comparison purposes.

The methodology presented in this paper builds primarily on the work of Choi and Varian (2009a) and Della Penna (2009). However, several differences exist in the methodologies employed, in particular as it relates to the construction of the search data indices. Firstly, this paper utilizes an index constructed from a specific set of Google search terms believed to possess predictive power for consumer bankruptcies. Previous papers have largely relied on data relating to Google's search categories, which are automatically populated with search terms by Google. Only a small list of the top 10 most popular search terms in each category are known to the researcher. While many of these search terms may possess predictive power, it is likely that many others will not.[13] Secondly, unlike the Choi and Varian paper, this paper introduces competing prediction indices, against which the robustness of the predictive power of the Google data is evaluated.

[13] For example, in the Google category for "Bankruptcy" the search term "Babcock" appears in the top 10 category searches, referring to the construction and engineering firm Babcock & Wilcox which has undergone a highly publicized bankruptcy filing. Also in the top 10 searches is the German term for insolvency, "insolvenz". Neither of these terms should be expected to possess any predictive power for consumer bankruptcies in the US, and are likely a source of significant noise in the prediction models. For this reason, I chose to use a collection of researcher-specified search terms instead of the categorical data utilized in much of the earlier research.

Results

I. Google Search Data Indices and the Consumer Bankruptcy Rate

At the state level, each of the three Google indices possesses a positive average contemporaneous correlation with the state bankruptcy rate, for both the pre- and post-policy subsamples. These correlations are significantly stronger for the larger post-policy subsample. The correlations for the post-policy subsample are presented in Table 1.

TABLE 1. Mean state-level correlation, Google search index and state bankruptcy rate, 32 state panel, Jan 2006-Mar 2010

	Least Specific Search Terms ---------------------------Most Specific Search Terms		
Lag on Google data	Index 1	Index 2	Index 3
Current	0.70	0.62	0.74
	(0.14)	(0.28)	(0.16)
1 Month Lag	0.68	0.60	0.73
2 Months Lag	0.64	0.59	0.72
3 Months Lag	0.52	0.56	0.61
4 Months Lag	0.45	0.57	0.56

Note: standard error in parentheses. Mean state-level correlation is the average of the correlations produced by the state-specific Google index value and the state-specific bankruptcy rate.

Of note, the correlations are strongest for Index 3, the index measuring the relative popularity of search terms believed to be highly specific to declaring bankruptcy. This index measures the relative popularity of search terms such as "how to declare bankruptcy" and "how to file bankruptcy". By contrast, Index 1 measures the relative popularity for these specific terms aggregated with more general terms such as "bankruptcy", "chapter 7" and "chapter 13". Index 2 is similar to Index 3 but notably lacks the very general search term "bankruptcy". Index 2 is somewhat anomalous compared to the other two indices. While the index still displays strong to moderate positive correlations with the state-level bankruptcy rate, these correlations are slightly lower on average and more variable than the other two indices.

Also noteworthy is the pattern of declining positive correlation as greater lags are applied to the Google indices. This result may be indicative of the variability of lead time in consumers' preparatory phase of online research prior to declaring bankruptcy. If individuals typically research bankruptcy one to several months in advance of an official declaration, this could produce the pattern of declining correlation observed.

A similar result is obtained by examining the goodness-of-fit of the fixed-effects panel regression model described in Model (1.1) for the post-policy period. This model uses a single Google variable as the sole regressor to explain the variation in the state bankruptcy rate. Estimating this model multiple times – substituting current to six-month lag values of Google index data – produces a pattern similar to that seen in the correlations. In each estimation, the Google index variable is highly significant. Index 3 again outperforms the other two indices in terms of goodness-of-fit (as measured by the Within R-squared value), producing R-squared values ranging from 0.23 for the six-month lag specification of (1.1), to 0.49 for the one-month lag specification.

Similar to the correlation results, the same pattern of declining goodness-of-fit is observed, as the model is re-estimated using greater lags on the Google data. The one exception to this is for Index 3, for which the one-month-lagged value of Google Index 3 produces a very slightly higher R^2 than the model using the current Google Index 3 value. The R-squared values from these regressions are presented in Appendix 1.2.

Of the three Google indices examined, both the correlation and estimation results provide support for Index 3 as the strongest predictor of the current and future bankruptcy rate. This result is in line with the *a priori* assumption that the index measuring the relative popularity of search terms believed to be most specific to preparatory bankruptcy research should also possess the strongest potential for accurately predicting the bankruptcy rate.

II. Google Search Data as a Predictor of the Consumer Bankruptcy Rate

Detailed results of the nowcasting and forecasting panel regressions for the pre-policy and post-policy subsamples are presented in Appendices 2.1 to 2.4. These appendices display the estimation results from models (2.2) and (2.3) using Google Index 3 search data. In addition, the appendices present the estimation results of the baseline models, and estimates of models (2.2) and (2.3) computed using survey-based indices as predictors in place of the Google data. The results of these 28 panel regressions are summarized in Tables 2 and 3. Table 2 presents the R-squared values as a measure of goodness-of-fit; Table 3 presents the coefficient estimates for the Google and competing survey-based variables.

TABLE 2. Goodness-of-fit comparisons, nowcasting and forecasting models, pre-policy and post-policy subsamples

	Pre-Policy		Post-Policy	
	Nowcasting	Forecasting	Nowcasting	Forecasting
Baseline Model	0.343	0.343	0.861	0.861
Google	0.469	0.395	0.878	0.871
Index of Consumer Sentiment	0.596	0.422	0.871	0.868
Index of Consumer Expectations	0.620	0.414	0.869	0.865
Index of Current Economic Conditions	0.557	0.526	0.872	0.869
Poor Expectations Index 1	0.440	0.367	0.866	0.863
Poor Expectations Index 2	0.475	0.413	0.867	0.864
N	497	497	1513	1545

Goodness-of-fit measured by within R-squared value of the panel regression. Pre-policy subsample covers Jan. 2004-Sept 2005. Post-policy subsample is Jan. 2006 – Mar. 2010. Google data uses Google Index 3.

Generally, the estimated coefficients for the Google search data variables are moderate to highly significant for all models estimated. For the pre-policy nowcasting subsample, the estimated coefficients for the current and one-month-lagged Google search data variables are significant at the one percent level, as is the one-month-lagged

Google coefficient in the forecasting model. The two-month-lagged Google search data is not significant in either model. For the post-policy nowcasting model, the current and two-month-lagged Google coefficients are significant; in the forecasting model, both the previous month and two-month-lagged Google search data coefficients are highly significant.

TABLE 3. Coefficient estimates of Google and survey-based indices, nowcasting and forecasting models, post-policy subsample, 32-state panel

	Current	1 Month Lag	2 Month Lag
Nowcasting Model **Model (2.2)**			
Google	0.116**	0.032	0.040*
Index of Consumer Sentiment	-0.145**	0.029	-0.146**
Index of Consumer Expectations	-0.145**	0.116**	-0.172**
Index of Current Economic Conditions	-0.102**	-0.121**	-0.041
Poor Expectations Index 1	0.039**	-0.016	0.043**
Poor Expectations Index 2	-0.114*	0.007	-0.235**
Forecasting Model **Model (2.3)**			
Google	-	0.077**	0.072**
Index of Consumer Sentiment	-	-0.085*	-0.157**
Index of Consumer Expectations	-	0.010	-0.188**
Index of Current Economic Conditions	-	-0.189**	-0.051
Poor Expectations Index 1	-	-0.003	0.047**
Poor Expectations Index 2	-	-0.049	-0.249**

** significant at the 5% level. ** significant at the 1% level. Nowcasting Models are estimated using model (2.2) as described in the methods section. Forecasting models are estimated using model (2.3). For survey-based indices, the index is used in place of the Google index. Post-policy subsample is Jan 2006-Mar 2010.*

Competing survey-based indices are also highly significant in the nowcasting and forecasting regression specifications. Such high significance of all indices studied may represent that all indices are strong predictors of the current and future bankruptcy rate. However, it may also arise due to the minimalistic specification of the baseline forecasting model.

The goodness-of-fit of all models including the baseline model is markedly worse in pre-policy subsample period than in the post-policy period. The reason for this discrepancy is unknown, though it may reflect some influence of the US policy change enacted in October 2005, which limited the ability of consumers to file for chapter 7 bankruptcy. The policy change received significant media coverage as early as April 2005 when the bill was formally signed into law. The effect of the policy change in late 2005 can be clearly seen in the Google and bankruptcy rate data; however, the advance notice of the policy change may also have influenced several months of the Google and actual bankruptcy data at the end of the pre-policy period. These months were not identified in the Cook's distance test for outliers. Because the pre-policy subsample period consists of only 21 months of data, the policy announcement may have influenced a significant proportion of the entire subsample. This is one possible explanation for the poor fit of the pre-policy models.

In the pre-policy subsample models, the Google data performs poorly relative to the survey-based indices. Both the Google forecasting and nowcasting models produce R-squared values only slightly above that of the baseline model. All competing survey-based indices produce higher R-squared values than the Google models, with the exception of the Poor Expectations Index measuring the percentage of individuals who responded that they expected their family financial situation to decline over the next year.

An opposite pattern is observed in the larger post-policy subsample. In this subsample, the baseline seasonal AR forecasting model produces a substantially higher R-squared value of 0.861. Moreover, the addition of the various indices in the nowcasting and forecasting model only produces slight improvements to the baseline goodness-of-fit. This differs from the pre-policy subsample, where large improvements in the R-squared values were observed. In the post-policy subsample, the Google nowcasting and forecasting models outperform all other survey-based indices in terms of goodness-of-fit, though the difference is modest.

Table 3 summarizes the coefficient estimates for the nowcasting and forecasting specifications (models (2.2) and (2.3), respectively) for the post-policy period, using the Google and competing survey-based indices. The nowcasting model specification includes the current month value of the index, as well as values for the preceding two months as regressors. The forecasting model includes only index values from the preceding two months to predict the current bankruptcy rate.

Of note, the estimated coefficients for the current value of the indices are all significant and display the expected sign. The expected positive coefficient is observed for the Google search index and Poor Expectations Index 1, which measure preparatory online bankruptcy research behaviour and the percentage of individuals expecting their financial situation to degrade in the future, respectively. Both would be expected to be positively associated with the bankruptcy rate. By contrast, the remaining indices measure positive consumer sentiment and expectations, and display the expected negative association with the current state bankruptcy rate. This pattern of expected coefficient signs is also observed in the pre-policy nowcasting model estimates.

Both the one-month and two-month-lagged Google index values are highly significant in the forecasting model, and both display the expected positive sign on their coefficient estimate. The Google data is the only index to display highly significant coefficients for both the one-month and two-month-lagged observations. With few exceptions, the survey-based indices also display the expected signs for the forecasting models.

The predictive power of the Google model was evaluated against the baseline and competing indices using mean absolute error (MAE) comparisons. The errors were calculated for an in-sample estimation period as well as a nine month out-of sample prediction period spanning April 2009 to December 2009. For the out-of-sample predictions, the monthly bankruptcy rate predictions were generated using information available in the preceding month (forecasting model), or the current month (nowcasting model). In total, 576 out-of-sample state bankruptcy rate predictions were generated for the 32 states over the nine-month prediction period. A transformation was applied to the errors from the forecasting and nowcasting log-log models to obtain model errors expressed as deviation from the true bankruptcy rate observed in month *t*. The mean of these errors is presented as the mean absolute error in Table 4. These absolute errors were also calculated as a percentage of the true bankruptcy rate for each of the in-sample and out-of-sample predictions. The mean of these percentage errors is also presented in Table 4 as the mean absolute percentage error. For each index model, a difference-of-means test was conducted to test the significance of the MAE reductions relative to the baseline model.

As with previous results, relatively poor prediction results are observed during the pre-policy period. In particular, the Google model performs more poorly than five of the six competing indices during the pre-policy period. The Google model only produces marginally lower prediction errors in this period. Compared to the mean baseline model absolute error of 7.027 bankruptcies per 100,000 individuals, the addition of the Google search data produces a 12.5 percent reduction in MAE, reducing it to 6.147 per 100,000. For this pre-policy subsample, this improvement is modest relative to the competing indices, which produce MAE improvements in the range of 8.9 to 27.2 percent as compared to the baseline model.

TABLE 4. Mean absolute and mean absolute percentage errors of bankruptcy rate predictions, 32-state panel

	In-Sample (Pre-Policy)		In-Sample (Post-Policy)	
	Nowcasting	Forecasting	Nowcasting	Forecasting
Baseline Model	7.027 (11.95%)	7.027 (11.95%)	2.976 (10.64%)	2.976 (10.64%)
Google	6.147* (10.88%)	6.651 (11.51%)	2.752** (9.90%)	2.88 (10.26%)
Index of Consumer Sentiment	5.341*** (9.81%)	6.51 (11.30%)	2.861 (10.28%)	2.919 (10.39%)
Index of Consumer Expectations	5.114*** (9.4%)	6.615 (11.44%)	2.854 (10.26%)	2.918 (10.40%)
Index of Current Economic Conditions	5.671** (10.07%)	5.769** (10.08%)	2.858 (10.23%)	2.907 (10.31%)
Poor Expectations Index 1	6.401 (11.34%)	6.864 (11.86%)	2.885 (10.38%)	2.946 (10.48%)
Poor Expectations Index 2	6.240 (11.02%)	6.599 (11.51%)	2.918 (10.42%)	2.956 (10.49%)
N	497	497	1513	1545

			Out-of Sample (Post-Policy)	
			Nowcasting	Forecasting
Baseline Model	-	-	3.325 (7.99%)	3.325 (7.99%)
Google	-	-	2.982*	3.026

			(7.20%)	(7.29%)
Index of Consumer Sentiment			4.012	3.771
	-	-	(9.34%)	(8.75%)
Index of Consumer Expectations			4.074	3.838
	-	-	(9.57%)	(8.96%)
Index of Current Economic Conditions			3.761	3.511
	-	-	(8.68%)	(8.09%)
Poor Expectations Index 1			3.504	3.411
	-	-	(8.35%)	(8.11%)
Poor Expectations Index 2			3.875	3.741
	-	-	(9.14%)	(8.80%)

*Difference of Mean Test Results (unpaired t-test): *MAE significantly smaller than the baseline model (10% level); ** MAE significantly smaller than baseline model (5% level); *** MAE significantly smaller than baseline model (1% level). Mean absolute percent errors in parentheses. Mean absolute error is measured as error in the prediction of the bankruptcy rate per 100,000. Mean absolute percent error is measured as the absolute error as a percentage of the true bankruptcy rate. The baseline model is estimated using model (2.1) as described in the methods section. Nowcasting and forecasting models are estimated using models (2.2) and (2.3), respectively. Out-of sample errors are generated individually for each month* t *using information available up to month* t-1 *(forecasting) or up to month* t *(nowcasting).*

An opposite result is again seen in the post-policy period. Here, the Google indicators outperform all other competing indices in terms of MAE reduction. Again, the reason for the discrepancy between the performance of the Google index in the pre-policy and post-policy subsamples is unknown. As previously stated, it may be a result of the advance influence of the federal bankruptcy policy change, though this should presumably affect the competing indices in a similar if not more pronounced way. It may also result from underlying changes to the volume of internet traffic, or the nature of search traffic. If increases in total Internet-use are driven by a particular type of Internet use – entertainment, for example – this may introduce a bias into the Google search data time series due to Google's data normalization process. It is important to note that the baseline model also performs poorly in the pre-policy period relative to the post-policy period. Because the baseline model contains no index data whatsoever, this supports the conclusion that bankruptcy rate volatility in the pre-policy period is the ultimate driver of these anomalous results.

In the post-policy period, the in-sample nowcasting and forecasting models produce modest improvements to the baseline prediction model for all indices. For this subsample, the Google in-sample models produce more accurate predictions than all five competing indices. In particular, the Google nowcasting model performs markedly better than the competing indices. As compared to the baseline model, the Google nowcasting and forecasting models generate MAE improvements of 7.5 percent and 3.2 percent, respectively. A difference-of-means test shows that the MAE difference between the Google nowcasting model and the baseline model is significant at the 5 percent level. This is the only model that produces significant in-sample MAE reductions in the post- policy period.

The Google model also outperforms all of the competing indices in the out-of-sample estimation results. This is an important result, as these out-of-sample estimation results represent a set of genuine predictions of state-level bankruptcy rates, using only the information available up to the time of prediction. Moreover, these out-of-sample predictions occurred during the final nine months of 2009, during which the United States was in a period of significant recession. It is during such volatile periods when accurate forecasting models are most valuable. For these out-of-sample predictions, the Google nowcasting and forecasting models produce 10.3 and 8.9 percent MAE improvements respectively, as compared to the baseline model. The difference-of-means test confirms that the Google nowcasting MAE significantly differs from the baseline model at the 10 percent confidence level. The Google nowcasting model is the only model to produce significant out-of-sample reductions as reported by the unpaired difference-of-means test.

The out-of-sample nowcasting MAE improvement of 10.3 percent is similar to those reported by Choi and Varian (2009a) in their models examining automobile sales (3 percent MAE improvement) and retail sales of automotive parts (15-18 percent MAE improvement).

Interestingly, the Google models are the only models to outperform the baseline model in the out-of-sample predictions. That is, the addition of the competing indices to the baseline model degraded the out-of-sample predictive power of the model.[14] The result may indicate that the relationship between the competing indices has changed

[14] At first glance, it may seem counterintuitive that a more heavily specified index models could perform worse than the less-specified baseline model. However, this is certainly a possibility for out-of-sample predictions. Even with in-sample estimation, a more heavily-specified model is not assured to have a lower mean absolute error; the ordinary-least-squares estimation process only guarantees an equal or lower mean-squared error.

during the prediction period in 2009. Whether this result is driven by the volatility experienced during the recessionary period or by a generally poor predictive power of the survey-based indices is unknown. Media coverage may influence consumer responses to surveys, and this may degrade their predictive ability during recessionary periods. Presumably, the Google index method would be better insulated from such effects.

On a state-by-state level, the out-of-sample Google index predictions typically outperform the baseline model and all competing indices. Compared to the baseline model, the Google out-of-sample nowcasting model produced more accurate predictions in 26 of 32 states, as measured by MAE. It outperformed the competing nowcasting model indices in 24 to 27 states of the 32-state panel, depending on the competing index chosen. Similarly, the Google forecasting model produced more accurate forecasts in 24 of 32 states, as compared to the baseline model, and outperformed the competing indices in 21 to 27 of the 32 states.

A notable trend in the out-of-sample forecasting predictions was the tendency for the baseline models and the five survey-based indices to overforecast the bankruptcy rate for the prediction period. Of the 288 forecast predictions made, these six estimation models over-predicted the bankruptcy rate between 61 and 69 percent of the time, depending on the index. By contrast, the Google forecasting model showed an equal tendency to overforecast and underforecast during the nine-month prediction period.

Prediction errors for the nine-month out-of-sample prediction period are presented for California and Massachusetts in Appendix 3.1 and 3.2. For comparison purposes, California was chosen as a representative "poor-fitting state", for which the Google predictions are poor, and Massachusetts was chosen as a representative "strong fitting" state.[15] In each figure, the Google forecasting model is compared to the baseline model and the most accurate of the five competing index models, as measured by MAE.

TABLE 5. Mean absolute errors and mean absolute percent errors of bankruptcy rate predictions, entire US, January 2006-March 2009

	In-Sample		Out-of-Sample	
	Nowcasting	Forecasting	Nowcasting	Forecasting
Baseline Model	2.337 (8.18%)	2.337 (8.18%)	3.627 (8.61%)	3.627 (8.61%)
Google	2.030 (6.90%)	2.084 (7.20%)	3.538 (8.35%)	3.425 (8.05%)
Consumer Confidence Index	2.034 (7.13%)	2.308 (7.95%)	4.19 (9.78%)	4.343 (10.33%)
Present Situation Index	2.071 (7.28%)	2.270 (7.91%)	3.475 (8.03%)	4.137 (10.08%)
Expectations Index	2.115 (7.37%)	2.366 (8.16%)	5.414 (12.83%)	4.997 (11.88%)
N	50	50		

[15] The two states were chosen based on the percentage MAE improvement when comparing the Google out-of-sample forecasting model with the baseline model. California displayed the fifth-poorest MAE improvement, and Massachusetts displayed the fifth-strongest MAE improvement of the 32-state panel. California also displays the second worst goodness-of-fit for the baseline model, as measured by MAE.

Mean percent errors in parentheses. Mean absolute error is measured as error in the prediction of the bankruptcy rate per 100,000. The baseline model is estimated using model 2.1 as described in the methods section. Nowcasting and forecasting models are estimated using models (2.2) and (2.3), respectively. Out-of sample errors are generated individually for each month t *using information available up to month* t-1 *(forecasting) or up to month* t *(nowcasting). Out-of-sample errors are calculated for a 12-month period from April 2009 to March 2010.*

Finally, the MAE results for the extension using the entire US as a single data series are presented in Table 5. These results are calculated as an additional robustness test of the Google prediction method, using a separate set of survey-based indices. In particular, this test addresses some of the ambiguity created in the panel model due to the non-homogeneous geographical disaggregation of the Google and survey-based indices. Using a single data-series for the entire US, I am able to compare models using national-level monthly data for both the Google and survey-based indices. This table compares the results of the Google models against three competing indices from the US Conference Board. These results use the same estimations models discussed previously, adapted for use with the non-panel national data series. Because the pre-policy period contains few observations, only the post-policy results are presented.

For all indices, the in-sample mean absolute errors are smaller for the national series than for the panel results. This is not surprising, as the panel model is more restrictive, imposing equal coefficients across states. Similar results are observed for the Google models in the national-level data as in the 32-state panel model. The Google index again produces moderate reductions to the in- and out-of-sample errors, as compared against the baseline model. Similarly, the Google model also performs well when compared to predictions utilizing the competing indices. For the in-sample estimation results, the Google model produces mean absolute errors of 2.030 per 100,000 in the nowcasting model, and 2.084 per 100,000 in the forecasting model. These represent 13.1 percent and 10.8 percent MAE reductions relative to the baseline model. The Google models produce larger MAE reductions than all three competing indices for the in-sample results. Once again, the out-of-sample nowcasting MAE improvement is similar to those reported by Choi and Varian (2009).

I also observe small MAE improvements for the Google out-of-sample prediction results. For these predictions, which were produced using a 12-month sample of predictions from April 2009 to March 2010, the Google nowcasting and forecasting models produce MAE reductions of 2.4 and 5.5 percent, respectively. Though the improvement to the baseline model is small, the out-of-sample Google forecast predictions are more accurate than all three competing indices and the nowcasting predictions are more accurate than two of the three. Similar to the panel results, several of the competing index predictions degrade the out-of-sample prediction results of the baseline model. Potential explanations for this result include the possibility that the survey-based indices are generally poor predictors of consumer bankruptcies, or that the effects of a recession during the prediction period has affected the prediction accuracy of survey-based indices.

Conclusions

The results of this paper support Google search data as a predictor of consumer bankruptcies. Correlations results demonstrate that correlation strength between consumer bankruptcy rates and Google index data is strongest for indices that contain search terms most specific to preparatory behaviour for declaring bankruptcy. In addition, these correlations decline as greater lags are applied to the Google data. This pattern supports the *a priori* belief that such a correlation pattern should be expected if the Google data is indeed measuring preparatory behaviour towards bankruptcy declaration. Simple panel estimation models which utililize the Google data as a single regressor to explain the bankruptcy rate also displayed the expected coefficient signs and moderate goodness-of-fit results. The goodness-of-fit of these models declined as further lags were applied to the single Google regressor, further demonstrating the expected pattern of preparatory behaviour towards bankruptcy declaration.

The results also support the Google search data index as a superior predictor of consumer bankruptcy rates relative to competing survey-based indices since January 2006. The strength of this support depends heavily on the predictive power of these competing indices. The strength of the competing indices as predictors of consumer bankruptcy is not clearly established by this paper. Generally, these competing indices were highly significant when added to a simple AR model of consumer bankruptcies, and displayed the expected sign on their coefficient estimates. However, they also performed poorly in prediction exercises. Whether these poor predictions are a result of poor predictive power or a result of the volatile conditions during the prediction period is unclear.

The strength of the Google results is largely dependent on the strength of the competing indices as predictors of the bankruptcy rate. If the survey-based indices are simply poor predictors of consumer bankruptcies, then they are also poor comparisons for the Google-based models. In this case, the results have shown that the Google-based indices produce superior forecasts to a set of indices possessing little prediction ability. However, because the Google index is superior to these competing indices, I can conclude that the Google data possesses at least *some* ability to improve a simple AR forecasting model of consumer bankruptcies. If, however, the survey-based indices perform poorly due to the increased volatility caused by a strong recession during the prediction period, then this would provide significantly stronger support for the predictive power of the Google index. In this case, the Google index would have demonstrated its ability to retain its prediction accuracy during a period of economic volatility, where competing survey-based indices have failed. This is a necessary requisite for a good predictor of economic activity, as it is during periods of volatility that leading economic indicators are most valuable.

Because the ability of these competing indices is not well established, the ability of this paper to assess the strength of the Google data as a predictor is limited. However, regardless of the prediction power of the competing indices, this paper has demonstrated that the Google index possesses at least a modest ability to improve predictions of consumer bankruptcies.

It is important to note that the Google data was able to demonstrate this modest predictive power in spite of several serious confounders. In addition to the economic volatility of the prediction period, there was a significant policy-change during the sample period which greatly affected both bankruptcy rates and the Google index. This policy change potentially influenced several months of observations preceding and following the change.

Perhaps more importantly, significant limitations on the quality of the Google data also limit its ability to accurately predict consumer behaviour. The secretive scaling and normalization procedures applied by Google make comparisons across search terms and geographic regions challenging. Minimum search volume requirements reduce data availability and limit analyses to search terms that receive sufficient amounts of traffic. Most concerning from a research perspective is the unavailability of raw data levels for search traffic. Because the Google index data measures "relative popularity" of a search term or set of terms, the data is highly influenced by the total volume of search traffic, which is unknown to the researcher. As such, it is impossible to know how much of the variation in an index of search terms is driven by variation in the search terms and how much is driven by variation in the underlying levels of total traffic. Moreover, there is very little reason to assume that bankruptcy related searches have risen proportionally with total Internet usage, so this should be seen as a potentially large source of noise in the data. As such, one should view the Google indices measuring the "relative popularity" of bankruptcy searches as only a crude proxy to the total number of individuals searching for bankruptcy information online.

Despite these limitations, the Google index data still produces modest predictive power in forecasting and nowcasting consumer bankruptcy applications, and outperforms competing survey-based indices since 2006. Ultimately, this supports the principal objective of this paper – to further demonstrate the potential of Internet search data for measuring and predicting of various types of economic behaviour, many of which are infeasible to measure using alternate means.

As of July 2010, the Google Insights for Search web platform remains in an early beta release form. Presumably, some of these limitations may improve in future iterations of the application. Ideally, access to raw search data levels for research purposes will be forthcoming. Google search engine data represents an immense and complex network of consumer behaviour. The information contained in the raw search data presents enormous potential for research and practical applications across many fields.

This paper focuses on consumer bankruptcies; however, this represents only one of many possible measurement and prediction applications for this burgeoning area of research. For economics more generally, raw online search data possesses strong potential to one day serve as a window to observe and aggregate individual consumer behaviour on a massive scale.

Bibliography

Askitas, Nikos & Zimmermann, Klaus F. (2009). "Google Econometrics and Unemployment Forecasting," IZA Discussion Papers 4201, Institute for the Study of Labor (IZA).

Burnham, Kenneth P. (2004). "Multimodel inference: Understanding AIC and BICin model selection," Proceedings of the Amsterdam Workshop on Model Selection. http://www2.fmg.uva.nl/modelselection/presentations/AWMS2004-Burnham.pdf

Choi, Hyunyoung & Varian, Hal (2009a). "Predicting the Present with Google Trends," technical report, Google Inc. http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf

Choi, Hyunyoung & Varian, Hal (2009b). "Predicting initial claims for unemployment benefits," technical report, Google Inc. http://research.google.com/archive/papers/initialclaimsUS.pdf

Della Penna, Nicolas & Huang, Haifang (2009). "Constructing Consumer Sentiment Index for U.S. Using Google Searhes," University of Alberta Working Paper No. 2009-26, University of Alberta.

Ginsberg, Jeremy, et al. (2009). "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* 457, 1012-1014.

Kholodilin, Konstantin et al. (2010). "Do Google Searches Help in Nowcasting Private Consumption? Real-Time Evidence for the US," DIW Discussion Paper 997. Deutsches Institut für Wirtschaftsforschung.

Schmidt, Torsten & Vosen, Simeon (2009). "Forecasting Private Consumption: Survey-based Indicators vs. Google Trends," Ruhr Economic Paper #155. Ruhr-Universität Bochum.

Tierney, Heather L. R. & Pan, Bing (2009). "A Poisson Regression Examination of the Relationship between Website Traffic and Search Engine Queries," MPRA Paper 18413, University Library of Munich, Germany.

Webb, G. Kent (2009). "Internet Search Statistics as a Source of Business Intelligence: Searches on Foreclosure as an Estimate of Actual Home Foreclosures," *Issues in Information Systems* 10(2), 82-87.

Wu, Lynn & Brynjolfsson, Erik (2009). "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales," Proceedings of the International Conference on Information Systems, Phoenix, Arizona.

Appendices

Appendix 1.1 Google Search Data Indices

Name	Search Terms Measured by the Index		
Google Index 1	bankruptcy chapter 7 bankruptcy chapter 13 bankruptcy chapter 7	chapter 13 declare bankruptcy file bankruptcy	file for bankruptcy bankruptcy lawyer how to declare bankruptcy
Google Index 2	chapter 7 chapter 13 chapter 7 bankruptcy file for bankruptcy	bankruptcy lawyer how to file bankruptcy filing bankruptcy filing for bankruptcy	declaring bankruptcy declare bankruptcy file bankruptcy bankruptcy trustee
Google Index 3	file for bankruptcy bankruptcy lawyer how to file bankruptcy	filing bankruptcy filing for bankruptcy declaring bankruptcy	declare bankruptcy file bankruptcy bankruptcy trustee

Note: Search terms are aggregated using the Boolean "OR" function to create a single index. Indices extracted from Google Insights for Search web application at a state-level. Each index measures the relative popularity over time of all search terms combined, as measured by the share of online Google searches containing at least one of these search terms.

Appendix 1.2 Goodness-of-fit using Google index data as the single regressor to explain bankruptcy rates – Model (1.1)

	Least Specific Search Terms --Most Specific Search Terms		
Lag on Google data	Index 1	Index 2	Index 3
Current	0.451**	0.429**	0.485**
1 Month Lag	0.439**	0.419**	0.488**
2 Month Lag	0.398**	0.396**	0.478**
3 Month Lag	0.277**	0.339**	0.354**
4 Month Lag	0.216**	0.352**	0.302**
5 Month Lag	0.185**	0.328**	0.274**
6 Month Lag	0.144**	0.284**	0.233**

**Google coefficient significant at 1% level. Goodness-of-fit measured by within-R-squared value from estimation of Model (1.1)

Appendix 2.1: Nowcasting Regression Results, Pre-Policy Subsample, Model (2.2)

	(Baseline)	(Google Index 3)	(CSI)	(ICE)	(ICC)	(PEI1)	(PEI2)
				Bankruptcy Rate			
Bankruptcy (lag1)	0.5665	0.48562	0.31597	0.25511	0.52214	0.47543	0.4671
	(0.05201)**	(0.04775)**	(0.04888)**	(0.04640)**	(0.04359)**	(0.04978)**	(0.04977)**
Bankruptcy (lag 12)	0.46641	0.48835	0.52108	0.58848	0.37459	0.53687	0.48643
	(0.06456)**	(0.05849)**	(0.05474)**	(0.05217)**	(0.05399)**	(0.06219)**	(0.06154)**
Google (current)		0.24565					
		(0.03076)**					
Google (Lag 1)		0.18583					
		(0.03239)**					
Google (lag 2)		0.02693					
		(0.03416)					
Index of Consumer Sentiment (ICS)			-1.4463				
			(0.10279)**				
ICS (lag 1)			-0.30991				
			(0.13030)*				
ICS (lag 2)			0.39971				
			(0.13219)**				
Index of Current Expectations (ICE)				-1.03445			
				(0.06539)**			
ICE (lag 1)				-0.35006			
				(0.08786)**			
ICE (lag 2)				0.20971			
				(0.08134)*			
Index of Current Economic Conditions (ICC)					-0.75113		
					(0.13054)**		
ICC (lag 1)					0.62174		
					(0.14133)**		
ICC (lag 2)					1.66882		
					(0.14355)**		
Poor Expectations Index 1 (PEI1)						0.14982	
						(0.01928)**	
PEI1 (lag 1)						0.07376	

						(0.01994)**	
PEI1 (lag 2)						-0.01549	
						(0.02092)	
Poor Expectations Index 2(PEI2)							-0.90603
							(0.12255)**
PEI2 (lag 1)							-0.51117
							(0.13132)**
PEI2(lag 2)							0.59594
							(0.13585)**
Constant	-0.09024	-1.67743	6.79123	5.82333	-6.76276	-0.48655	4.1882
	(0.26553)	(0.29645)**	(1.06199)**	(0.60006)**	(1.16964)**	(0.26552)	(1.10654)**
Observations	497	497	497	497	497	497	497
Number of states	31	31	31	31	31	31	31
R-squared	0.3427	0.4688	0.5957	0.6201	0.5574	0.44	0.4752

Standard errors in parentheses
* significant at 5% level; ** significant at 1% level. Pre-policy subsample is January 2004-March 2010. Poor Expectations Index 1, measures the percentage of consumers who responded that they expect to be worse off financially in the coming year. Poor Expectations Index 2 measures the gap between the percentage of people who responded that they expect to be better off financially and the percentage who responded that they expect to be worse off. Google data is measured using Index 3. ICS = "Index of Consumer Sentiment. ICE= Index of Consumer Expectations. ICC = Index of Current Economic Conditions. PEI = Poor Expectations Index.

Appendix 2.2: Forecasting Regression Results, Pre-Policy Subsample, Model (2.3)

	(Baseline)	(Google Index 3)	(CSI)	(ICE)	(ICC)	(PEI1)	(PEI2)
				Bankruptcy Rate			
Bankruptcy (lag1)	0.5665	0.51877	0.60554	0.49697	0.56857	0.55086	0.5586
	(0.05201)**	(0.05069)**	(0.05296)**	(0.05435)**	(0.04430)**	(0.05187)**	(0.05093)**
Bankruptcy (lag 12)	0.46641	0.46717	0.3485	0.45586	0.36901	0.45435	0.41445
	(0.06456)**	(0.06227)**	(0.06372)**	(0.06389)**	(0.05583)**	(0.06509)**	(0.06419)**
Google (Lag 1)		0.21533					
		(0.03430)**					
Google (lag 2)		0.03002					
		(0.03641)					
ICS (lag 1)			-0.52732				
			(0.15452)**				
ICS (lag 2)			1.06512				
			(0.14742)**				
ICE (lag 1)				-0.72405			
				(0.10499)**			
ICE (lag 2)				0.38378			
				(0.09999)**			
ICC (lag 1)					0.59652		
					(0.14608)**		
ICC (lag 2)					1.80928		
					(0.14629)**		
PEI1 (lag 1)						0.07874	
						(0.02117)**	
PEI1 (lag 2)						-0.03738	
						(0.02203)	
PEI2 (lag 1)							-0.66267
							(0.13704)**
PEI2 (lag 2)							0.81009
							(0.14022)**
Constant	-0.09024	-0.87853	-2.23537	1.72089	-10.96287	-0.07896	-0.5817
	(0.26553)	(0.29741)**	(1.01080)*	(0.67140)*	(0.94511)**	(0.2765)	(0.94974)
Observations	497	497	497	497	497	497	497
Number of states	31	31	31	31	31	31	31
R-squared	0.3427	0.3954	0.422	0.4139	0.5257	0.3666	0.413

Standard errors in parentheses. * significant at 5% level; ** significant at 1% level. Poor Expectations Index 1 measures the percentage of consumers who responded that they expect to be worse off financially in the coming year. Poor Expectations Index 2 measures the gap between the percentage of people who responded that they expect to be better off financially and the percentage who responded that they expect to be worse off. Google data is measured using Index 3. ICS = "Index of Consumer Sentiment. ICE= Index of Consumer Expectations. ICC = Index of Current Economic Conditions. PEI = Poor Expectations Index.

Appendix 2.3: Nowcasting Regression Results, Post-Policy Subsample, Model (2.2)

	(Baseline)	(Google Index 3)	(CSI)	(ICE)	(ICC)	(PEI1)	(PEI2)
				Bankruptcy Rate			
Bankruptcy (lag1)	0.82387	0.71587	0.76047	0.79313	0.73247	0.80961	0.80025
	(0.00853)**	(0.01171)**	(0.01132)**	(0.01003)**	(0.01236)**	(0.00977)**	(0.00981)**
Bankruptcy (lag 12)	0.04324	0.02424	0.03334	0.03471	0.03331	0.03596	0.03731
	(0.00603)**	(0.00588)**	(0.00597)**	(0.00605)**	(0.00589)**	(0.00620)**	(0.00602)**
Google (current)		0.11614					
		(0.01708)**					
Google (Lag 1)		0.03198					
		(0.01715)					
Google (lag 2)		0.04036					
		(0.01696)*					
Index of Consumer Sentiment (ICS)			-0.14539				
			(0.04307)**				
ICS (lag 1)			0.02853				
			(0.0498)				
ICS (lag 2)			-0.14608				
			(0.04184)**				
Index of Current Expectations (ICE)				-0.1452			
				(0.03522)**			
ICE (lag 1)				0.11576			
				(0.04151)**			
ICE (lag 2)				-0.17225			
				(0.03461)**			
Index of Current Economic Conditions (ICC)					-0.10189		
					(0.03907)**		
ICC (lag 1)					-0.12131		
					(0.04085)**		
ICC (lag 2)					-0.0409		
					(0.03893)		
Poor Expectations Index 1 (PEI1)						0.03929	
						(0.01110)**	
PEI1 (lag 1)						-0.01613	
						(0.01075)	

PEI1 (lag 2)						0.0429	
						(0.01016)**	
Poor Expectations Index 2 (PEI2)							-0.11412
							(0.04823)*
PEI2 (lag 1)							0.00744
							(0.04965)
PEI2 (lag 2)							-0.23544
							(0.04718)**
Constant	0.45515	0.12329	1.82586	1.43222	1.95612	0.35297	2.1768
	(0.03436)**	(0.04014)**	(0.15411)**	(0.14360)**	(0.14735)**	(0.03791)**	(0.28442)**
Observations	1545	1513	1513	1513	1513	1513	1513
Number of states	32	32	32	32	32	32	32
R-squared	0.8606	0.8782	0.8705	0.8687	0.8723	0.866	0.8668

Standard errors in parentheses

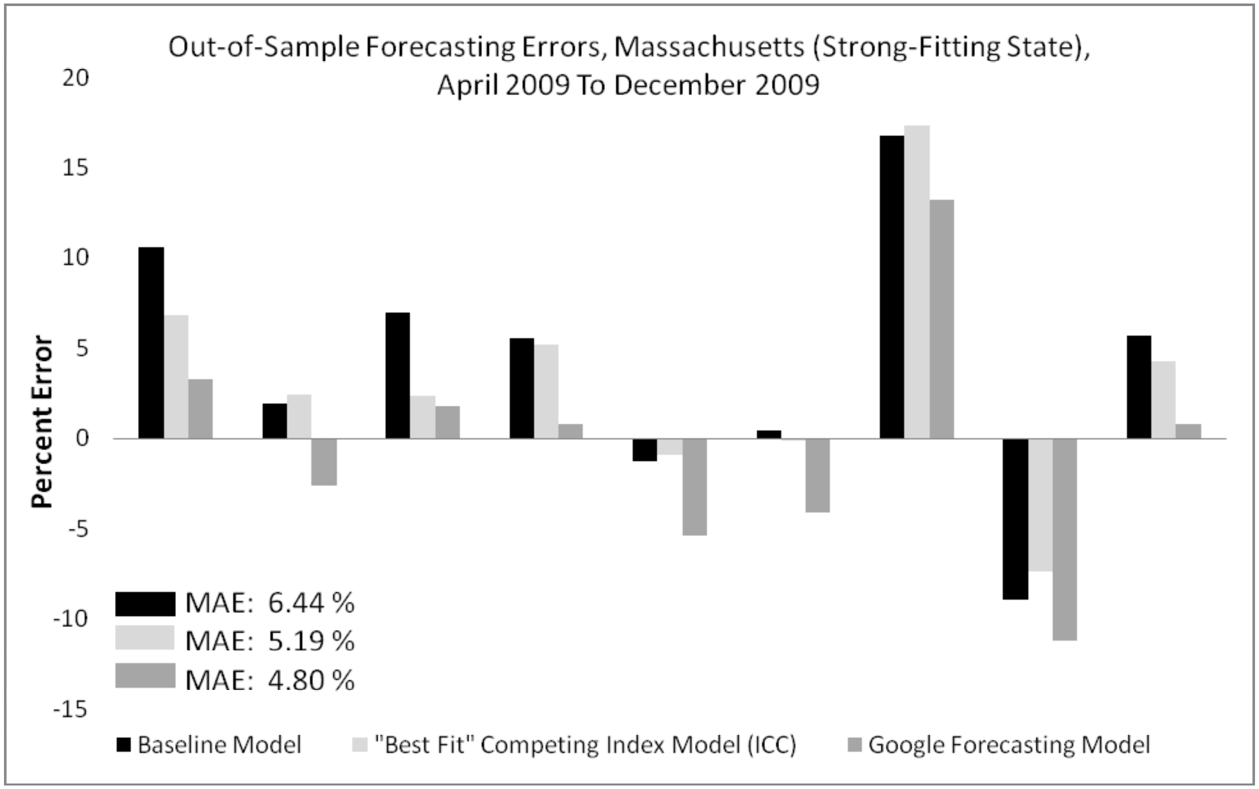
* significant at 5% level; ** significant at 1% level. Post-policy subsample is January 2006-March 2010. Poor Expectations Index 1 measures the percentage of consumers who responded that they expect to be worse off financially in the coming year. Poor Expectations Index 2 measures the gap between the percentage of people who responded that they expect to be better off financially and the percentage who responded that they expect to be worse off. Google data is measured using Index 3. ICS = "Index of Consumer Sentiment. ICE= Index of Consumer Expectations. ICC = Index of Current Economic Conditions. PEI = Poor Expectations Index.

Appendix 2.4: Forecasting Regression Results, Post-Policy Subsample, Model (2.3)

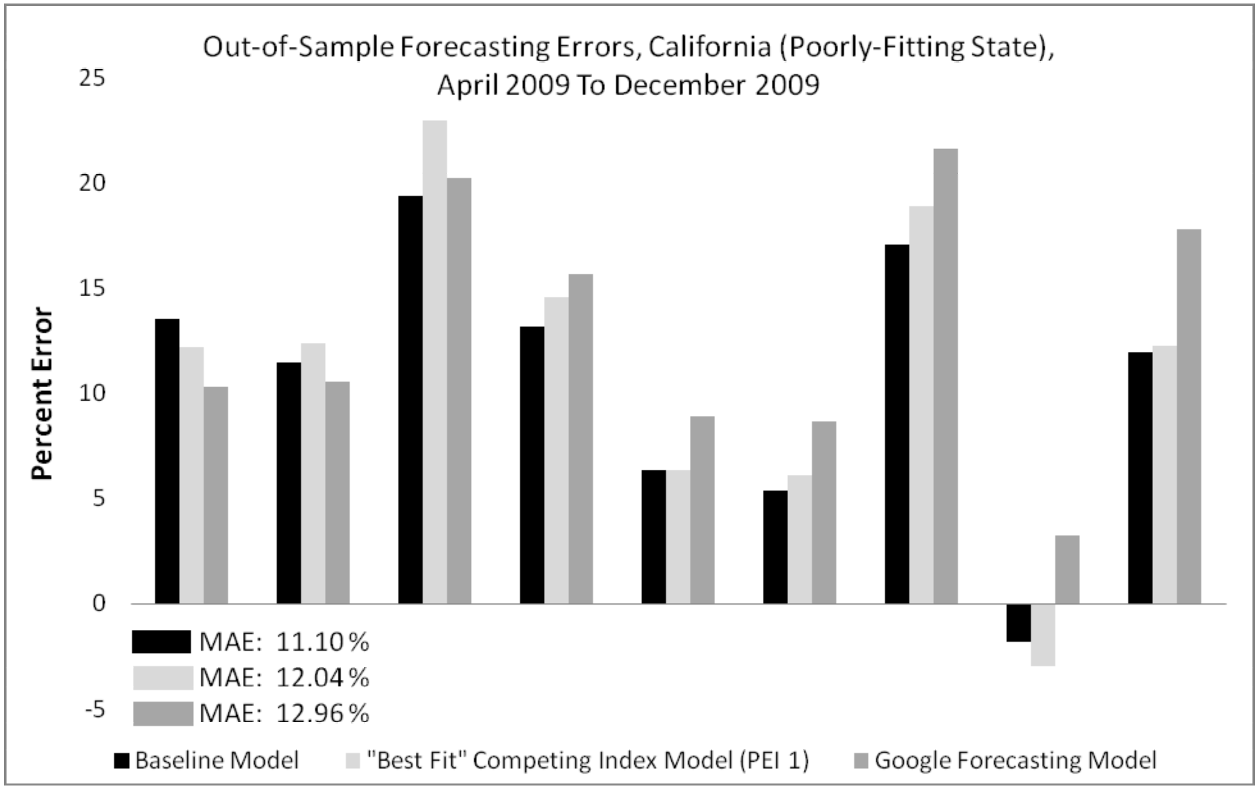
	(Baseline)	(Google Index 3)	(CSI)	(ICE)	(ICC)	(PEI1)	(PEI2)
				Bankruptcy Rate			
Bankruptcy (lag1)	0.82387	0.73421	0.75933	0.78997	0.73564	0.81024	0.79774
	(0.00853)**	(0.01166)**	(0.01114)**	(0.00989)**	(0.01207)**	(0.00953)**	(0.00961)**
Bankruptcy (lag 12)	0.04324	0.02623	0.03296	0.03367	0.03403	0.03717	0.0371
	(0.00603)**	(0.00602)**	(0.00600)**	(0.00608)**	(0.00591)**	(0.00622)**	(0.00604)**
Google (Lag 1)		0.07734					
		(0.01623)**					
Google (lag 2)		0.07187					
		(0.01585)**					
ICS (lag 1)			-0.08509				
			(0.04192)*				
ICS (lag 2)			-0.15736				
			(0.04069)**				
ICE (lag 1)				0.00964			
				(0.035)			
ICE (lag 2)				-0.18819			
				(0.03456)**			
ICC (lag 1)					-0.18947		
					(0.03631)**		
ICC (lag 2)					-0.05175		
					(0.03509)		
PEI1 (lag 1)						-0.00278	
						(0.01037)	
PEI1 (lag 2)						0.04701	
						(0.01003)**	
PEI2 (lag 1)							-0.04944
							(0.04692)
PEI2 (lag 2)							-0.24884
							(0.04606)**
Constant	0.45515	0.21115	1.7407	1.34644	1.84023	0.40226	1.97509
	(0.03436)**	(0.04001)**	(0.15026)**	(0.13851)**	(0.14260)**	(0.03667)**	(0.26341)**
Observations	1545	1545	1545	1545	1545	1545	1545
Number of states	32	32	32	32	32	32	32
R-squared	0.8606	0.8707	0.8675	0.8654	0.8694	0.8626	0.8643

Standard errors in parentheses
* significant at 5% level; ** significant at 1% level. Post-policy subsample is January 2006-March 2010. Poor Expectations Index 1 measures the percentage of consumers who responded that they expect to be worse off financially in the coming year. Poor Expectations Index 2 measures the gap between the percentage of people who responded that they expect to be better off financially and the percentage who responded that they expect to be worse off. Google data is measured using Index 3. ICS = "Index of Consumer Sentiment. ICE= Index of Consumer Expectations. ICC = Index of Current Economic Conditions. PEI = Poor Expectations Index.

Appendix 3.1: Out-of-Sample Forecasting Errors, Select States, Post-Policy Subsample



Appendix 3.2: Out-Of-Sample Forecasting Error, Poor-Fitting State, Post-Policy Subsample


Out-of-Sample Forecasting Errors, California (Poorly-Fitting State),
April 2009 To December 2009
Percent Error
MAE: 11.10%
MAE: 12.04%
MAE: 12.96%
■ Baseline Model
■ "Best Fit" Competing Index Model (PEI 1)
■ Google Forecasting Model