Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2009 Proceedings

Americas Conference on Information Systems (AMCIS)

2009

Forecasting U.S. Home Foreclosures with an Index of Internet Keyword Searches

G. Kent Webb
San Jose State University, webb_k@cob.sjsu.edu

Forecasting U.S. Home Foreclosures with an Index of Internet Keyword Searches

G. Kent Webb
San Jose State University
webb_k@cob.sjsu.edu

ABSTRACT

Finding data to feed into financial and risk management models can be challenging. Many analysts attribute a lack of data or quality information as a contributing factor to the worldwide financial crises that seems to have begun in the U.S. subprime mortgage market. In this paper, a new source of data, key word search statistics recently available from Google, are applied in a experiment to develop a short-term forecasting model for the number of foreclosures in the U.S. housing market. The keyword search data significantly improves forecast of foreclosures, suggesting that this data can be useful for financial risk management. More generally, the new data source shows promise for a variety of financial and market analyses.

Keywords

Risk management, financial forecasting, internet keyword search, mortgage foreclosures.

INTRODUCTION

Acquiring current information on the concerns of customers, potential customers, and market participants often requires expensive survey research. This type of data can be quite valuable to many types of organizations for planning and market analysis [17]. The growing interest in behavioral financial models has also encouraged an interest in this type of information [15, 18]. A relatively new source of data has become freely available from the internet search company, Google. Its "Google Trends [9]" and "Google Insights [8]" programs provide current and historical data on monthly and weekly keyword search volumes worldwide and by geographic regions. This paper presents a simple, short-term forecasting model for U.S. home foreclosures, finding that forecasts are significantly improved by the addition of keyword search data as an independent variable.

The rising number of foreclosures in the U.S. has become a key metric in understanding the current economic recession and the implications for future economic growth. In assessing what went wrong with the risk management models designed to evaluate the U.S. housing market, the former Chairman of the U.S. Federal Banking system commented that improved models would need to incorporate data that would signal when basic consumer attitudes have changed from feelings of euphoria to periods of fear [10]. Keyword search data, not currently used in standard risk or financial management models, may provide one source of signals on changes in attitudes, concerns, and interests.

Google researchers have recently reported success in using the keyword search statistics to detect outbreak of flu by tracking the increase in volume of searches in Google on the keyword "flu" [6]. The logic of the research hypothesis is that as individuals around the world start to experience flu symptoms, they will turn to their computers to get information on their ailment. The Google researchers demonstrate that the search statistics can identify the risk of a flu outbreak up to two weeks faster than the current best detection system, a surveillance program managed by the U.S. Centers for Disease Control (CDC) and Prevention

As noted on the Google site, the pattern of search volume also seems to match with some seasonal patterns. One financial example describes how search volume on "internal revenue service", the tax collection agency for the U.S., increases each year around April 15, the deadline for filing taxes [7]. This example suggests that internet user's interests in financial services or information seem to be reflected by the type of searches they undertake. Of course, this data is limited to internet users and hasn't been collected with the statistical rigor characteristic of some market research, but the data is feely available, very current, and extensive. In a more formal study, a researcher was able to duplicate with statistical significance the results of a market survey identifying green technology investment opportunities using the Google search data [21].

Risk Management Models

Credit information systems play a critical role in housing finance. Countries with extensive information systems have broader and deeper housing markets [20]. Inferior information quality negatively affects risk management [16]. A lack of good information and the resulting poor econometric forecasts of the risk of foreclosure for the subprime mortgage market were critical issues in the development of the recent financial crises [3]. Some lenders developed forecasts for the probability of foreclosure based on historical transactions, data prior to the housing boom of the mid 2000s and the expansion of mortgage availability. Since lenders were compensated based on the number of loans which they then sold off in secondary markets, they had an incentive to ignore or understate future risks.

Many of the standard information system analytical tools are used in current risk management models that evaluate the potential of default for mortgage holders. Monte Carlo simulation has been extensively applied [2, 11]. Given the large amount of internal data generated by banks, data mining tools have been developed to make risk management more effective [4, 13, 22]. Honohan [12], however, criticizes these models as being "too mechanical, albeit sophisticated." Nevertheless, as early as 2003 a statistical analysis of subprime lending resulted in a conclusion that lenders and regulators need to train their attention and understanding on this growing segment of the credit market given the risk of default [5].

RESEARCH OBJECTIVE

As with internet users who experience flu symptoms and so search the internet for information, it is proposed that homeowners in the U.S. who begin to feel financial strain will likewise search for financial information. As the strain increases, owners will begin to face the prospect of foreclosure. They may no longer be able to meet their mortgage payments, defaulting on their loan, with the result that a financial institution takes ownership of the house. This scenario suggests the following research hypothesis:

H1: Keyword search volume on the term "foreclosure" will be positively correlated with actual U.S. home foreclosures and will provide useful data for forecasting.

More generally, a finding that the keyword search data can be used to forecast foreclosures also suggests many other applications where trends in internet user's interests might be beneficial.

SOURCES OF DATA

The U.S. market research firm Realtytrac releases a monthly summary of total U.S. home foreclosures by aggregating government data [16]. This data is commonly referenced in the press when discussing the U.S. foreclosure crises and appears to be the best and most current available data on this financial market. Realtytrac has been collecting this data since January 2005 and provides much geographic detail for the foreclosure data to subscribers of their service.

To encourage the use of their advertising programs, Google Inc. provides a website allowing users to type in a keyword and get a graph and an index of the volume of weekly searches (www.google.com/trends) back to 2004 for terms with significant volume. Figure 1 was created by typing in the word "foreclosure". Letters in boxes on the graph refer to specific news events associated with changes in search volumes. The graph was generated in April, 2009 in response to a reviewer's comments and so contains more data than was used in the analysis.

Rather than providing actual search volume, Google "normalizes" the data to create an index that the website suggests will make the data more useful for analysis. This data can be downloaded in a .csv format, compatible with Excel. Two types of indexes are available: relative and fixed. Both indexes involve dividing all the data by the volume of data for one point in time. With the fixed index, all volumes are divided by the first week for which data is available, usually the first week of January 2004. With the relative index, the first week of the period selected by the user becomes the base period for the index. The fixed index was used for this study.

The Google Trends keyword search index is available as a weekly time series, but the Realtytrac foreclosure data is only available on a monthly time period. In order to match the periodicity of the data, the weekly index was averaged for each month to create a monthly approximation of the keyword search data.

Basic descriptive statistics for the data appear in Table 1. The minimum number of actual foreclosures occurred in March 2005 while the minimum number of searches was in December of that year. Actual foreclosures fell again a few months later in 2006, suggesting a possible lagged relationship in the data. The Pearson Correlation between the two variables is 0.931, significant at the 0.01 level based on a two-tailed test. Month by month, the search index tracks very closely to actual foreclosures.

Table 1: Descriptive Statistics for Actual U.S. Home Foreclosures and the Keyword Search Index of Foreclosures

Monthly Data from January 2005 to December 2008

	Search Index	**Actual Foreclosures**
Minimum	0.7	62,422
Maximum	2.765	303,868
Mean	1.4526	156,266
Pearson Correlation: 0.931, significant at 0.01 level using a two-tailed test.		

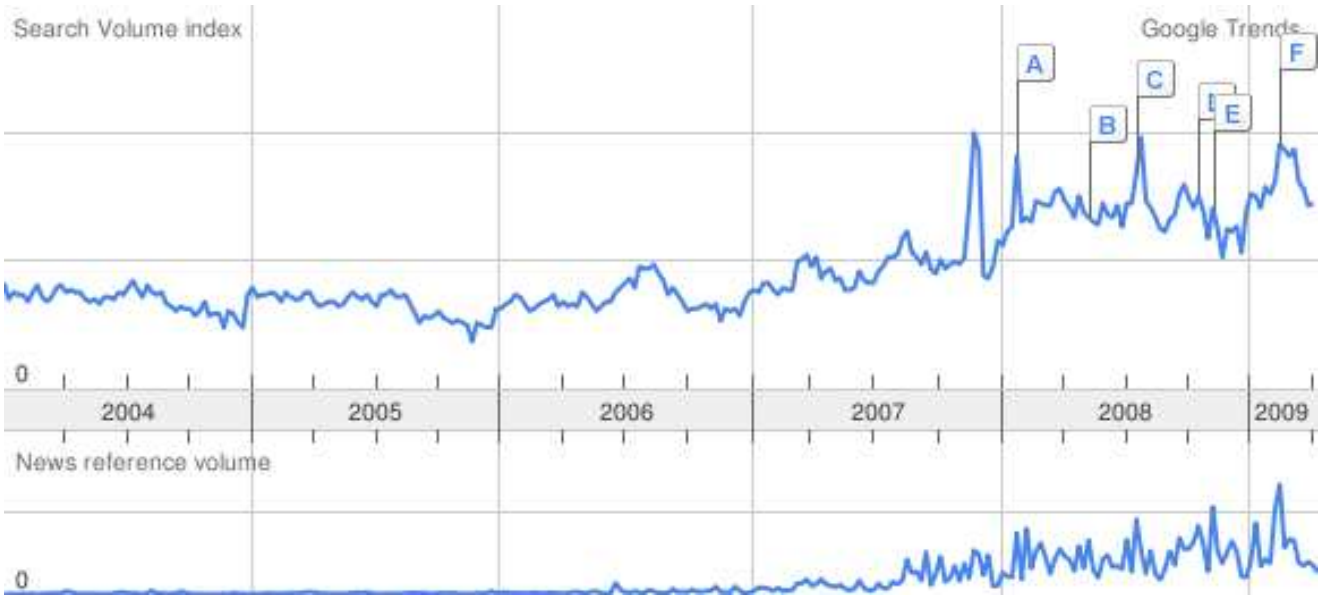

Figure 1: Graphical Results of Typing the Word "Foreclosure" Into Google Trends (Letters in boxes refer to specific news reports: Box A, a news report in ReportonBusiness.com that "US homes in foreclosure soared 79% in 2007"; Box B a report from WTOL that "US foreclosure filings surge 65% in April; Box C, a report in Forbes that "US foreclosure filings more than double in 2Q; Box D, a report from KTVN that "US foreclosure filings up by 71% in 3Q; Box E, a report in the Greater Baton Rouge Business Report that "Foreclosure rates up 25% year-over-year"; Box F, a report in the Tort Deform: The Civil Justice Defense Blog about Obama's foreclosure plan)

RESULTS

The output of a regression model with actual U.S. home foreclosures as the dependent variable appears in Table 2. There are three independent variables. First, the dependent variable, foreclosures, is lagged by one month. The logic of the model is to see if the keyword search index will be an improvement over a simple forecast using just the lagged dependent variable. The second independent variable, a difference variable, is the month-to-month change in the search volume index. The logic is that an increase in searches suggests an increase in concern about foreclosure. The keyword search index is also included as a five-month trailing average, calculation of which reduces the number of observations available for the regression. The data for the regression begins in June rather than January allowing for enough data to calculate the five-month trailing average.

As the table indicates, the two variables based on search index volume are significant at better than the 0.05 level, providing good evidence that as homeowners get close to foreclosure they begin to signal their distress by searching on the internet for related information. Variations using a different number of months to calculate the prior moving average index yielded about the same results, with the five-month indexing having a slightly better significance.

Dependent Variable: Actual U.S. Home Foreclosures R = .968; R-Square = .937; Adjusted R-Square = .933; Number of Observations = 42			
Independent Variables	Estimated Coefficient	t-statistic	Significance
Constant	-4313.121	-0.500	0.620
Previous Month Foreclosures	0.679	5.675	0.000
Monthly Difference in Search Index	36144.291	2.674	0.027
Prior five-month Moving Average of Search Index	42847.12	2.304	0.011

Table 2: U.S. Home Foreclosures Forecast by Previous Month Foreclosures, Current Search Volume Index, and Previous Search Volume Index

Since the equation in Table 2 uses the current month search index to calculate the difference variable, it provides a very short-term forecasting horizon. The search index data comes out almost real-time while the market research data is released two weeks after the end of the month. This is about the lead-time that the Google researchers demonstrated in their study of flu outbreaks.

To increase the forecasting horizon, the difference variable is removed from the model. The results are summarized in Table 3 for an equation using only previous month foreclosures and the prior five-month moving average of the search index. This equation will forecast up to six weeks ahead the release of the actual foreclosure data. The significance for the five-month moving average of search index volume degrades somewhat, but it is still significant at better than the 0.05 level.

The research hypothesis the search index is positively correlated with future foreclosures is supported, suggesting that mortgage holders begin to search for information on foreclosure as they approach defaulting on their mortgages. It seems that this data does improve the forecast for foreclosures over a simple model using a lagged dependent variable.

Dependent Variable: Actual U.S. Home Foreclosures R = .964; R-Square = .929; Adjusted R-Square = .925; Number of Observations = 42			
Independent Variables	Estimated Coefficient	t-statistic	Significance
Constant	-1415.058	-0.158	0.876
Previous Month Foreclosures	0.712	5.693	0.000
Prior five-month Moving Average of Search Index	37700.599	2.257	0.030

Table 3: U.S. Home Foreclosures Forecast by Previous Month Foreclosures and Previous Search Volume Index

Since this model is simply a weighted average approach to forecasting the estimated coefficients have no specific economic meaning except that they represent the weights used to combine them into the forecast. For example, based on table 3 a forecast of foreclosures next month would be 71.2 percent of current month foreclosures plus 27700.599 times the prior five-month moving average of the search index.

LIMITATIONS OF THE APPROACH

This simple model merely creates a weighted average of keyword searches and previous actual foreclosures to forecast in an autoregressive style without taking account of structural issues in the market. For example, rule changes in the U.S. financial industry during early 2009 designed to delay foreclosures will probably result in lower actual foreclosures than forecast. However, the forecast may indicate how many foreclosures might have happened without the policy change.

No attempt is made to compare this model with sophisticated models developed for the credit industry that may rely on other data or more detailed efforts at financial modeling. The goal of this effort is just to demonstrate that the keyword search data provides an interesting proxy variable that might be useful in model development. Further work could be done to integrate this data into a more sophisticated model.

Changes in keyword search for "foreclosures" may be the result of factors other than internet users facing foreclosure and looking for information. For example, since a large proportion of sales in the U.S. are homes that have gone through foreclosure, buyers may be searching for information, increasing searches but not increasing actual foreclosures. Also, over most of the time period for the model estimation foreclosures were rising, however there were a few periods in the early part of the data when foreclosures were falling.

CONCLUSION

A keyword search index for the term "foreclosures" significantly improves the short-term forecasting for U.S. home foreclosures in a simple, short-run model. This result indicates that the Google Trends data can provide useful input for financial forecasting and risk management models. It also suggests that the search data may be useful for analysis in a broader set of applications.

Finding that a five-month moving average works well to forecast foreclosures suggests some mortgage holders at risk of default may be searching for information on the internet well before they fall into foreclosure. Future research on the foreclosure market could focus on examining in more detail how early the search data signals a problem. As the former chair of the U.S. Federal Reserve Bank, Alan Greenspan, recently suggested, credit models would benefit from data that could signal a change in consumer attitudes and concerns.

REFERENCES

1. Baker, H.K., Mukherjee, T.K. (2007). Survey research in finance: views from journal editors. *International Journal of Managerial Finance*, 3,1, 11-25.
2. Calem, P.S. and LaCour-Little, M. (2003). Risk-based capital requirements for mortgage loans. *Journal of Banking & Finance*. 28, 3, 647-672.
3. Carey, W.P. (2008). The subprime crisis and its role in the financial crisis. Journal *of Housing Economics*. 17,4, 254-261.
4. Costa, G., Folino, F., Locane, A., Manco, G. Ortale, R. (2007). Data mining for effective risk analysis in a bank intelligence scenario. *Data Engineering Workshop, 2007 IEEE 23rd International Conference*. 904-911.
5. Cowan, A. M., and Cowan, C.D. (2003). Default correlation: an empirical investigation of a subprime lender. *Journal of Banking & Finance.* 28, 4, 753-771.
6. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature.* Published online 19 November 2008. Available at http://www.nature.com/nature/journal/vaop/ncurrent/full/nature07634.html
7. Google, available at: http://www.google.com/trends?q=IRS
8. Google Insights, available at: http://www.google.com/insights/search
9. Google Trends, available at: http://www.google.com/trends
10. Greenspan, Alan (2008). We will never have a perfect model of risk. *Financial Times*, March 16.

11. Hall, J. & Berry, M., (2006). Making housing assistance more efficient: a risk management approach. *Urban Studies*, 43, 9, 1581-1604.
12. Honohan, P. (2008). Risk management and the costs of the banking crisis. *National Institute Economic Review*. 206, 1, 15-24.
13. Koh, H.C., Wei, C.T., Chwee, P.G., (2006). Two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information*. 1, 1, 96-118.
14. Mark, R. M., and Krishna, D., (2008). How risky is your risk information. *Journal of Risk Management in Financial Institutions.* 1,4, 439-451.
15. Neuhauser, K.L. (2007). Survey research in finance. *International Journal of Managerial Finance* 3, 1, 5-10.
16. Realtytrac, data available by subscription at: http://www.realtytrac.com
17. Sahay, B.S., Ranjan, J (2008). Real time business intelligence in supply chain analytics. *Information Management & Computer Security*. 16,1,28-48.
18. Shiller, R.J. (2002). From efficient market theory to behavioral finance. *Cowles Foundation Discussion Paper No. 1385*. Available online at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=349660
19. Stokes J.R., Gloy, B.A. (2007). Mortgage delinquency migration: an application of maximum entropy econometrics. *Journal of Real Estate Portfolio Management* 13, 2, 153-160.
20. Warnock, V.C., and Warnock, F.E. (2008). Markets and housing finance. *Journal of Housing Economics*. 17,3, 239-251.
21. Webb, G.K. (2007). Analysis of pages and metrics related to global environmental management. *Issues in Information Systems*, 9(2), 111-116
22. Zhang, D., Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *IEEE Transactions on systems, Man, and Cybernetics, Part C: Applications and Reviews.* 34, 4, 513-522.