Peter Fieger
University of New England, Australia

John Rice
Zayed University, United Arab Emirates

Modeling Chinese Inbound Tourism Arrivals into Christchurch

New data and modeling approaches are improving the usefulness of Internet search data for forecasting inbound tourist arrivals. Previous research has focused on Google Trends as a source of search data to augment tourism forecasting capabilities. In the context of rapidly increasing Chinese outbound tourism Google data lacks the market penetration in China to produce reliable auxiliary data for tourism forecasting. This short paper provides evidence of the usefulness of Baidu search data in predicting Chinese inbound tourist arrivals into a specific region in New Zealand. It also compares three modeling approaches, finding a Vector Autoregressive approach the most useful.

Key words: Forecasting, Outbound Tourism, China, New Zealand, Internet Search Data, Modelling, Baidu

Peter Fieger
Business School
University of New England
Armidale, NSW
Australia
Phone: [61] 2 6773 3333
Email: pfieger2@une.edu.au

John Rice
College of Business
Zayed University
PO Box 144 534
Abu Dhabi
U.A.E.
Phone: [971] 2 599 3213
Email: john.rice@zu.ac.ae

Introduction

The prediction of future demand is important for both tourism industry stakeholders like airlines and accommodation providers, and for national policymakers alike (Divisekera and Deegan 2010). Tourism, as an industry, can be both cyclical and volatile. Major external shocks (natural and man-made) can result in significant negative demand for tourism services. Cyclicality can relate to seasons, holiday times and other factors. While these factors persist, making planning challenging, globally overall tourism demand continues to grow.

Outbound markets often have strong preferences for certain accommodation and experience attributes, making an understanding of future demand from key markets important (Razzaq, Hall and Prayag, 2016). This is especially true from emerging source (outbound) countries like China. New analytical techniques and newly available data sources have assisted in improving tourism forecasting's predictive power (for example, D'Urso *et al*., 2015). Among the most important adjuncts to traditional predictive data has been web search data. This data aggregates the use of key search terms in outbound market locations to better predict inbound tourism demand.

Forecasting visitor arrivals to a destination using web search data is not new, although modeling advances and improved data availability make an assessment of alternative modeling and recently available data options a useful exercise. Recent research in the inbound tourism forecasting arena has emphasized the importance of both push and pull factors (Prayag and Soscia, 2016; Zhou-Grundy and Turner, 2014) – push in relation to source country economic factors and pull factors in relation to recipient country infrastructure. Relatively persistent factor conditions relating to weather and natural attractions also play a fundamental role in the emergence of inbound tourism demand.

Developing better forecasting models is thus an important challenge with significant benefits. Increasingly, researchers are identifying the information gap that exists between potential visitors and their intended location as a focal area of research. In this regard, in tourism research, search engine data is proving to be of strong value in improving the predictive ability of models that focus only on pull and push factors (Bangwayo-Skeete and Skeete, 2015). Google search data has proven useful in enhancing predictive models, but Google is not universally available. Due to historical and current geopolitical factors, Google is widely unavailable in China (for more background information, see Tan & Tan, 2012). This limits the usefulness of Google search data in predictive models of outbound Chinese tourism demand.

In this paper, we use arrival data of Chinese nationals at Christchurch International Airport, as well as search data from Baidu, the major Chinese search engine, to develop an approach that compares more traditional univariate time series analysis methods such as Autoregressive Integrated Moving Average (ARIMA) and Holt-Winters seasonal forecasting with multivariate Vector Autoregression. While Google has achieved the status of the predominant search engine in most international markets, its success, or lack thereof, in China has been volatile. After allegations of interference in various Gmail accounts of Chinese human rights activists, Google withdrew its already censored services from the Chinese market altogether in 2010, only providing a link to the search engine of its Hong Kong sister site on its website. As a result, Google's share of Chinese search engine queries declined to below 2 percent by 2013 (Statcounter, 2013). The Chinese search engine market is now dominated by Baidu, whose search engine accounts for a market penetration of 92.1 %, compared to Google's 27.4% (CNNIC, 2015).

Like Google Trends, Baidu Index provides an indicator of the volume of a certain proportion of its more common search queries available to the public for analysis. Also like Google, the algorithm that produces the search popularity figures is somewhat opaque, but unlike Google, Baidu produces its search figures on a continuous scale. This contrasts Google, that provides only indexes of its search data to 100 (with 100 representing the point of maximum searches during the time interval that is being analyzed). A more thorough discussion about the similarities and differences between Google Index and Baidu Trends can be found in Vaughan and Chen (2015). In any event, it appears reasonable to assume that the data that both search engines providers supply can be considered as reasonable proxies for actual search volume patterns for a particular term.

For this study, we aimed to utilize a number of search terms that we assumed would be used by intending visitors to Christchurch. These are the term of 'Christchurch' itself, as well as the search for several other relevant tourist locations in New Zealand. These terms included Christchurch (基督城), Queenstown (皇后镇), Lake Tekapo (蒂卡波湖), New Zealand (新西兰), Arrowtown (箭镇), Lake Wanaka (瓦纳卡湖), and Auckland (奥克兰市). There are some difficulties with the use of the Chinese transcriptions of some of these location names. For instance, the Chinese term for Auckland, 奥克兰市, may in Chinese also be mistaken for the Californian city of Oakland. Similarly, the name of Christchurch, generally translated as 基督城, may also be translated as Christ's city. Due to the ambiguity of the translation, we decided not to use the Chinese term for Auckland in any subsequent analysis. We further assumed that searches for Christ's City would only represent a small and negligible proportion of searches for Christchurch, and thus the Chinese translation for Christchurch would be appropriate to use.

As Baidu Index makes data for analysis only available for a limited number of search terms we eliminated those terms for which no search volume data were offered. This yielded the following search terms as candidates for inclusion in our forecasting model: Christchurch, New Zealand, and Queenstown. We obtained the search volume scores for these locations for the period from January 2011 to August 2016. Our rationale for the chosen search terms is that the motivations underpinning searches for these terms comprise three elements: the search for the location itself with the possibility of visiting, the search in relation to a specific one-off event, and the search for the search term for other random reasons.

Taking the example of Christchurch, we would assume that the variation in search frequency for this term would largely be a result of potential visitor interest, plus the outliers created by searches for specific events related to Christchurch (such as the 2011 earthquake), plus random searches for Christchurch (or Christ's city) that can be described as statistical noise. The search volume time series for Christchurch included three obvious outliers (Figure 1), indicating special one-off events (for which Baidu Index also provides the context): The major Christchurch earthquake in February 2011, a promotional campaign for a Christchurch educational institution in June 2013, and the Valentine's Day earthquake of February 2016.

Figure 1: Baidu search patterns for Christchurch

We smoothed all three events over with moving average data from the immediate adjacent periods. Search volume data for the other locations did not contain any obvious outliers. Thereafter we calculated monthly averages of search volume scores, in order to be able to synchronize search volume data with visitor arrival data. The time series for the monthly arrival of Chinese visitors at Christchurch airport was obtained from Statistics New Zealand. For the purpose of this analysis, we used the data from early 2011 to July 2016. Our paper thus has three aims: first to evaluate the predictive capacity of all three approaches, second to assess the value of including search data in the multivariate model and third to assess the value of Baidu search data for predicting Chinese outbound tourism demand.

Empirical Models and Results

This research utilises three different approaches to forecasting Chinese visitor arrivals into Christchurch. The first univariate approach employs the Holt-Winters forecasting method. This method was specifically designed to forecast univariate time series that exhibit a clear seasonal component. The underlying concept of this method is to draw on trend and seasonality inherent in the time series to produce forecasts. For an in-depth description of this approach see Holt (2004) and Winters (1960). As the visitor arrivals time series displays a clearly increasing pattern of seasonality we employed the multiplicative version of the Holt Winters forecasting method (see Chatfield, 1978).

The second univariate method utilizes the autoregressive integrated moving average (ARIMA) approach as devised by Box and Jenkins (1968). This methodology requires the existing times series (e.g. arrival) to be stationary (e.g. the means and standard deviations should be similar at different intervals of the series). After inspecting the visitor arrival time series for existing auto-correlations and partial autocorrelations we decided to build an ARIMA model of the form of (0,1,1)(0,1,1)x12, which mimics the structure of the well-known ‘Airline Model’.

Finally, we specified a vector autoregressive model to forecast visitor arrivals. This methodology has emerged relatively recently and was originally pioneered by Watson (1994) and Hamilton (1994). Calibration of a VAR model is significantly more complex than univariate models and requires a number of pre- and post-model building tests. We utilised the visitor arrival series as well as the Baidu keyword search series for the Chinese terms for ‘Christchurch’ and ‘New Zealand’ and ‘Queenstown’. Additionally, to account for the seasonality in the data we added the month as an exogenous term. In preparation, all four endogenous time series were transformed using their natural logarithm and were differenced to achieve stationarity.

The necessary lag order for this particular model was determined to be 4, using Akaike’s Information Criterion. The model (see Appendix I) showed a high degree of predictive capacity (R-squared= 0.86) and post-model building analysis revealed that the model exhibited satisfactory statistical properties e.g. there was no remaining autocorrelation at the chosen lag order and that disturbances were largely normally distributed.

The prediction and forecasting results of all three models were superimposed over the original visitor arrival times series and can be seen in Figure 2. To the left of the vertical red line, it is obvious that all three models produce predictions that follow the general pattern of the time series very well. We calculated the relative mean square error (RMSE) for each of the three model predictions via

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$

We found that of the three forecasting models used, the two univariate models tended to produce inferior results to the vector autoregressive model (Holt Winters provided an RMSE of 601.353, ARIMA provided an RMSE of 854.222 and our Vector Autoregressive (VAR) model delivered an RMSE of 443.603). While prediction of past arrival data is no guarantee for the accurate forecast of future arrivals, it is the VAR model appears most instructive for forecasting purposes (Figure 2). The model was built using data covering the period from April 2011 to June 2016 (to the left of the vertical line). We then added forecast and post-model arrival data for another 12 months.

In the forecasting period, to the right of the vertical line, we can see a sustained rise in visitor numbers along with a solidified seasonal pattern. The ARIMA and Holt-Winters models appear to be slightly more optimistic than the VAR model in their forecasts. In the forecasting period of four or five months immediately following we can see that the VAR data follows the actual observed pattern significantly closer than the forecasts obtained via the Holt-Winters and ARIMA methods. However, it is also apparent that following the initial forecasting period actual and predicted data diverge substantially. The very stark contrast in predicted and observed peak visitation that can be observed in January/February can most likely be ascribed the occurrence of Chinese New Year, which occurred in the previous year in February, while it was celebrated in 2017 in the month of January. Chinese New Year is usually associated with significantly increased travel activities of Chinese residents (Pan et al., 2006).

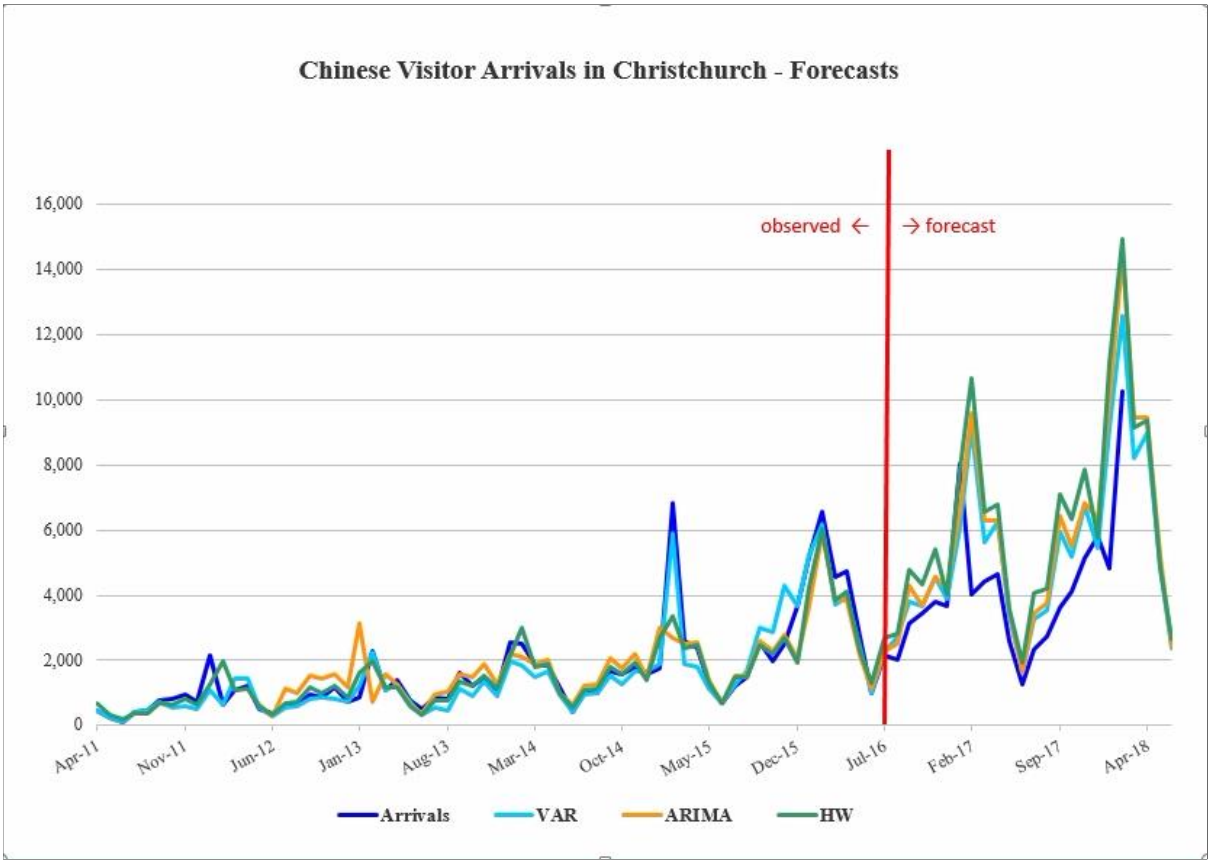

Figure 2: Predictions and forecasts of visitor arrivals

The forecast of the VAR model is of particular interest as it includes the element of Internet search data in its prediction. We performed a Granger causality test on our VAR model in an effort to determine whether the endogenous variables in this model are indeed useful for forecasting visitor arrivals. The null hypothesis in this test is that the tested variable does not cause the outcome variable. As can be easily seen, all three hypotheses could be rejected and the predictive value of Internet search data for visitor arrival can indeed be confirmed (Table 2).

Table 2 Granger causality tests

Granger Causation	Chi2	df	P-value
Baidu 'Christchurch' --> Christchurch arrivals	15.45	4	0.0039
Baidu 'New Zealand' --> Christchurch arrivals	11.55	4	0.0210
Baidu 'Queenstown' --> Christchurch arrivals	19.45	4	0.0005

It is also of interest whether the direct impact of Internet search activity can be further defined and quantified. To investigate this question, we estimated the impulse response function that measures the effect that the search for the Chinese term for Christchurch has on Christchurch visitor arrivals of Chinese nationals (Figure 3).

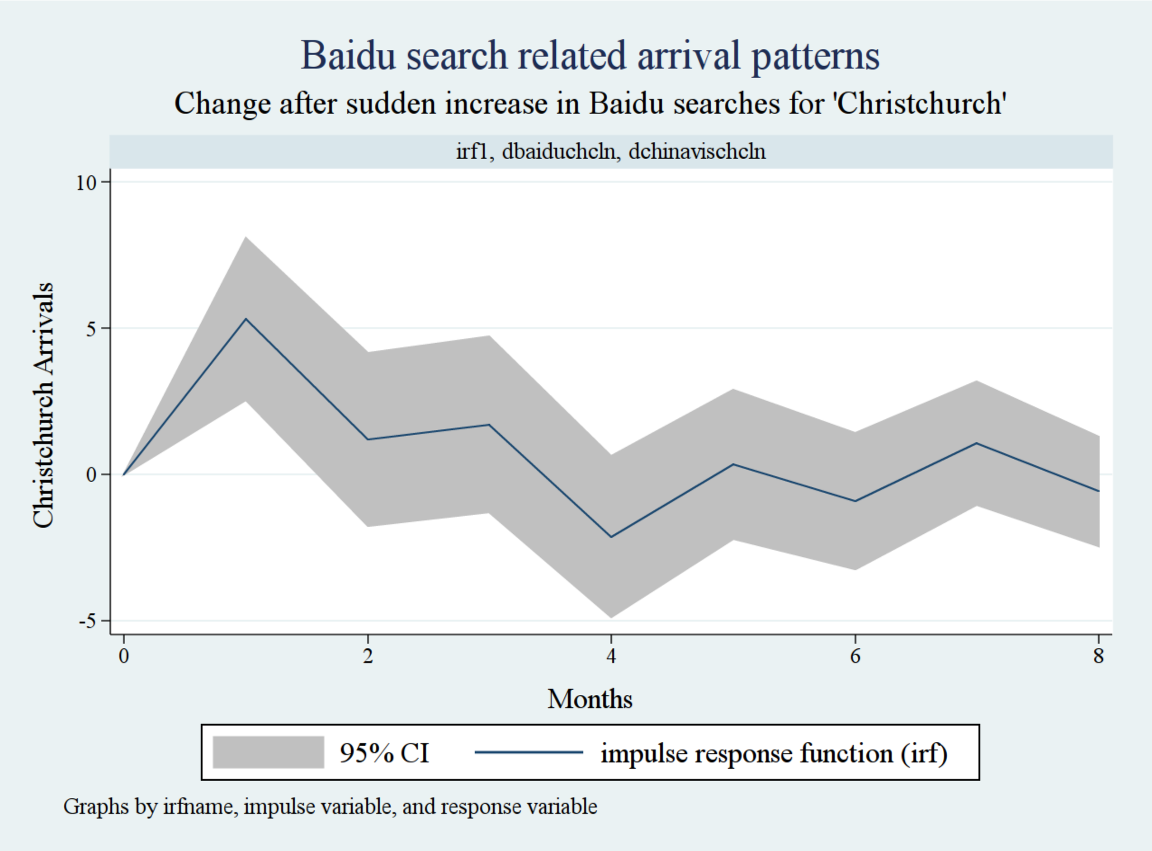

Figure 3: Impulse response function, Christchurch search on visitor arrivals

Figure 3 indicates that surges in search volumes for Christchurch are followed by an immediate significant increase in visitor arrivals in the following month, followed by a continued positive impact (albeit non-significant) for two further months. This finding confirms that Internet search data are suitable for supplementing short-term visitor forecasting in respect to visitor arrivals originating from China.

Conclusions

The purpose of this study was threefold – to provide a forecast for Chinese demand for a localized inbound tourism market, to show that Baidu search data is a useful adjunct to other evidence in developing forecasts for future Chinese tourism demand and finally to compare the predictive power of three quantitative forecasting approaches using this data to actual visitor arrivals during the period investigated here. In turn, the paper first found that the current growth pattern in visitor arrivals from China into Christchurch is likely to continue in the immediate future. Growth in Chinese interest in Christchurch, which we have shown is a significant predictor of future arrivals, has continued to increase in the most recent period.

Second, in assessing the predictive power of the models presented here we found that Baidu search behaver is indeed a significant predictor of later tourism activity. This is an important finding. It reinforces other research that has focused on Google Trend data which, due to historical and regulatory issues, is generally of limited use in the Chinese context.

Finally, we employed three conceptually different methods to undertake the forecasts, all of which yielded similar results. We noted, however, that a Vector Autoregression (VAR) approach was demonstrably superior in predicting later demand than other forecasting models commonly used in the relevant literature.

References

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, *46*, 454-464.

Box, G. & Jenkins, G. (1968). Some recent advances in forecasting and control. *Applied Statistics*, 17(2), 91-109.

Chatfield, C. (1978). The holt-winters forecasting procedure. *Applied Statistics*, 264-279.

CNNIC (China Internet Network Information Center). (2015). The penetration rates of search engines inChina as of 2014. The 35th Statistical Report on Internet Development in China Available at: https://www.cnnic.cn/hlwfzyj/hlwxzbg/201502/P020150203551802054676.pdf.

Divisekera, S., and J. Deegan. (2010). "An Analysis of Consumption Behaviour of Foreign Tourists in Ireland." Applied Economics 42 (13): 1681–1697.

D'Urso P., Disegna M., Massari R. and Prayag G. (2015) Bagged fuzzy clustering for fuzzy data: An application to a tourism market. *Knowledge-Based Systems 73*: 335-346.

Hamilton, J. D., (1994*) Time Series Analysis*, Princeton, Princeton University Press.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5-10.

Pan, G. W., Scott, N., & Laws, E. (2006). Understanding and sharing knowledge of new tourism markets: The example of Australia's inbound Chinese tourism. *Journal of Quality Assurance in Hospitality & Tourism*, *7*(1-2), 99-116.

Prayag G. and Soscia I. (2016). Guilt-Decreasing Marketing Appeals: The Efficacy of Vacation Advertising on Chinese Tourists. *Journal of Travel & Tourism Marketing, 33*(4): 551-565.

Statcounter (2013), Top 5 Desktop, Tablet & Console Search Engines in China from Q3 2008 to Q2 2015 | StatCounter Global Stats

Tan, J., & Tan, A. E. (2012). Business under threat, technology under attack, ethics under fire: The experience of Google in China. *Journal of business ethics*, *110*(4), 469-479.

Vaughan, L., & Chen, Y. (2015). Data mining from web search queries: A comparison of google trends and baidu index. *Journal of the Association for Information Science and Technology*, *66*(1), 13-22.

Watson, M. W., (1994), Vector Autogregression and Co-integration. In Vol. IV of *Handbook of Econometrics*, ed. R.F. Engle and D.L. McFadden. Amsterdam: Elsevier.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324-342.

Zhou-Grundy, Y., & Turner, L. W. (2014). The Challenge of Regional Tourism Demand Forecasting: The Case of China. *Journal of Travel Research*, DOI: 0047287513516197.

Appendix 1

Vector Autoregression model Chinese Visitor Arrivals
(secondary equations omitted)

	Coefficient	StdErr	z	P>z	95% Conf. Interval	
Chinese visitor arrivals (L,D)						
lag 1	-0.79	0.13	-6.22	<0.01	-1.04	-0.54
lag 2	-0.46	0.16	-2.92	<0.01	-0.77	-0.15
lag 3	-0.06	0.15	-0.37	0.71	-0.35	0.24
lag 4	0.05	0.12	0.45	0.65	-0.17	0.28
Baidu 'Christchurch' (L, D)						
lag 1	3.55	1.20	2.97	<0.01	1.21	5.90
lag 2	0.02	0.99	0.02	0.98	-1.92	1.97
lag 3	0.59	1.13	0.52	0.60	-1.62	2.80
lag 4	1.38	0.96	1.43	0.15	-0.51	3.27
Baidu 'New Zealand' (L, D)						
lag 1	0.50	0.33	1.52	0.13	-0.15	1.14
lag 2	0.42	0.45	0.93	0.35	-0.47	1.30
lag 3	-0.43	0.48	-0.91	0.37	-1.38	0.51
lag 4	0.01	0.36	0.03	0.98	-0.70	0.72
Baidu 'Queenstown' (L, D)						
lag 1	0.10	0.45	0.22	0.83	-0.78	0.98
lag 2	-0.86	0.37	-2.32	0.02	-1.58	-0.13
lag 3	0.27	0.36	0.74	0.46	-0.44	0.98
lag 4	0.03	0.41	0.08	0.93	-0.77	0.84
Month						
February	0.18	0.18	1.01	0.31	-0.17	0.54
March	-0.52	0.17	-3.08	<0.01	-0.85	-0.19
April	-0.25	0.20	-1.24	0.22	-0.65	0.15
May	-1.01	0.22	-4.53	<0.01	-1.45	-0.57
June	-1.35	0.24	-5.63	<0.01	-1.82	-0.88
July	-0.54	0.29	-1.90	0.06	-1.11	0.02
August	-0.15	0.28	-0.53	0.60	-0.69	0.39
September	0.45	0.23	1.95	0.05	0.00	0.91
October	0.07	0.22	0.34	0.73	-0.35	0.50
November	-0.26	0.18	-1.39	0.16	-0.62	0.10
December	-0.59	0.16	-3.62	<0.01	-0.91	-0.27
Intercept	0.35	0.14	2.45	0.01	0.07	0.62

Models statistics: RMSE: 0.281861 R-sq: 0.8639 Chi-2: 368.1708 P>Chi-2: <0.001