Prediction in Projection Using Google Search Trends

Allison Morgan*
University of Colorado, Boulder

(Chaotic Dynamics Final Project)

(Dated: May 5, 2017)

I. INTRODUCTION

In this paper, we examine three Google search trends and ask whether or not we can predict them using delay-coordinate embedding. Usually we cannot measure, or may not even know, every state variable of a dynamic system. Often, and in this case, all we have is a time series of just one of the system's observable state variables. A method for trying to reconstruct the state space of the system is called delay-coordinate embedding. The technique dates back to Packard et al.'s [17] work constructing a set of vectors using time-delayed coordinates of the original time series. Shortly thereafter, Takens [19] proved that this method resulted in a topologically correct representation of the true dynamics of the system. Building off this work, researchers have used the Lorenz Method of Analogues (LMA) to generate successful forecasts of a wide range of time series: from childhood rates of the chicken pox and measles, to predictions of computer performance [5, 14, 16]. In this paper, we use delay-coordinate embedding and LMA to see if we can accurately forecast Google search trends.

II. METHODS

The code used to generate the data and plots for this paper can be found on Github [15].

A. Data Collection

In 2007, Google Trends [11] was published to allow users to explore trending search topics. As a search engine, Google uses search history to return personalized search results [9] and predict search requests [10]. These time series are generally non-linear, with an unknown number of state variables.

In this paper, we use an unofficial Google Trends client written in Python [7]. We pass the API a keyword and a time frame for which to return search results across. The trend returned represents relative search interest, scaled across the window of time requested. Furthermore, the resolution of the time series varies with the window requested (see Fig. 1).

In order to obtain the highest resolution data set possible, while still reflecting the underlying trend, we needed to identify the scaling factor between points across two time windows. By making requests by the week, we were able to access hourly data. We also asked for the same hour at the beginning and ending of two requests, to identify the scaling factor. Then we stitched together a coherent time series reflecting relative search interest of a topic within hourly precision.

Hourly data is only available dating as far back as January 1st, 2015. Our data contains hourly snapshots up to April 20th, 2017 resulting in each time series containing 20,280 points total. It should be noted that, unlike calculations of the Lyapunov exponent or fractal dimension which require preservation of the diffeomorphism between real and reconstructed dynamics, there is no hard limit on the number of data points required [18]. These time series can be generated for any keyword. In this paper we study three periodic trends in particular: (1) "influenza," (2) "full moon," and (3) "baseball".

B. Weighted Permutation Entropy

Permutation entropy (PE) involves constructing sets of ranked ordinals across a time series [1]. Calculating PE requires the tuning of three parameters: the number of points in the ranked ordinal set l, the spacing between each point in the ranking τ, and the number of points to look across N. A variant of PE, weighted permutation entropy (WPE), takes into account not only the ordering of the time series but also its amplitude [3]. The probability of seeing an ordinal ranking π is

$$p_w(\pi) = \frac{\sum_{i \leq N-l} w(x^l_{i+1}) \cdot \delta(\phi(x^l_{i+1}, \pi))}{\sum_{i \leq N-l} w(x^l_{i+1})}$$

where the weight associated with an ordinal ranking is

$$w(x^l_{i+1}) = \frac{1}{l} \sum_{j=1}^{i+1} \left(x_j - \bar{x}^l_{i+1}\right)^2$$

the average difference between each point in the series and the average of the whole series. Altogether, weighted permutation entropy over a window of length N from the set $\mathcal{S}_l$ of ordinal rankings π is

$$\text{WPE} = -\sum_{\pi \in \mathcal{S}_l} p_w(\pi) \log_2 p_w(\pi)$$

* Allison.Morgan@colorado.edu

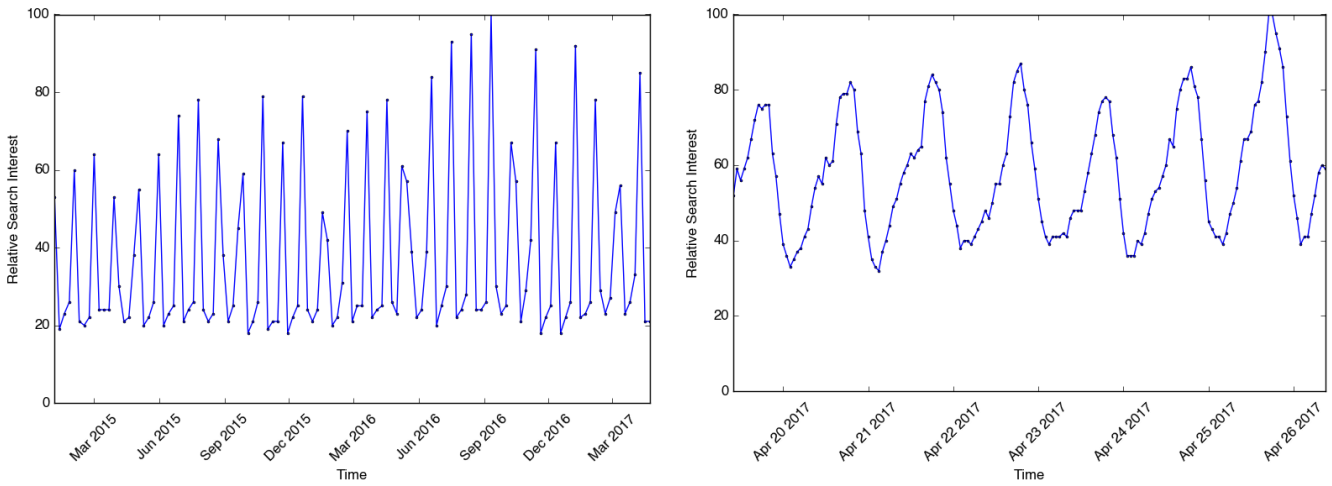

FIG. 1. Search interest in "full moon" from January 1st 2015 to present day (left), and from April 19th through 26th (right). Notice the change in scale and different resolutions. In Fig. 4, the stitched together trend is shown.

For WPE ≈ 0, the trend is totally predictable, whereas for WPE ≈ 1, the trend is unpredictable. For all three calculations of weighted permutation entropy, we chose $l = 4$, $\tau = 24$, and $N = l! * 100 = 2400$ [4].

C. Lorenz Method of Analogues

Lorenz method of analogues (LMA) is essentially a k-nearest neighbor search in the embedding space [14]. First we divide our time series into training (80% of the time series) and prediction (remaining 20% of the time series) sets. Then the prediction set is delay-coordinate embedded, by appropriate choices of delay parameter τ and embedding dimension m using the TISEAN package [12, 13]. Next, for each point in our prediction set, we will choose the k nearest neighbors to that point in our full embedding. The choice of k represents a trade-off between bias and variance in our prediction task. As k increases, we better suppress noise in our data, but we also lose precision. Once we have chosen the k nearest neighbors to a point in our test set, we follow those points from our training set and return their average. In the simplest case we follow the points for one step in time. We could also generate a multi-step forecast, following those points multiple steps forward in time [6]. In this paper, we present single-step forecasts with $k = 5$.

D. Mean Absolute Scale Error

To measure our prediction accuracy, we will calculate the mean absolute scaled error (MASE) between our prediction and the true signal as such:

$$MASE = \sum_{j=n+1}^{k+n+1} \frac{|p_j - c_j|}{\frac{k}{n-1}\sum_{i=2}^{n} |x_i - x_{i-1}|}$$

The numerator of MASE represents the average prediction error ($|p_j - c_j|/k$), and the denominator represents the average error under a random walk prediction ($|x_i - x_{i-1}|/(n-1)$), using the previous value in the observed signal as the forecast. A MASE < 1 means our prediction error was on average smaller than a random walk forecast. On the other hand, MASE > 1 means that our prediction method did worse on average when compared to a random walk. Furthermore, (1/MASE) represents the factor by which we did better than a random walk [6].

III. RESULTS

A. Influenza

In this paper, search interest in influenza was chosen because there is scientific interest in predicting how many people will be sick next year, how much of next year's flu vaccine should be made, etc. Search interest in the flu has shown to be well correlated with historical cases of the flu [8], but also have been shown to be poor predictors of the 2011-2013 US flu seasons [2]. The trend in relative search interest is shown in Fig. 2.

We plotted WPE across the trend in Fig. 3 using $l = 4$, $\tau = 24$, and $N = 100 * l! = 2400$. Since WPE is so high (average of 0.7946), this time series appears largely unpredictable.

In order to run LMA, we need to delay-coordinate embed our data under appropriate choices of delay τ and

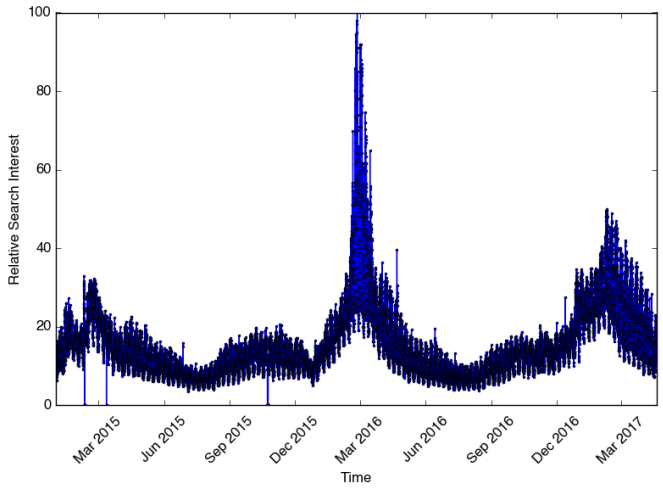

FIG. 2. Hourly search interest in "influenza" from January 1st 2015 to April 20th 2017

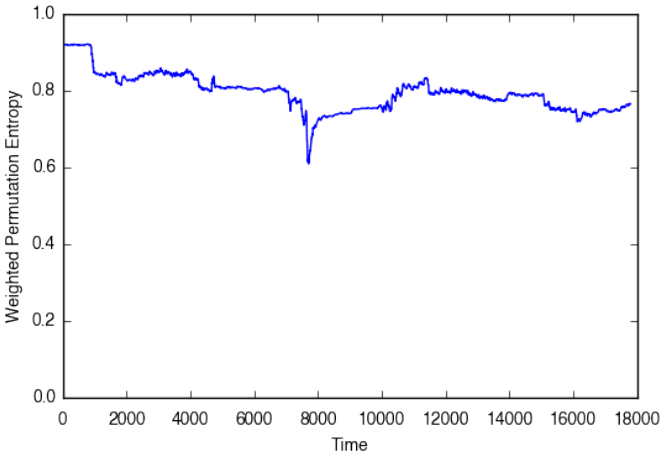

FIG. 3. Weighted permutation entropy for search interest in "influenza"

embedding dimension m. The mutual information oscillates quite a lot around its overall shape. We chose the very first minimum at $\tau = 13$. Using this value of τ, we plotted the percentage of false nearest neighbors (FNN) as a function of embedding dimension. We chose $m = 5$ such that FNN dropped below 10%.

Our prediction is shown at the top of Fig. 8. The MASE is 1.2518, meaning that on average we did worse than a random walk.

B. Full moon

Search interest in the full moon was chosen because we were curious if a trend with a much smaller period of oscillation would affect prediction accuracy. The trend is shown in Fig. 4.

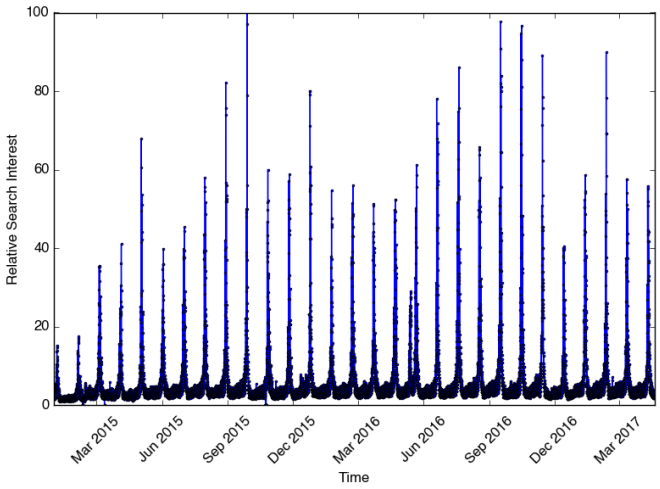

FIG. 4. Hourly search interest in "full moon" from January 1st 2015 to April 20th 2017

In Fig. 5, WPE for our trend is shown under the same parameters l, t and N as mentioned earlier. WPE is significantly lower than it was for influenza (average of 0.5360). This suggests that search interest in full moon is more predictable than interest in influenza.

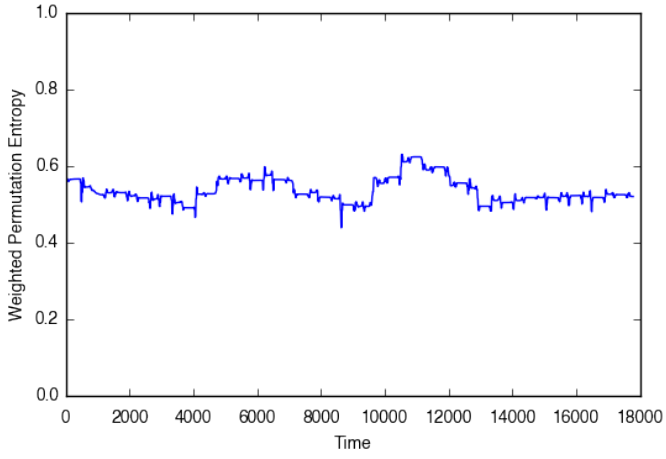

FIG. 5. Weighted permutation entropy for search interest in "full moon"

Before running LMA, we plotted mutual information with respect to the location within the data. The first minimum occurs at $\tau = 14$. Using this τ, we plotted the percentage of FNN as a function of embedding dimension. We chose $m = 5$, where FNN just begins to drop below 10%.

Our prediction is shown in the middle plot of Fig. 8. The MASE is 0.9079, meaning than on average our prediction did slightly better than a random walk.

C. Baseball

Finally, the last data set we considered was search interest in baseball. As you can see in Fig. 6, search interest rises over the summer months.

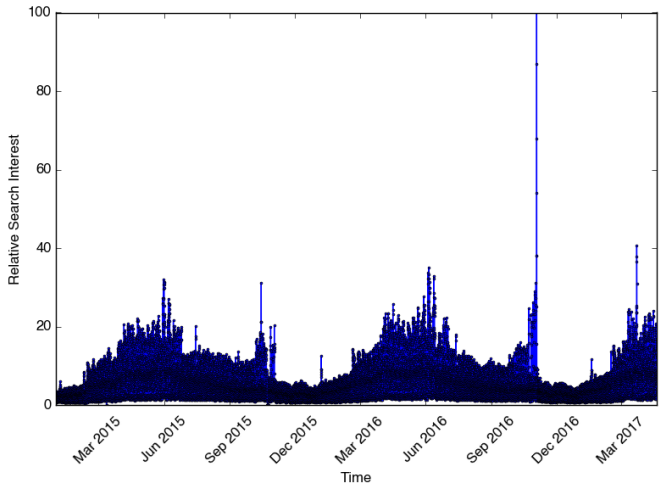

FIG. 6. Hourly search interest in "baseball" from January 1st 2015 to April 20th 2017

From Fig. 7 we can see how WPE varies across the series. The WPE is the highest for all of the trends considered (average of 0.8647), suggesting that baseball should be least predictable.

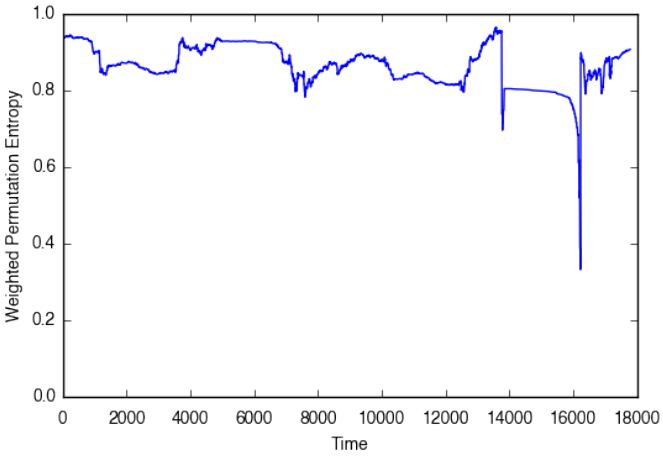

FIG. 7. Weighted permutation entropy for search interest in "baseball"

Again, we first need to embed our data. The delay parameter τ was found by plotting mutual information. The first minimum value occurs for $\tau = 5$. The percentage of FNN drops below 10% for an embedding dimension of $m = 4$.

Our prediction is shown at the bottom of Fig. 8. The MASE is 0.3872, the lowest error of all predictions made. In other words, we performed 2.5826 times better than a random walk forecast.

IV. DISCUSSION

Circling back to our original research question: Are Google search trends predictable? We have only considered three search trends, but it appears that some search trends are predictable. Interest in baseball, even though it has the highest weighted permutation entropy, has the lowest prediction error. Predictions of search interest in full moon also perform relatively well as compared to a random walk. Though, like Google, we could not predict search interest in influenza better than a random walk.

We would expect that low WPE would be correlated with low MASE, but interestingly, this was not observed. There are several reasons why this may have occurred. Firstly, it is worth stating that choosing parameters for calculating WPE and for generating our embeddings is an art. Garland and Bradley [6] showed that topologically inaccurate embeddings can still yield accurate predictions, so perhaps our choice of parameters for WPE is not quite correct. Secondly, it is possible that our data is just too noisy. A choice of larger k within LMA, may help curb error due to noise. In addition, it may be interesting to apply these same methods over less resolved data (daily search interest), and ask whether or not our prediction accuracies improve. Thirdly, the baseball trend has a much smaller range than full moon in relative search interest. Perhaps our measure of prediction accuracy unfairly punishes time series with large variance. Future work should involve examining other measures of accuracy.

This work could be extended in many ways. We use a very simple forecasting method. A quick extension would test accuracy under a multi-step prediction using LMA. Future work should include more search trends. In doing so, perhaps we could better understand what features of search trends lend to better prediction accuracy. It would also be interesting to study how search trends are correlated. Can we predict interest in baseball from interest in the Colorado Rockies, and vice versa?

[1] Christoph Bandt and Bernd Pompe. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.*, 88:174102, Apr 2002.

[2] Declan Butler. When google got flu wrong. *Nature*,

494(7436):155–6, Feb 14 2013.

[3] Bilal Fadlallah, Badong Chen, Andreas Keil, and José Príncipe. Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. *Phys. Rev. E*, 87:022911, Feb 2013.

[4] Joshua Garland. *Prediction in Projection: A new paradigm in delay-coordinate reconstruction.* PhD thesis, University of Colorado, Boulder, 2016.

[5] Joshua Garland and Elizabeth Bradley. Predicting computer performance dynamics. In *International Symposium on Intelligent Data Analysis*, pages 173–184. Springer, 2011.

[6] Joshua Garland and Elizabeth Bradley. Prediction in projection. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(12):123108, 2015.

[7] GeneralMills. pytrends. `https://github.com/GeneralMills/pytrends`, 2013.

[8] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4, Feb 19 2009.

[9] Google. How Google Search Works — Search Algorithms. `https://www.google.com/search/howsearchworks/algorithms/`.

[10] Google. Search using autocomplete — Search Help. `https://support.google.com/websearch/answer/106230?hl=en`.

[11] Google. Google Trends. `https://trends.google.com/`, 2007.

[12] Rainer Hegger, Holger Kantz, and Thomas Schreiber. Practical implementation of nonlinear time series methods: The tisean package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, 1999.

[13] Holger Kantz and Thomas Schreiber. chapter Phase space methods, pages 30–47. Cambridge University Press, Cambridge, 2005.

[14] Edward N Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4):636–646, 1969.

[15] Allison Morgan. google_trends. `https://github.com/allisonmorgan/google_trends`, 2017.

[16] LF Olsen and WM Schaffer. Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics. *Science*, 249(4968):499–504, 1990.

[17] Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

[18] Tim Sauer. Time series prediction: forecasting the future and understanding the past. volume 15., pages 175–193, Reading, MA, 1993. Addison-Wesley Pub. Co.

[19] Floris Takens. Detecting strange attractors in fluid turbulence, 1981.

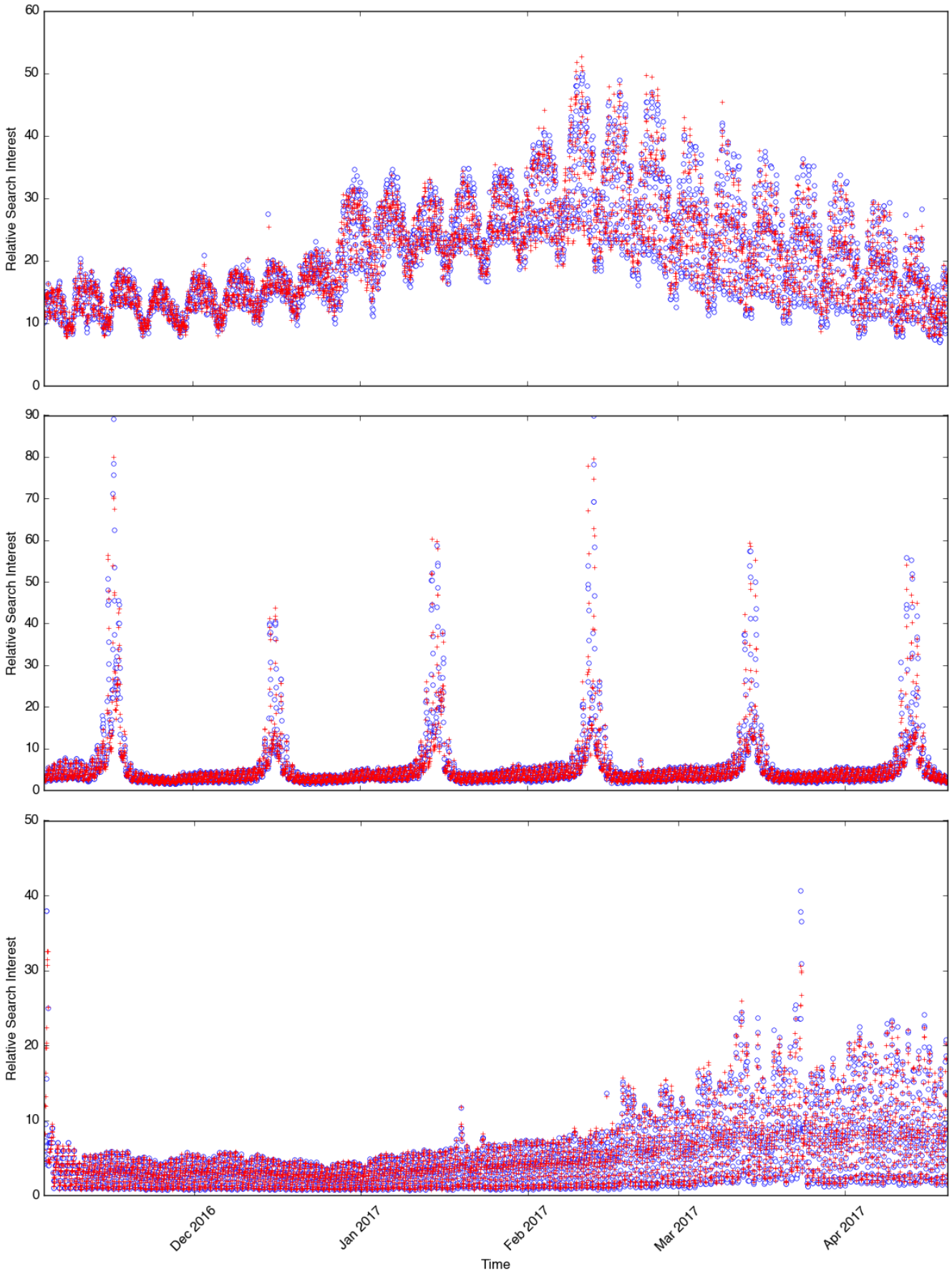

FIG. 8. The true points in our test data (blue circles), and predictions (red pluses) using one-step forecasts of LMA under $k = 5$. From top to bottom, search interest in "influenza," "full moon," and "baseball."