Predicting Fine-Grained Tourism Visitation using Social Media Data

Amir Khatibi, Jussara M. Almeida, Marcos A. Gonçalves
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, MG, Brazil
{amirkm, jussara, mgoncalv}@dcc.ufmg.br

Abstract

The ability to perform accurate predictions about future events is interesting in many areas, one of them being the Tourism Industry. Usually, countries and cities invest a huge amount of money for planning and preparation in order to welcome the incoming tourists. By having an accurate prediction of future visitations changes in the next days or months, the benefits of this profitable industry could be maximized. In most previous works, the forecasting is performed for a whole country and not for fine-grained areas of a country or for touristic places with unavailable visitation census. In this work, we suggest that accessible data in social networks and travel websites can be used to support the inference of visitation changes in any touristic point in the world much quicker and cheaper. To test our hypothesis we analyze visitation, climate and social media data in more than 80 National Parks in U.S during the last 5 years. Furthermore, we train a simple linear regression model and by computing the proportion of social media along climate data versus official visitations, we could predict with relatively high accuracy the actual visitation count for each of these parks in the next months.

Key words: Social Media Data, tourists reviews, tourism demand, climate data, unavailable visitation census, U.S national parks

Introduction

Decision makers of industries like transportation companies, accommodation facilities, hotels and traveling agencies, all would like to have good estimates of the future demand in the weeks, months, seasons and even years regarding the number of incoming tourists to their regions. In such context, the development of models to predict future visitation demand to specific places and regions can be of great benefit. It is important to note that such predictions are not trivial as many factors could interfere in the cyclic and/or trending behavior of visitation counts. For example factors like exchange rate (Webber 2001), epidemics, fuel price, climate changes (Hengyun Li and Li 2016), local and global financial crisis (Maditinos and Vassiliadis 2008) and hit movies (Riley and Van Doren 1992) could cause drastic deviations in tourism demand forecasts if we do not properly weight these elements. However, most of these factors are reflected quickly in social media (Asur and Huberman 2010), (Chunara, Andrews, and Brownstein 2012) due to the huge amount of involved users and vast amount of daily produced content in these websites.

Nevertheless, one of the challenges in tourism prediction is data gathering. Indeed, the official visitation for many touristic points is not well-documented and easily available. Moreover, conducting surveys at entrances of major attractions is expensive and provides only limited spatial and temporal coverage. The situation is worse in the case of developing countries and even more complicated regarding remote touristic sites. This is why in most prior related works (Wang 2004), (Cankurt and Subasi 2015), (Chang Jui Lin 2011), the proposed prediction models for forecasting touristic activities are built and tested over a whole country and not for specific regions or attractions. Here, it is necessary to mention that there are works in the area of attraction recommendation (Borras, Moreno, and Valls 2014) but their main focus is the users/tourists and not the attractions. In other words, this field of the research is focused on recommending to the tourist the proper attractions based on the history of her preferences, her social network and other personal factors while in our problem the main focus is on the visitations to the attractions themselves.

In this context, our main contribution in this work is evaluating the feasibility of exploiting accessible data in social media networks alongside climate data in order to infer the visitation percentage change in attractions. To this aim, first, we choose sites with available monthly visitation census as the ground-truth of our analysis. In this way, more than 80 National Parks in United States of America were chosen. Then we use available reviews by tourists and their ratings in a famous travel website besides monthly average temperature in each site to correlate this information with visitation counts in the corresponding attractions in the same period of time. Next, we train a simple prediction model using a linear regression approach; the model is trained for each attraction separately using the last 5 years of social media and climate data. Finally, by computing the proportion of social media reviews along climate data average temperature versus official visitation counts, we could predict with relatively high accuracy the actual visitation count for each of these parks in the next months. Such results are very important, mainly for attractions with "difficult to gather" official visitation census data.

The rest of this paper is as follows. We first present related work. Next, the datasets used in our analysis, the data cleaning phase and the metrics we used in our study are discussed. Following we detail our experimental characterization and exploit the results to forecast the touristic demands using a simple linear regression model. Finally we conclude the paper with glimpses in future work.

Related Work

There are plenty of works using Location Based Social Networks (LBSN) data such as Foursquare and Yelp to study the mobility behavior of tourists and citizens (Li and Chen 2009), (Cho, Myers, and Leskovec 2011), (Hasan, Zhan, and Ukkusuri 2013) and (Hossain et al. 2016). On the other hand, there is only a few works analyzing the social media data such as tourists' reviews, ratings and check-ins to infer the visitation density over time. Moreover, those that use this information to make an estimation of future touristic demands, only do it in a coarse-grained fashion (e.g. country or city). There is a few exceptions though. For instance, in (Spencer A. Wood and Lacayo 2013), the authors use the locations of photographs in Yahoo Flickr social media website to estimate visitation rates in some recreational sites around the world; they use information from the profiles of the photographers to derive travelers' origins in order to compare their estimations to empirical data and conclude that the crowdsourced information can indeed serve as a reliable proxy for empirical visitation rates.

In (Nicholas A. Fisichelli 2015), the authors analyze the climate and visitation data for U.S. national parks using a third-order polynomial temperature model and argue that it explains 69% of the variation in historical visitation trends. Albeit their interesting results, we show in this work that by exploiting the social media data, a higher accuracy of 83% can be achieved.

The work described in (Hengyun Li and Li 2016) links the climate and seasonal tourism demand to study the effects of home climate, destination climate, and climate differences between destinations and hometown on touristic demands. In their study, other features of a destination include access to sea or lakes, availability of cultural and historical places, price, hospitality, accommodations, ease of access and cuisine are also considered. Anyhow, statically, they show that the features: home climate, destination climate, and climate difference count for most of the Hong Kongers' tourism demand considering 19 major tourism cities in Mainland China.

In this context, in here we provide a more thorough study by analyzing the climate history of each attraction along with social media data. To the best of our knowledge, we are the first to perform such joint analysis. By mixing the climate and social media data in order to predict touristic demands we could produce better forecasting models especially for touristic places with no availability of visitation census due to various reasons such as being costly surveys, difficulty to access remote places to collect data and so on.

In (Cankurt and Subasi 2015), the authors use multilayer perceptron (MLP) regression and support vector regression (SVR) models in order to make multivariate tourism forecasting for Turkey. The authors use features such as: wholesale prices index, US Dollar selling, hotel bed capacity of turkey, number of tourism agency in Turkey. Using robust models like SVR and MLP let them gain a high accuracy prediction results. However, they again work of the coarse-level of country. And we show that simpler linear models, which are easy to train and to interpret, can produce relatively accurate predictions.

Data Analysis

In our study, we used various datasets collected from different sources in order to produce more robust results. In this Section, we first present our dataset sources with some details, and then we present the metrics we utilize in our characterization and prediction experiments.

Dataset

The social media data in our experiments are collected from the famous world-wide social network - Trip Advisor which is the largest travel website for more than 11,000 reviews per day and 315 million mobile application users[1]. We collected the monthly number of reviews along average rating scores of reviewers during the period of January 2011 till September 2016 [2].

In addition, to improve our analyses, climate data including the monthly minimum, maximum and average temperature aside with the monthly precipitation of all the 83 National Parks in the period of January 2000 to November 2016, were collected from the U.S National Climate Data Center [3].

Finally, we obtained the official visitation statistics from the U.S National Park Service website. In their portal, the monthly touristic demands for National Parks in U.S. is provided. We downloaded the monthly total number of visitors in each national park in the period of January 1996 to February 2016 for 83 parks in U.S [4] to use as the ground truth dataset for our study. In Table 1, the list of datasets is presented with details.

Data Cleaning

For our experiments, we chose 124 National Parks in U.S. with both, available Social Media data and monthly Official Visitation census. In a further analysis, we discarded some parks with very few reviews in the Social Media. The reason for this cut is that low number of reviews for an attraction in a long period of time indirectly shows the few contributions of the community to the social media page of this park in our specified travel website. Such contribution are key for our study. As a result, we filtered out all parks with less than

[1]Based on TripAdvisor's fact sheet available at $http://www.tripadvisor.com/PressCenter-c4-FactSheet.html$

[2]available at $https://www.tripadvisor.com/Attractions-g191-Activities-c57-t67-United_states.html$

[3]available at $https://www.ncdc.noaa.gov/cag/time-series/us/$

[4]available at $https://irma.nps.gov/Stats/$

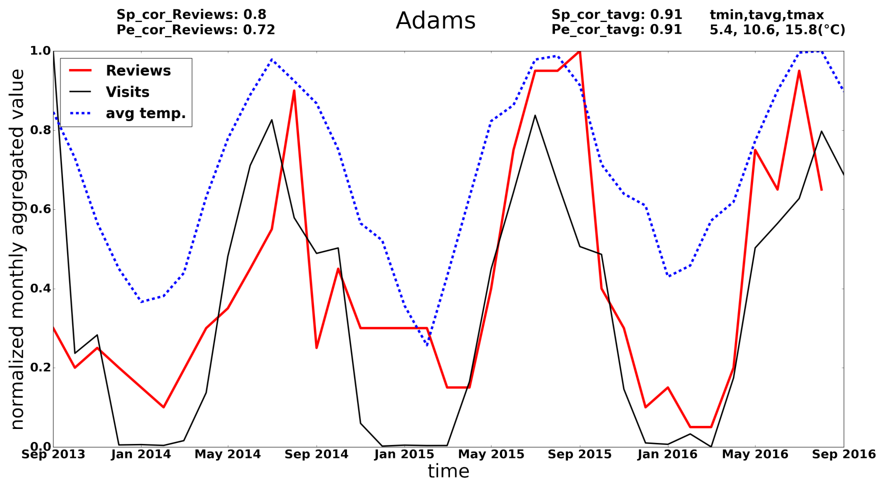

Figure 1: Correlation results for Adams National park in Massachusetts, US. Meaning of the abbreviations on top of the Figure: Sp= Spearman; Pe=Pearson; tavg= average temperature. All correlations are related to the official number of visits from the ground truth.

Table 1: Datasets

Dataset	provider	granularity	attributes	Data Range
VIS	U.S National Park Service	monthly	total number of visitors	1996-01 to 2016-08
CLM	U.S National Climate Data Center	monthly	min,avg,max temperature, avg precipitation	2000-01 to 2016-10
SOC	Trip Advisor travel website	monthly	No. reviews, avg ratings	2011-01 to 2016-09

300 reviews in the last 5 years (an average of 60 per year). After the data cleaning process, we remained with 83 National Parks.

Correlation Coefficients

In the next section, we correlate the three collected datasets. For correlation evaluation of time-series, there is a set of metrics like Spearman (Bonett and Wright 2000), Pearson (Gautheir 2001), Kendell (Yue, Pilon, and Cavadias 2002) and DTW (DJ Berndt 1994). The most widely used type of correlation coefficient is Pearson. It assumes that the two variables being analyzed are measured on at least interval scales, meaning they are measured on a range of increasing values. The coefficient is calculated by taking the covariance of the two variables and dividing it by the product of their standard deviations. However, the Spearman correlation is more robust than the Pearson coefficient in correlating time-series(Bonett and Wright 2000). Pearson measures linear dependence whereas Spearman measures are invariant by monotonous transforms of the variables.

In more details, if we have one dataset $\{x_1, x_2, ..., x_n\}$ containing n values and another dataset $\{y_1, y_2, ..., y_n\}$ containing n values then the formula for Pearson correlation is:

$$P = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where $\bar{x}$ is the mean of x_i and analogously for $\bar{y}$.

In order to calculate Spearman's rank correlation, first we should rank the observations in the two samples separately from smallest to largest. Equal observations are assigned the mean rank for their positions. Let u_i be the rank of the i-th observation in the first sample and v_i be the rank of the i-th observation in the second sample. Spearman correlation coefficient is a measure of the correlation between ranks, calculated by using the ranks in place of the actual observations in the formula for calculating Pearson correlation coefficient. Equation 2 presents the formula for Spearman; this equation can be approximated as 3 which is not exact when there are tied measurements but the approximation is good when the number of ties is small in comparison to n.

$$S = \frac{n\sum_{i=1}^{n}(u_i v_i) - (\sum_{i=1}^{n} u_i)(\sum_{i=1}^{n} v_i)}{\sqrt{[n\sum_{i=1}^{n} u_i^2 - (\sum_{i=1}^{n} u_i)^2][n\sum_{i=1}^{n} v_i^2 - (\sum_{i=1}^{n} v_i)^2]}} \tag{2}$$

$$S = 1 - \frac{6\sum_{i=1}^{n}(u_i - v_i)^2}{n(n^2 - 1)} \tag{3}$$

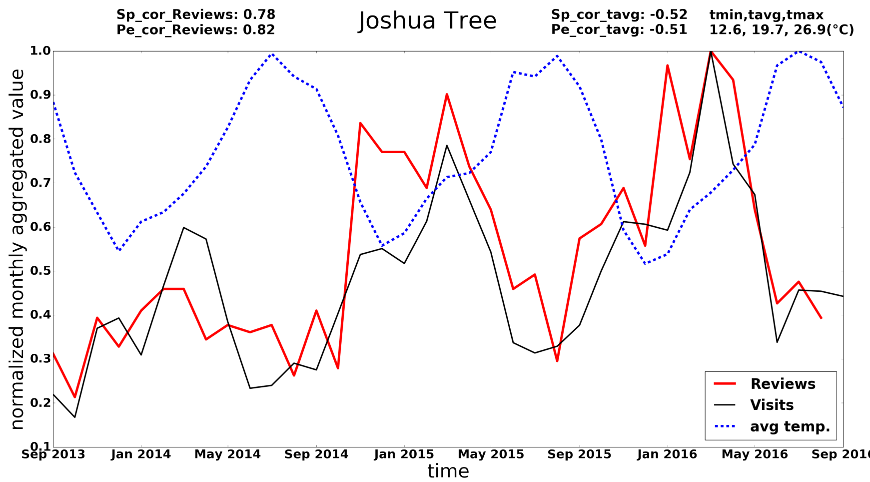

Figure 2: Correlation results for Joshua Tree National park in California, US. Meaning of the abbreviations on top of the Figure: Sp= Spearman; Pe=Pearson; tavg= average temperature. All correlations are related to the official number of visits from the ground truth.

In our experimental results, we report on both Spearman and Pearson correlation values.

Experimental Characterization

This section includes the results of our experimental characterization that contrasts the touristic demands versus social media and climate data. After normalizing the official visitation census data, average temperature data and the social media reviews and ratings by the maximum value of each time-series, we correlated the correspondent monthly data values using both Pearson and Spearman correlation coefficients in the period of January 2011 till September 2016.

Our results show that over 83% of the parks show a moderate to high correlation of more than 50% between the monthly total number of reviews and the monthly total number of visitors for each National Park. In the following, we categorize the correlation results into three categories: (A) the parks with high correlation within their social media and official visitation data (more than 65%); (B) those parks with moderate correlation within the social media reviews and the number of visits (between 50% till 65%) ; (C) parks with low correlation (less than 50%) within their social media and official visitations but moderate correlation within the climate data and the official visits (over 50%); and (D) Parks with low correlation (less than 50%) within their social media, climate and official visitation data.

Overall, 54 of the 83 considered parks (65%) were classified in category A, 19 in category B, 6 in category C and only 4 in category D. This shows the high potential of using both types of data - social media and climate - simultaneously in prediction tasks.

To illustrate these results, Figures 1 and 2 present the correlation graphs of the National parks Adams in Massachusetts and Joshua Tree in California in United States of America. As it can be seen in the (high) correlation values above the figures, correlation category of these parks belong to the category (A). This is graphically illustrated by the similarity patterns in the three temporal series.

Figure 3, shows the correlation results for the Big Cypress National park in southern Florida, U.S. with a correlation category (B).

As mentioned, in a few parks (4), correlation of social media and touristic demand census was not high. Figure 4 – Cabrilio National Monument in California, U.S. – illustrates of cases in category (D).

In order to better understand the reasons for this phenomenon, i.e, the low correlation between the number of reviews versus the number of visits in a few parks, we decided to analyze their average monthly temperature. To do so, we plotted the average monthly minimum, average and maximum temperature of all parks alongside the value of the correlation between the social media and official data. In Figure 5, temperature effects on correlation of social media and official data is presented using a scatter plot. In the fourth column and the fourth row, the range of x and y values is between (0,1) which represents the correlation value while in the rest of rows and columns, x and y values shows the temperature in Celsius degree.

As it can be inferred from this figure, the social media is more representative of the real visitations when the average temperature of the park is moderate. In other words, for the parks with a high temperature climate (whether min, avg or max temperature), the correlation value decreases. The blue circles in Figure 5 shows this behavior in the scatter plot. It could be seen in this figure that there is a linear correlation between the min, max and average monthly temperature in all these parks so we could use any of these temperature-based variables in our correlation analysis.

We turn our attention now to the predictive capability of the features and some seasonal aspects we are analyzing. Figure 6 represents the potential effectiveness of predicting

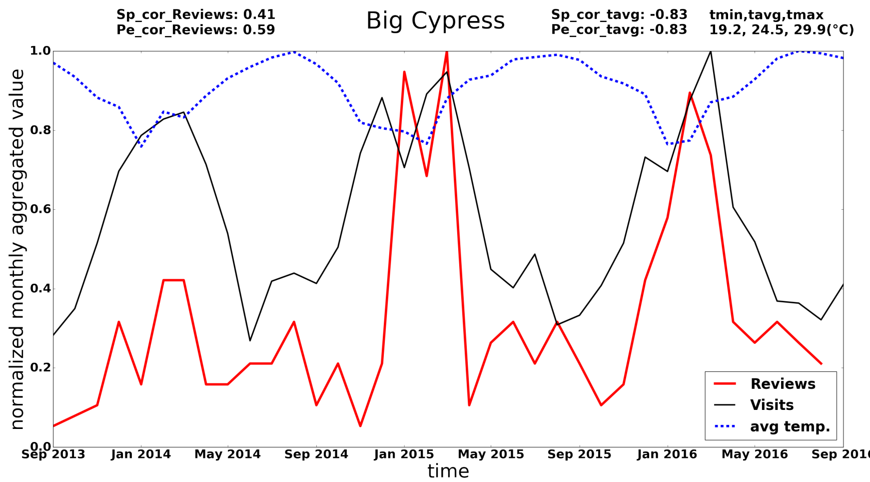

Figure 3: Correlation results for Big Cypress National park in southern Florida, US. Meaning of the abbreviations on top of the Figure: Sp= Spearman; Pe=Pearson; tavg= average temperature. All correlations are related to the official number of visits from the ground truth.

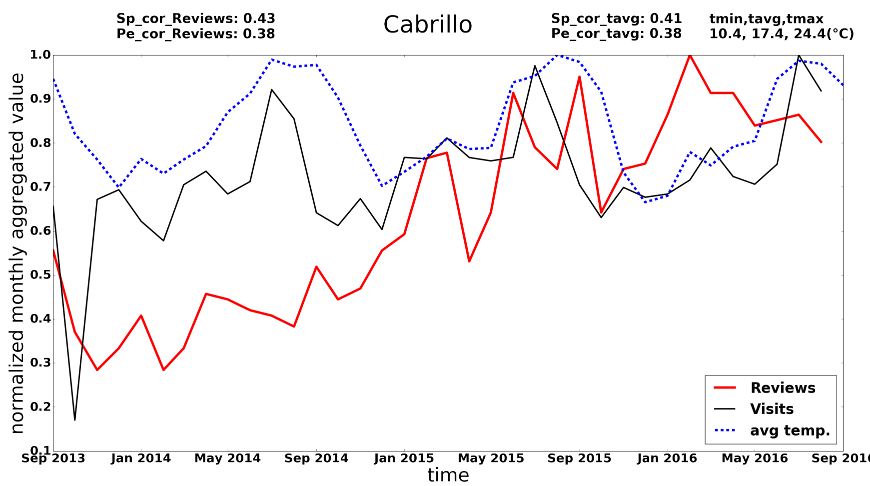

Figure 4: Correlation results for Cabrilio National Monument in California, US. Meaning of the abbreviations on top of the Figure: Sp= Spearman; Pe=Pearson; tavg= average temperature. All correlations are related to the official number of visits from the ground truth.

touristic demands in different seasons of the year. It can be seen that in some seasons, the monthly number of reviews in social media is more representative of the total number of visits. For example in some parks we have a higher correlation value in summer or winter.

In Figure 7, we plot the visitation forecasting potential, based on the calculated correlations, of the different features. Note that for each feature, the values are sorted in an ascending order to provide a better view of their performance in comparison with the other features.

Considering the features 'Reviews_Spring', 'Reviews_Summer', 'Reviews_Autumn', 'Reviews_Winter', 'all_Reviews', 'avg_temperature' and 'Precipitation', the avg_temperature is the feature that produces the best correlation with the official visits. However, by itself , it cannot be considered a good estimator as there are many cases in which the correlation value of average temperature is very low. In other words, when this feature is well correlated, the value is very low. The problem is that we do not know for which parks this could happen. On the other hand, we can see from the Figure and previous analysis that the feature 'social media reviews' is more stable and reliable. The reason is that when this feature is not well correlated with the official visitation, it still represents more than 30% of the behavior of the touristic demand over the time. And when the value is high, it is still very close to the best representative feature in the same park.

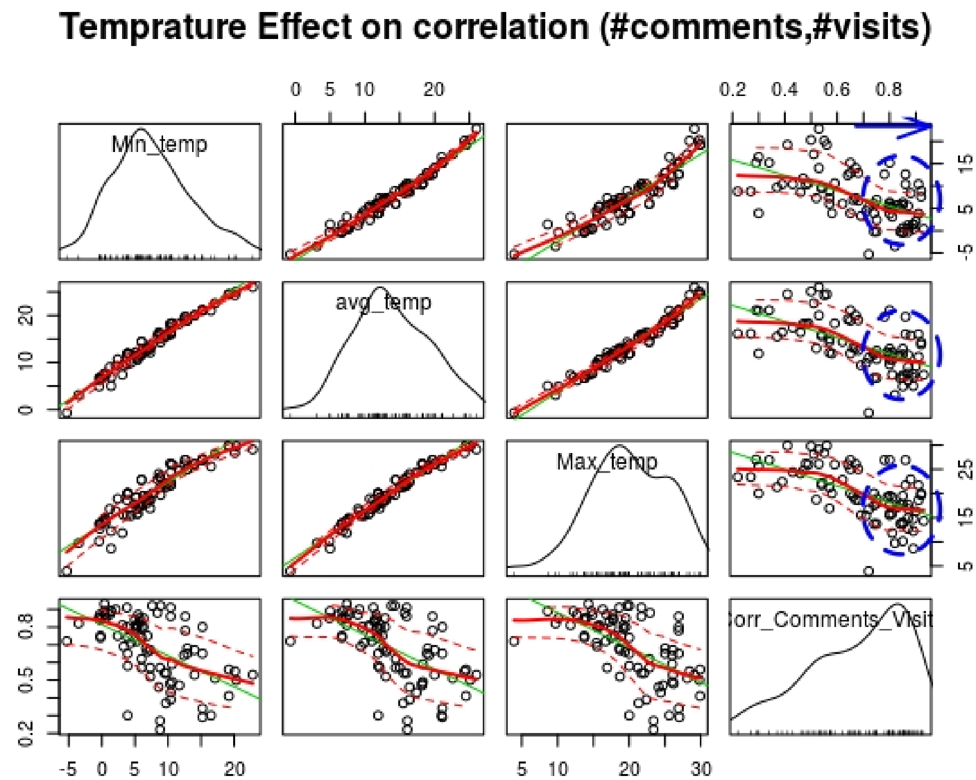

Figure 5: Temperature effect on correlation of social media and official data

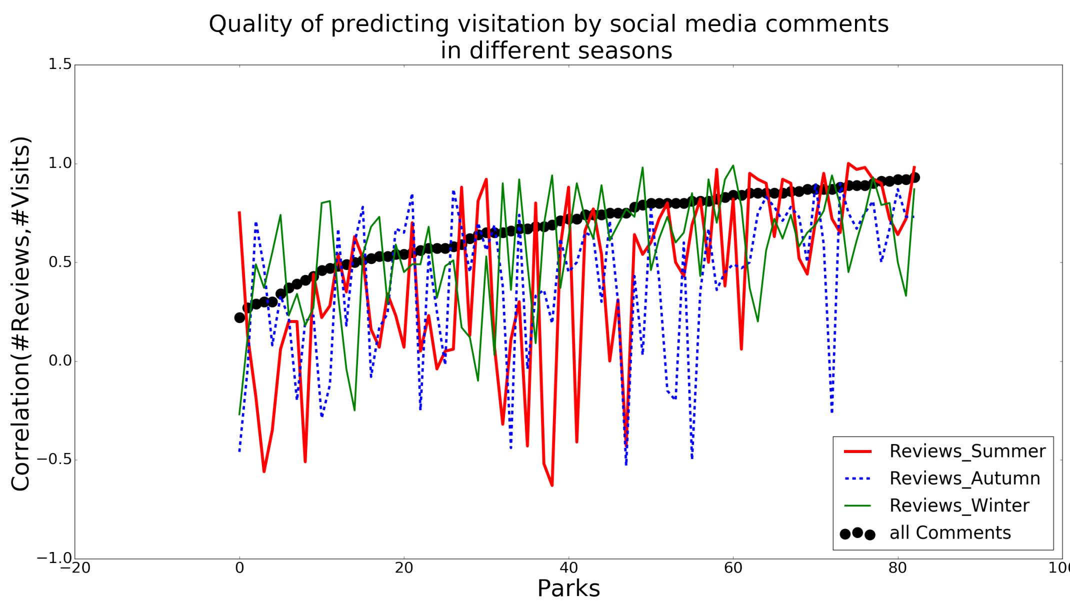

Figure 6: Predicting visitation in different seasons of the year

Prediction

For time-series forecasting, we should first check the main trend of the time-series; whether it is decreasing, increasing or constant over time. Then, we should figure out the cyclic behavior of the time-series or, in other words, its seasonal-

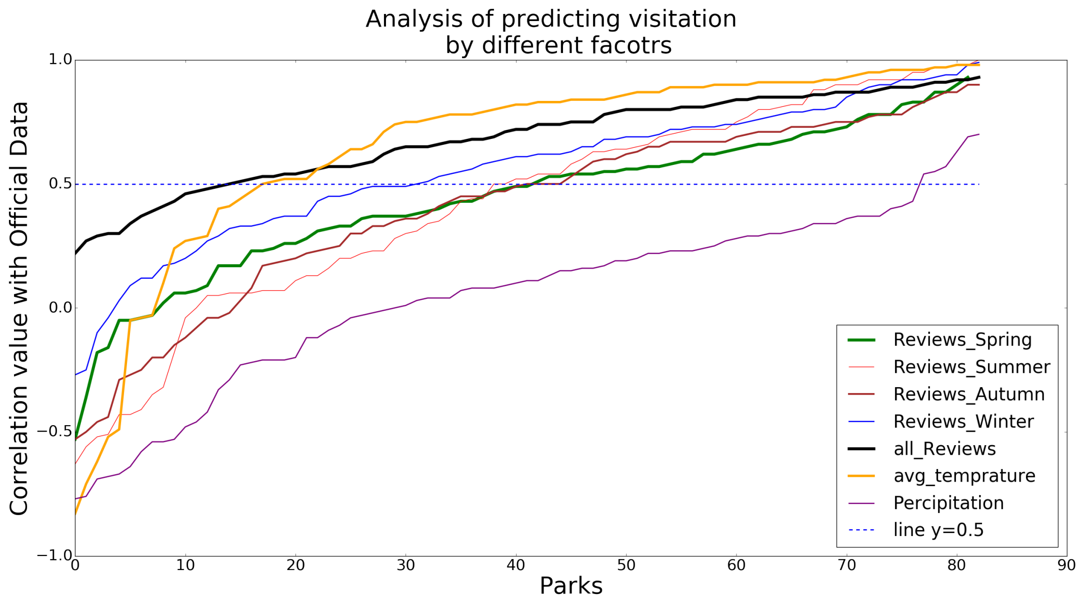

Figure 7: Visitation forecasting quality by different factors

ity. Next, in order to join these two components and summarize the original time-series, a random (residual) component should be summed up with the former two components. This is the component which causes the prediction of time-series to always be uncertain and the necessity of having a confidence interval. The random component is a kind of noise which should be taken apart from the main trend and seasonality of the time-series at the time of forecasting. There are many stochastic, adaptive and regression models to use on time-series forecasting such as Support Vector Regression (SVR) and artificial neural network (Cankurt and Subasi 2015), autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) (Wei 1994), Holt-Winters (Kalekar 2004) and Hidden Markov Models(HMM) (Wesley Mathew 2012). Nevertheless, in this work we start with a simple prediction model, i.e. a linear regression approach (Montgomery, Peck, and Vining 2015) to show the potential of using social media data along with climate data to forecast the touristic demand in the following months. In our view, this simple solution will better emphasize the predictive capability of the features, as the model is simpler and more explainable, besides being faster to train. Moreover, recent work has showed that correlations are high, linear models have been shown to be quite effective in other popularity prediction tasks (Vasconcelos, Almeida, and Gonçalves 2015), (Pinto, Almeida, and Gonçalves 2013), (Szabo and Huberman 2010). Anyway, we leave the use of more complex prediction models for future work.

.

For prediction accuracy measurement, we use relative error; assuming that the true value of a quantity be x and the measured value x_0. Then the relative error is defined by equation 4. In this equation, Δx is the absolute error while the percentage error is relative error multiplied by 100.

$$relative\ error = \frac{\Delta x}{x} = \frac{x_0 - x}{x} \tag{4}$$

To obtain the prediction results in the dataset of National Parks, we trained a linear regression model with the last 5 years of the social media, climate and official visitation data for each park, during the period of January 2011 till January 2016. Then we used each model to forecast the number of visits in the next 1, 3 and 6 months in the period of January to August 2016 using the monthly quantity of reviews and average temperature in the same period (January to August) but from the previous year (2015); Equation 5 presents the linear regression formula where C_1 and C_2 are linear regression coefficients have been trained separately for each park. In Table 2 we report the learned coefficients. Next we compared the result with the available official visitation to calculate the relative error of each model. Figure 8 shows the relative error in a Cumulative distribution function (CDF) plot (Anderson and Darling 1954).

$$\begin{aligned} Visitation\ Count = (C_1 * No\ Reviews) + \\ (C_2 * Average\ Temperature) \end{aligned} \tag{5}$$

Our experiments show that the results are actually very good. For one fourth of the parks, the percentage error is quite small (below 10%). In fact, for the majority of the sites (60%), the relative error is under 30%, which is quite reasonable. And for jsut a small percentage of the sites (10%), the error can be quite large (above 100%) which is mostly the parks with low correlation between their social, climate and official data,. i.e., categories C and D.

Conclusion

In this paper, we analyzed use of Social Media data as a way for tourism demand forecasting for places and attrac-

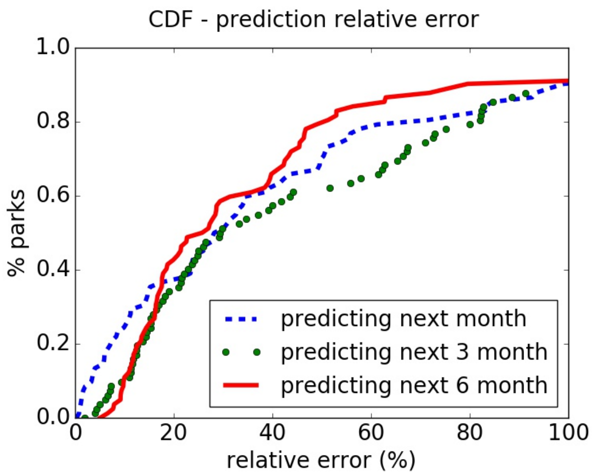

Figure 8: prediction relative error - Cumulative distribution function

tions with unavailable official visitation census. We took advantage of correlating various datasets gathered from different sources due to evaluating our hypothesis of forecasting tourism demands using the Social Media reviews and ratings along with climate data. A dataset of official visitation in more than 80 National Parks in U.S. has been used as the ground truth basis for our analysis.

In the future works, we aim to improve the accuracy of the prediction results using more robust prediction models like SVM and ANN, in addition to clustering the attractions based on their locations, climate and the intervals of their monthly average number of visitation. In addition, we would like to study Location-based Social Networks as another way to collect the information of check-ins of the tourists in different attractions.

References

Anderson, T. W., and Darling, D. A. 1954. A test of goodness of fit. *Journal of the American statistical association* 49(268):765–769.

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, 492–499. IEEE.

Bonett, D. G., and Wright, T. A. 2000. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika* 65(1):23–28.

Borras, J.; Moreno, A.; and Valls, A. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications* 41(16):7370–7389.

Cankurt, S., and Subasi, A. 2015. Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components. *Balkan Journal of Eletrical and Computer Engineering* Vol.3, No.1.

Chang Jui Lin, Hsueh Fang Chen, T. S. L. 2011. Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: Evidence from taiwan. *International Journal of Business Administration* 2.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1082–1090. ACM.

Chunara, R.; Andrews, J. R.; and Brownstein, J. S. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86(1):39–45.

DJ Berndt, J. C. 1994. Using dynamic time warping to find patterns in time series. *KDD workshop* 10:359–370.

Gautheir, T. D. 2001. Detecting trends using spearman's rank correlation coefficient. *Environmental forensics* 2(4):359–362.

Hasan, S.; Zhan, X.; and Ukkusuri, S. V. 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, 6. ACM.

Hengyun Li, H. S., and Li, L. 2016. A dynamic panel data analysis of climate and tourism demand: Additional evidence. *Journal of Travel Research* I-14.

Hossain, N.; Hu, T.; Feizi, R.; White, A. M.; Luo, J.; and Kautz, H. 2016. Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities. *arXiv preprint arXiv:1603.03181*.

Kalekar, P. S. 2004. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology*.

Li, N., and Chen, G. 2009. Analysis of a location-based social network. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, 263–270. Ieee.

Maditinos, Z., and Vassiliadis, C. 2008. Crises and disasters in tourism industry: happen locally, affect globally. In *Management of International Business and Economics Systems, MIBES Conference*, 67–76.

Montgomery, D. C.; Peck, E. A.; and Vining, G. G. 2015. *Introduction to linear regression analysis*. John Wiley & Sons.

Nicholas A. Fisichelli, Gregor W. Schuurman, W. B. M. P. S. Z. 2015. Protected area tourism in a changing climate: Will visitation at us national parks warm up or overheat? *PLoS ONE* 10(6).

Pinto, H.; Almeida, J. M.; and Gonçalves, M. A. 2013. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 365–374. ACM.

Riley, R. W., and Van Doren, C. S. 1992. Movies as tourism promotion: A pull factor in a push location. *Tourism management* 13(3):267–274.

Spencer A. Wood, Anne D. Guerry, J. M. S., and Lacayo, M. 2013. Using social media to quantify nature-based tourism and recreation. *Scientific Report* 3.

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM* 53(8):80–88.

Vasconcelos, M.; Almeida, J. M.; and Gonçalves, M. A. 2015. Predicting the popularity of micro-reviews: A foursquare case study. *Information Sciences* 325:355–374.

Wang, C.-H. 2004. Predicting tourism demand using fuzzy time series and hybrid grey theory. *Tourism Management-Elsevier* 25:367–374.

Webber, A. G. 2001. Exchange rate volatility and cointegration in tourism demand. *Journal of Travel research* 39(4):398–405.

Wei, W. W. S. 1994. *Time Series Analysis Univariate and Multivariate Methods*, volume 2. Pearson - Addison wesley.

Wesley Mathew, Ruben Raposo, B. M. 2012. Predicting future locations with hidden markov models. *ACm UbiComp 2012*.

Yue, S.; Pilon, P.; and Cavadias, G. 2002. Power of the mann–kendall and spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of hydrology* 259(1):254–271.

Table 2: Linear regression model -learned coefficients

Park	C_1	C_2	Park	C_1	C_2
Adams	-0.11622161	1.21801019	Hovenweep	0.22203335	0.61022916
Antietam	0.19862209	0.75558285	Independence	0.3005342	0.58650049
Assateague Island	0.61370474	0.94431147	Jean Lafitte	0.11838278	0.20356979
Aztec Ruins	0.27601232	0.92056559	Joshua Tree	0.29362107	-0.45262861
Bandelier	0.38373783	0.49929536	Kenai Fjords	0.64419066	0.5104998
Big Bend	0.56907618	-0.30224516	Kennesaw Mountain	0.0905915	1.08612672
Big Cypress	0.24665886	-1.95850352	Kings Canyon	0.17301444	1.29675599
Black Canyon	0.21464489	0.86746823	Lake Mead	0.14387742	1.0257716
Bryce Canyon	0.27027008	0.80103843	Lincoln Home	-0.09781215	0.98047734
Cabrillo	0.23920407	0.36264714	Little Bighorn Battlefield	0.36546409	0.86730278
Canaveral	0.28359523	0.82084556	Lyndon B. Johnson	0.69984867	-0.53071302
Cape Cod	0.17761348	1.04776102	Mammoth Cave	1.20638814	0.23701912
Cape Hatteras	0.13125016	1.42497049	Minute Man	0.07082508	0.99523212
Capitol Reef	0.38626972	0.77473519	Montezuma Castle	0.40192158	0.08963126
Cedar Breaks	0.64278239	0.37511449	Mount Rainier	0.48845374	0.92329412
Chiricahua	0.51042524	-0.48652055	Mount Rushmore	0.50542264	0.78962759
Colorado	0.20680582	0.62163805	Muir Woods	0.2820518	0.86121826
Crater Lake	0.26179579	1.00422405	Natural Bridges	0.34613147	0.58719186
Craters of the Moon	0.40812067	0.95581679	Ocmulgee	0.13853699	0.57454598
Cumberland Gap	0.12883619	0.85733041	Padre Island	0.66758626	0.76485867
Cumberland Island	0.35278411	0.21325232	Petroglyph	0.21075052	0.46272063
Cuyahoga Valley	0.06215396	0.8911931	Pictured Rocks	0.68378877	0.34962079
Denali	0.89282823	0.62369615	Pinnacles	-0.07526282	-0.17425336
Devils Postpile	0.8269298	0.73888461	Redwood	0.35034759	0.93102505
Devils Tower	0.44739327	0.86638349	San Antonio Missions	0.50098498	0.07445492
Dinosaur	0.04909874	1.17514854	San Francisco Maritime	0.01187019	1.0627824
Dry Tortugas	0.51548118	-0.21612369	Shiloh	-0.19419537	0.53664466
Everglades	0.29402736	-1.01963941	Sleeping Bear Dunes	0.93742206	0.22356636
Ford's Theatre	0.54452506	0.08389264	Statue of Liberty	0.5698896	0.79048766
Fort Davis	0.37197309	0.22935437	Stones River	0.10254982	0.68966152
Fort Matanzas	0.37889505	0.6953018	Sunset Crater Volcano	0.10078884	1.02448171
Fort McHenry	0.04109759	0.94529083	Theodore Roosevelt NP	0.35434726	0.68654889
Fort Pulaski	0.21922235	0.6221173	Valley Forge	0.16304726	0.68205947
Fort Smith	0.06697106	0.52004599	Vanderbilt Mansion	0.2625722	0.93153881
Fort Sumter	0.41360108	0.85800944	Vicksburg	0.20335343	0.74454298
Gettysburg	-0.07082087	1.18073001	Walnut Canyon	0.02534873	0.53758153
Glen Canyon	0.19205586	1.29033358	Washington Monument	0.93855068	0.33214409
Grand Teton	0.33968658	0.79983548	White Sands	0.34149341	0.47117086
Haleakala	0.23038558	-0.35894535	Wolf Trap	0.57417184	0.91170958
Harpers Ferry	0.14044576	1.08966384	Wright Brothers	0.09928977	1.52863421
Hawaii Volcanoes	0.23357509	-0.82910556	Wupatki	0.11606265	1.3231079
Hot Springs	-0.02703018	0.88043339			