Proceedings of the Twenty-Eighth AAAI Conference on Innovative Applications (IAAI-16)

Wikipedia in the Tourism Industry: Forecasting Demand and Modeling Usage Behavior

Pejman Khadivi
Discovery Analytics Center
Computer Science Department
Virginia Tech, Blacksburg, Virginia
Email: pejman@cs.vt.edu

Naren Ramakrishnan
Discovery Analytics Center
Computer Science Department
Virginia Tech, Arlington, Virginia
Email: naren@cs.vt.edu

Abstract

Due to the economic and social impacts of tourism, both private and public sectors are interested in precisely forecasting the tourism demand volume in a timely manner. With recent advances in social networks, more people use online resources to plan their future trips. In this paper we explore the application of Wikipedia usage trends (WUTs) in tourism analysis. We propose a framework that deploys WUTs for forecasting the tourism demand of Hawaii. We also propose a data-driven approach, using WUTs, to estimate the behavior of tourists when they plan their trips.

1 Introduction

Tourism is considered as one of the most profitable industries around the world and, hence, tourism demand forecasting is important in various domains such as business planning and assessing economic activity in a region. Various prediction techniques have been used in the literature for demand forecasting (Song and Li 2008). Recently, there have been different studies that suggest that the use of search engine datasets, such as Google Search Trend data can help improve forecasting accuracy (Bangwayo-Skeete and Skeete 2015; Yang and others 2015).

In recent times, online browsing activity has been identified as a precursor to planning travel and visits to far flung destinations. Analyzing such online activity can provide detailed information about places of interest, volume of interest, and spikes/troughs in attention. Our hypothesis in this paper is that, by looking at Wikipedia usage trends (WUT), we may be able to improve the accuracy of tourism demand forecasts.

Wikipedia usage trends have been previously used for forecasting applications in (Hickmann and others 2015) to forecast the Influenza season. Pertaining to the tourism industry, Wikipedia has been used previously for sightseeing recommendation and narrative generation purposes (Fang and others 2015; Hecht and others 2007). However, to the best of our knowledge, there has been no research on using WUTs for tourism demand forecasting. Beside demand forecasting, another important tourism-related issue is to extract the behavior of people when they make decisions about their trips, e.g., how early he/she reads about the National Mall or plans to book his/her flight. The answers to such questions are useful for resource allocation and marketing purposes.

In this paper, we aim to use Wikipedia usage trends for the purpose of tourism analysis. The main contributions of this paper are twofold. First, we propose forecasting algorithms to deploy WUTs and Google Search Trends for tourism demand forecasting. Second, we propose a framework for the extraction of tourists' behavior in trip planning. We performed various experiments on a Hawaii tourism dataset. Results show that WUTs can improve the accuracy of demand forecasting. Furthermore, the extracted tourist behaviors are consistent with the results of the surveys conducted by tourism departments. Thus, our contributions are:

- Developing a regression approach that deploys WUTs to improve the accuracy of tourism demand forecasting.
- Proposing a novel framework for tourist behavior extraction from WUTs and tourism demand time series.

2 Related Work

In order to forecast tourism demand, different univariate and multivariate regression methods have been deployed in the literature. Song and Witt (2006) used vector autoregressive (VAR) model to forecast tourist arrivals to Macau. In addition to a tourist arrival time series, they also incorporated the costs of living in Macau and the originating country. Gunter and Onder (2015) compared the performance of various forecasting models, such as VAR and ARMA, to predict international tourism demand in Paris. Athanasopoulos and de Silva (2012) deployed multivariate exponential smoothing methods for the forecasting of tourism demand for Australia and New Zealand. Chan and others (2005) used various GARCH models to address variations in tourism demand caused by economical instabilities. The use of gravity models for tourism demand analysis has been addressed in (Morley and others 2014). Song and Li (2008) provide a detailed survey on various forecasting techniques.

A number of tourism demand forecasting approaches are based on usage trend data that are supplied by search engines. Xiang and Pan (2011) performed behavior analysis to understand how tourists use search engines for travel planning. They determined which keywords are more important when tourists use search engines. Choi and Varian (2012) used Google search trends (GST) to forecast the number of

monthly visitors of Hong Kong. They use data from the first two weeks of a month to forecast the total number of visitors in the whole month. In (Bangwayo-Skeete and Skeete 2015), GST data about *hotels* and *flights* search terms for popular tourist destinations in the Caribbean are used for forecasting tourism demand. Yang and others (2015) use GST and Baidu trends to predict tourism demand in the Hainan Province of China. In (Artola and others 2015), GST is used to improve the forecasting of tourism demand for Spain.

3 Problem Formulation

In this paper, we use three sets of time series: tourism demand, WUTs, and GSTs. We assume that we have a set of P Wiki-pages related to a specific tourism destination, $\mathbb{W} = \{W_1, \cdots, W_P\}$. The WUT of a specific page, W_i, is a time series denoted by $\mathcal{W}^i = w_1^i, \cdots, w_t^i$, where w_t^i is the total number of times that W_i has been used at time t. We also assume that we have a set of Q Google search terms, $\mathbb{G} = \{G_1, \cdots, G_Q\}$ and that the GST data corresponding to G_j is represented by $\mathcal{G}^j = g_1^j, \cdots, g_t^j$. The tourism demand time series is illustrated by $\mathcal{Y} = y_1, \cdots, y_t$ where y_t is the total number of monthly visitors at time t. In a similar way, $\mathcal{Y}^D$ and $\mathcal{Y}^I$ are the time series that represent the total number of domestic and international visitors, respectively. For a specific time series $\mathcal{X}$, we denote the sequence of samples between time points 1 and t_0 by $\mathcal{X}_{t_0}$, i.e. $\mathcal{X}_{t_0} = [x_1 \cdots x_{t_0}]^T$. In the following subsections we introduce two problems that are studied in the rest of the paper.

3.1 The Demand Forecasting Problem

The demand forecasting problem is to forecast the volume of tourism demand (i.e. tourist arrival) in the future. In the most general setting, forecasting the tourism demand at time t based on time series $\mathcal{D}$ which is available uptil time $t-k$, i.e. $\mathcal{D}_{t-k} = d_1, \cdots, d_{t-k}$, can be formulated as follows:

$$\hat{y}_t = \mathcal{F}\left(\mathcal{D}_{t-k}\right) \tag{1}$$

where $\mathcal{F}$ is the forecasting function and $\mathcal{D}$ can be any univariate or multivariate time series. As an example, if we aim to forecast tourism demand based on the observed values of $\mathcal{Y}$ we will have $\hat{y}_t = \mathcal{F}\left(\mathcal{Y}_{t-k}\right)$.

The problem of tourism demand forecasting using WUT of a Wiki-page, W_i, is formulated in the following equation:

$$\hat{y}_t = \mathcal{F}\left(\mathcal{Y}_{t-k}, \mathcal{W}_{t-k}^i\right) \tag{2}$$

where $\mathcal{F}(\cdot)$ is the estimation function. In Eq. 2, the *forecasting lead time* is k months, i.e. we aim to forecast the demand k months earlier. The forecasting problem using GST and other data sources can be formulated in a similar way.

3.2 The Behavior Extraction Problem

Let us assume that a typical trip planning has n activities, $\mathbb{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_n\}$. As an example, these activities may include flight reservation, hotel booking, and planning to visit different attractions. The problem of *tourists behavior extraction* is to determine how long before the actual trip time, a typical tourist performs each of these activities and is reported as a distribution over time. We use $\mathrm{a}_j^{\mathbf{A}_i}$ to show the fraction of tourists that perform activity $\mathbf{A}_i$ in j months earlier than the actual trip time. Then, we will have:

$$\sum_{j=0}^{\infty} \mathrm{a}_j^{\mathbf{A}_i} = 1 \quad , \quad i = 1, \cdots, n. \tag{3}$$

In reality, a finite upper bound can be considered for j. The traditional approach to acquire the $\mathrm{a}_j^{\mathbf{A}_i}$ values is through conducting surveys (HTA Report 2002-2013). However, since people use online resources for trip planning, we expect that usage trends of online resources can be used to extract average tourist behavior during trip planning. Wiki-pages, based on their contents, can be classified into different categories. As an example, a Wiki-page that is about a national park can be classified as an *attraction*. This classification can be conducted based on the trip planning activities in $\mathbb{A}$. Then, the problem of tourists behavior extraction using WUTs is to determine how tourists visit the pages in each category prior to their trip and is formulated as follows:

$$\mathrm{a}_j^{\mathbf{A}_i} = \mathcal{B}\left(\mathcal{Y}^{\mathcal{D}}, \mathcal{Y}^{\mathcal{I}}, \mathbf{W}^{\mathbf{A}_i}\right) \tag{4}$$

where, $\mathcal{B}$ is the estimation function and $\mathbf{W}^{\mathbf{A}_i}$ is a set of WUT time series of pages that belong to the category of $\mathbf{A}_i$.

4 Dataset Description

In this paper, we mainly use Wikipedia pages and their usage trends for tourism in Hawaii. Monthly tourism demand time series for Hawaii are available through the website of the *Department of Business, Economic Development and Tourism* of the state of Hawaii[1]. We collected Wikipedia pages and their usage trends from English Wikipedia and Wikipedia Trends[2] websites. For this purpose, we conducted a breadth-first-search approach on the graph of Wikipedia pages, beginning from the page for Hawaii [3]. We retrieved all the Wiki-pages related to Hawaii by extracting pages that contain a minimum number of mentions of *Hawaii*. Using this approach we retrieved 3,126 pages. Also, we collected GST time series for *Hawaii*, *Hawaii flight*, and *Hawaii hotel* search terms. Considering the data availability, we focused our study on the time frame between January 2008 and April 2015.

Page Classification: For further analysis, Wiki-pages are classified into the following tourism-related categories:

- *Attractions*: pages that provide information about various types of attraction such as beaches and museums.
- *Flights*: pages that provide information about airlines and airports in Hawaii.
- *Schools and universities*: pages that provide information about schools, colleges, and universities in Hawaii.
- *Events*: pages that provide information about various events in Hawaii such as festivals and sport events.
- *Transportation*: pages that provide information about different types of transportation in Hawaii.

[1] dbedt.hawaii.gov/visitor/tourism/

[2] www.wikipediatrends.com

[3] en.wikipedia.org/wiki/Hawaii

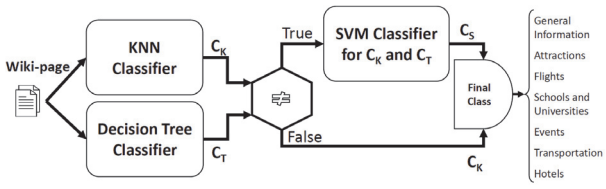


Figure 1: The classifier used for Wiki-page categorizing.

Table 1: Number of pages in each category.

Category	# Pages	Category	# Pages
General Information	2202	Attractions	518
Flights	53	Events	50
Transportation	81	Schools and universities	202
Hotels	20		

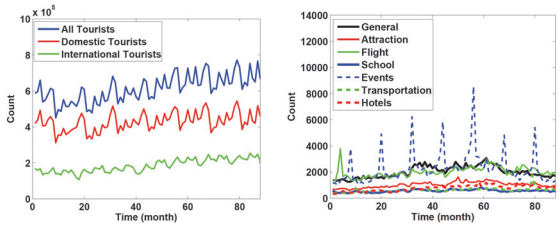
- *Hotels*: pages that provide information about hotels and accommodations in Hawaii.
- *General information*: other pages.

Using a dictionary of keywords for each category, Wiki-pages are mapped to the feature space using a bag-of-words approach. Furthermore, training, test, and validation datasets are constructed by random selection of 85, 60, and 150 pages, respectively. These pages are labeled manually based on their content. Preliminary experiments with various classifiers showed that for each of the pages in the validation dataset, at least one of the kNN and Decision Tree (DT) classifiers return the correct label of the page. Hence, we deployed a hierarchical classification model, illustrated in Fig. 1, to classify the rest of the pages. In the deployed ensemble classifier, we first train kNN and DT classifiers. Then, we classify each Wiki-page using the kNN and DT to have C_K and C_T classes, respectively. If $C_T = C_K$, we accept C_K as the final class of the page. On the other hand, when $C_T \neq C_K$, we train an SVM classifier to classify the page for two classes. The result of the SVM classifier is then accepted as the category of the page. We tested our method on the test dataset and the classification error was 8%. Table 1 shows the number of pages in each category. Tourism demand time series are illustrated in Fig. 2(a). We can observe that the monthly count of the domestic tourists is always higher than the monthly count of the international ones. For comparison purposes, the average usage trends of the above categories are illustrated in Fig. 2(b).

5 Forecasting using Wikipedia Usage Trends

In this section, we aim to explore how WUT influences the accuracy of tourism demand forecasts. A well known approach for tourism demand forecasting is to use autoregressive models, AR(m). Here, we deploy this approach as the building block of other regression methods. The general model of AR(m) is as follows:

$$\hat{y}_t^{AR} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} \qquad (5)$$



(a) (b)

Figure 2: (a) Tourism demand time series (b) Average usage trend of page categories.

where, k is the forecasting lead time, α_j's and β are regression coefficients, and m is the order of the model. In the following subsections we extend this model to incorporate WUT and GST time series in the forecasting process.

5.1 Regression with Wikipedia Usage Trends

The general model of tourism demand forecasting using WUT is defined in Eq. 2. Using a linear regression setting for the forecasting function $\mathcal{F}(\cdot)$, the forecasting model based on the usage trend of Wiki-page W_i is as follows:

$$\hat{y}_t^{W_i} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j w_{t-k-j}^i \qquad (6)$$

where β, α_j's, and γ_j's are regression coefficients. In the rest of the paper, we denote this model by ARW(m). Note that ARW(m) is an autoregressive exogenous (ARX) model where the external variable is a WUT time series.

Experiments show that each of the Wiki-pages results in a slightly different estimation and the accuracy of estimation based on each individual Wiki-page changes over time. To overcome this variation and in order to improve the accuracy of the forecasts, we use an ensemble method to aggregate the estimations that have been achieved based on different Wiki-pages. Let us assume that we have P Wikipedia pages and we aim to forecast the tourism demand at time t, y_t, based on the data that is available uptil time $t-k$. For this purpose, we name the latest v months of the available data as the *validation set*, $\mathbb{V} = \{y_{t-k-v+1}, \cdots, y_{t-k}\}$. Then, for each Wiki-page, the average performance of ARW(m) is measured on the validation set, $\mathbb{V}$. The top N pages that result in the highest accuracy are then selected to perform the final forecasting. We represent this set of top-N Wiki-pages as $\mathbf{W}_N^{Top}$. Then, the final forecast is performed as follows:

$$\hat{y}_t^{W} = \frac{1}{N} \sum_{W_i \in \mathbf{W}_N^{Top}} \hat{y}_t^{W_i} \qquad (7)$$

where $\hat{y}_t^{W_i}$ is determined based on Eq. 6. This algorithm, which we name it as Wiki-based Regression(m,N), is illustrated in Algorithm 1.

5.2 Regression with Google Search Trends

Let us assume that we have Q time series of GST and we aim to forecast the tourism demand at time t, y_t, based on

Algorithm 1 Forecasting with WUT Algorithm
Function ARW performs forecasting based on Eq. 6

function WIKI-BASED REGRESSION($\mathbf{W}$,$\mathcal{Y}$,N,t,m,k)
  for $p = 1$ to P **do**
    $E_{Sum} = 0$
    for $i = 1$ to v **do**
      $\hat{y}$ = ARW($\mathcal{W}^p_{t-i-k}$,$\mathcal{Y}_{t-i-k}$,m,k)
      $E_{Sum} = E_{Sum} + (y_{t-i} - \hat{y})^2$
    $RMSE_p = \sqrt{E_{Sum}/v}$
  $y_{Sum} = 0$
  for $i = 1$ to N **do**
    p = Wiki-page index with i^{th} lowest RMSE
    $\hat{y}$ = ARW($\mathcal{W}^p_{t-k}$,$\mathcal{Y}_{t-k}$,m,k)
    $y_{Sum} = y_{Sum} + \hat{y}$
  return $\frac{1}{N} y_{Sum}$

the data available uptil time $t-k$. Then, we will have:

$$\hat{y}^{G_i}_t = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j g^i_{t-k-j} \quad (8)$$

where β, α_j's, and γ_j's are regression coefficients. In the rest of the paper, we call this algorithm as ARG(m). Then, the final estimation of tourism demand using GST time series, $\hat{y}^G_t$, is calculated using the following equation:

$$\hat{y}^G_t = \frac{1}{Q} \sum_{i=1}^{Q} \hat{y}^{G_i}_t \quad (9)$$

where $\hat{y}^{G_i}_t$ is calculated using ARG(m) and the overall estimation method is called Google-based Regression(m).

Similar to Wiki-based and Google-based regressions, we can combine the forecasting results of individual regressions based on WUTs and GSTs in the following manner:

$$\hat{y}^{WG}_t = \frac{1}{N+Q} \left(\sum_{W_i \in \mathbf{W}^{Top}_N} \hat{y}^{W_i}_t + \sum_{i=1}^{Q} \hat{y}^{G_i}_t \right) \quad (10)$$

where $\hat{y}^{W_i}_t$ and $\hat{y}^{G_i}_t$ are the estimations of y_t based on Eq. 6 and Eq. 8, respectively. We call the overall ensemble method as Google-Wiki Ensemble.

5.3 Experimental Results

To study the performance of the forecasting techniques we performed experiments on three tourist arrival time series of Hawaii: $\mathcal{Y}^D$, $\mathcal{Y}^I$, and $\mathcal{Y}$. For each time series, the last 12 months are considered as the test dataset. The value of N in Algorithm 1 was set to 10 for $\mathcal{Y}$ and to 15 for $\mathcal{Y}^D$ and $\mathcal{Y}^I$. The size of the validation set in Algorithm 1, v, was 12 months. We used root-mean-square error (RMSE) of estimations as a measure of accuracy. Results are illustrated in Figures 3 and 4.

The average performance of various forecasting methods with different parameter settings are compared in Fig. 3. Here, k and m are changing between 1 and 9. In Fig. 3 (a), (c), and (e), for each value of k, we changed the value of m from 1 to 9 and reported the average RMSE of the forecasts. Similarly, in Fig. 3 (b), (d), and (f), for each value of m, we changed the value of k from 1 to 9 and reported the average RMSE of the forecasts. It is clear from Fig. 3 that almost everywhere, the accuracy of forecasting based Wiki-based Regression and Google-Wiki Ensemble is higher than the accuracy of AR(m) and Google-based Regression. The RMSE of the four forecasting methods for the best values of m are compared in Fig. 4. This figure compares the performance of the forecasting methods for four different lead times. Results show that in most of the cases either Wiki-based Regression or Google-Wiki Ensemble results in the lowest RMS errors.

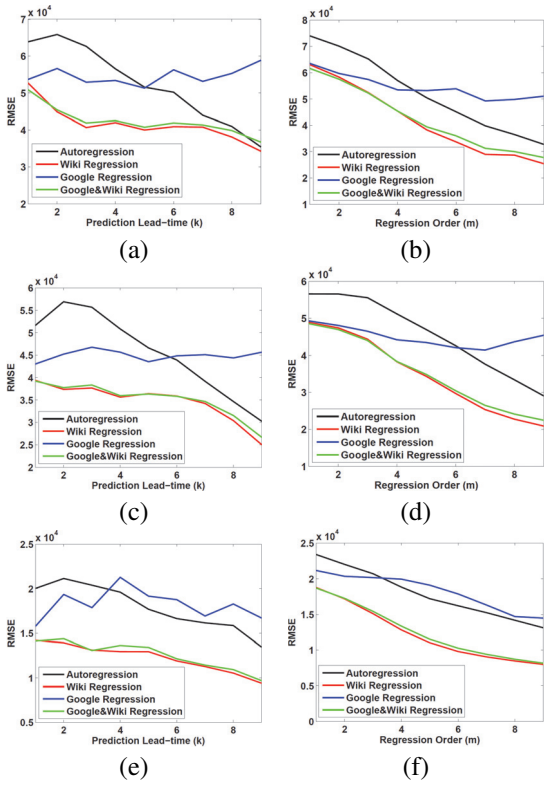


Figure 3: The average RMSE of different regression algorithms w.r.t. m and lead-time: (a) and (b) All tourists, (c) and (d) domestic tourists, (e) and (f) international tourists.

The importance of each category of Wiki-pages for domestic and international tourists is illustrated in Fig. 5. In this experiment, similar to Algorithm 1, we used the method of ARW(m) to determine the top-250 Wiki-pages, $\mathbf{W}^{Top}_{250}$. Then, we measured the ratio of pages that belong to each category in $\mathbf{W}^{Top}_{250}$. This figure depicts that domestic and international tourists have different behaviors. As an example, while for domestic tourists, the *Hotel* category is highlighted with lead time of 1, for international tourists it is highlighted with lead time 6. We can also observe from this figure that Wiki-pages in *Flight*, *Hotel*, *Events*, and *Transportation* categories are more important than the other three ones.

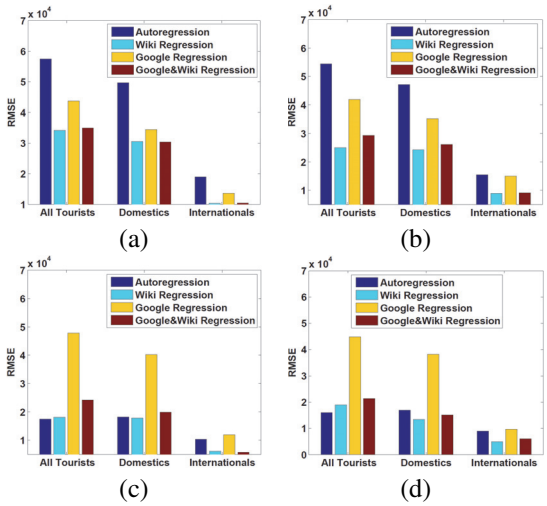


Figure 4: Forecasting RMS errors for different regression algorithms with (a) one month lead-time, (b) 3 months lead-time, (c) 6 months lead-time, and (d) 9 months lead-time.

Figure 5: Word clouds illustrating the importance of each class of Wikipedia pages in forecasting the tourism demand.

6 Wikipedia Usage Behavior

One of the important issues in tourism industry is to know when people start to plan their trips and how they use online resources for that purpose. In this section, we propose a data-driven framework based on WUTs to answer this question.

We categorize tourists into two groups: Wikipedia readers and non-Wikipedia readers. A Wikipedia reader tourist is a person that uses Wikipedia as an online resource before and during his trip. We assume that every Wikipedia reader tourist visits a Wiki-page only once and that the fraction of all the tourists that are Wikipedia reader, K_{W_i}, is constant for each specific page. Hence, if at time t we have y_t tourists, $y_t^{W_i}$ of them are Wikipedia readers that read Wiki-page W_i and we have: $K_{W_i} = y_t^{W_i}/y_t$. To simplify the analysis, we assume that all Wikipedia reader tourists start to read Wiki-pages at most M months before their actual trip time. Results in section 5.3 indicate that domestic and international tourists may have different planning behaviors. Therefore, in this framework, we disaggregate the behavior of domestic and international tourists and based on the above assumptions, we have:

$$w_t^i = \sum_{j=0}^{M} K_{W_i}^D b_j^i y_{t+j}^D + \sum_{j=0}^{M} K_{W_i}^I c_j^i y_{t+j}^I + \nu_t + e_t \quad (11)$$

where, w_t^i is the number of visits to the Wiki-page W_i, and y_{t+j}^D and y_{t+j}^I are the number of domestic and international tourists at time $t+j$, respectively. Furthermore, $K_{W_i}^D$ and $K_{W_i}^I$ represent the ratio of domestic and international Wikipedia reader tourists, respectively. Also, b_j^i's and c_j^i's are constant coefficients that determine ratio of Wikipedia reader tourists that have read W_i, j months before the trip time. Note that since coefficients b_j^i and c_j^i represent the reading distributions, these coefficients cannot be negative and each set should be summed up to 1. In Eq. 11, ν_t and e_t are two error components, illustrating the uncertainties that we have in the model. Uncertainty (or variation) in the assumptions and data (i.e. noise in K_{W_i}, w_t^i, etc.) is modeled by e_t and Wiki-page visits that have a non-tourist reason are shown by ν_t. The intuition behind this model is that the number of Wikipedia page-visits for a specific area is a linear function of the number of tourists that will visit that specific area (i.e. in this paper Hawaii) in the future, plus some noise.

Yearly trip-planning surveys conducted for Hawaii (HTA Report 2002-2013) show that trip planning activities follow a normal-like distribution. Since we assumed that Wikipedia-reader tourists read Wiki-pages during their trip planning, we expect that the reading distributions follow a normal-like distribution as well. Therefore, in order to apply this prior knowledge to the model and to dictate the estimation model to follow a normal-like distribution we need to add the following set of constraints:

$$b_0 \le b_1 \le \cdots \le b_{\beta-1} \le b_\beta \ge b_{\beta+1} \ge \cdots \ge b_M \quad (12)$$
$$c_0 \le c_1 \le \cdots \le c_{\zeta-1} \le c_\zeta \ge c_{\zeta+1} \ge \cdots \ge c_M$$

where β and ζ are indexes that the maximum value of distributions occur for b and c, respectively.

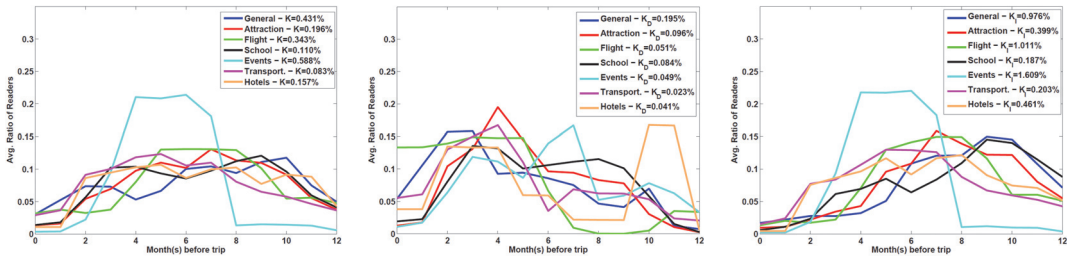
In order to estimate the Wikipedia reading distributions of domestic and international tourists, we first drop the error factors, ν_t and e_t, from Eq. 11 to calculate the estimated number of visits to the Wiki-page W_i, i.e. $\hat{w}_t^i$. Then, we solve the following optimization problem:

$$b_j^{i\,*}, c_j^{i\,*}, K_{W_i}^{D\,*}, K_{W_i}^{I\,*} = \arg\min \frac{1}{T} \sum_{t=1}^{T} \left(\hat{w}_t^i - w_t^i\right)^2 \quad (13)$$

$$\begin{aligned} s.t. \quad & 0 < K_I \le 1 \quad , \quad 0 < K_D \le 1 \\ & 0 \le b_j^i \le 1 \, , \, 0 \le c_j^i \le 1 \quad , j = 0, 1, \cdots, M \\ & \sum_{j=0}^{M} b_j^i = \sum_{j=0}^{M} c_j^i = 1 \\ & b_0 \le b_1 \le \cdots \le b_{\beta-1} \le b_\beta \ge b_{\beta+1} \ge \cdots \ge b_M \\ & c_0 \le c_1 \le \cdots \le c_{\zeta-1} \le c_\zeta \ge c_{\zeta+1} \ge \cdots \ge c_M \end{aligned}$$

In the optimization problem of Eq. 13, M, β, and ζ are model parameters which are estimated through cross validation.

Note that solving the problem of Eq. 13 results in two reading distributions: b_j^i's for domestic tourists and c_j^i's for international ones. It is easy to show that by taking the weighted average of these two distributions, one can calculate the average reading distribution of all tourists. In other



(a) (b) (c)

Figure 6: Reading distributions averaged over different page categories for (a) all, (b) domestic, and (c) international tourists.

words, if N is the total number of tourists and N_D is the number of domestic tourists we will have:

$$a_j^i = \frac{N_D}{N} b_j^i + \frac{N - N_D}{N} c_j^i \quad , \quad j = 0, 1, \cdots, M \qquad (14)$$

The average of reading distributions for category of pages (as defined in Section 4) are illustrated in Fig. 6. In order to calculate the average distribution of category $\mathbf{A}$ we use the following equation:

$$\bar{a}_j^{\mathbf{A}} = \frac{\sum_{W_i \in \mathbf{A}} K_{W_i} a_j^i}{\sum_{W_i \in \mathbf{A}} K_{W_i}} \qquad (15)$$

where K_{W_i} is the average ratio of Wikipedia reader tourists for page W_i. Similar to Eq. 14, K_{W_i} is calculated using $K_{W_i}^D$ and $K_{W_i}^I$. Fig. 6(a) shows the average reading distributions for all the tourists. The average reading distributions of domestic and international tourists for category $\mathbf{A}$, i.e. $\bar{b}_j^{\mathbf{A}}$ and $\bar{c}_j^{\mathbf{A}}$, can also be calculated in a similar way and are illustrated in Fig. 6 (b) and (c). Figure 6 also shows the average ratio of Wikipedia reader tourists for each category which is calculated as follows:

$$K = \frac{\sum_{W_i \in \mathbf{A}} K_{W_i}}{|\mathbf{A}|} \qquad (16)$$

where $|\mathbf{A}|$ represents the size of category $\mathbf{A}$.

Figure 6(a) depicts that on average, most of the Wikipedia reading activities have occurred about 4 to 8 months prior to the trip. This is inline with the surveys reported in (HTA Report 2002-2013) as the mean decision date for most of the activities are between 4 to 8 months before the actual arrival date. Figure 6(a) also indicates that the most popular categories are Events, General information, and Flights. A comparison of Fig. 6(b) and (c) shows that international tourists start planning their trip about 4 to 6 months before the domestic ones. Furthermore, we can observe that on average, the ratio of Wikipedia reader tourists is higher among international tourists than the domestic ones.

7 Conclusion

In this paper we showed that Wikipedia usage trends can be effectively used in tourism planning. Experimental results indicated that WUT time series improve the accuracy of tourism demand forecasts. We also used WUT and tourism demand time series to estimate the behavior of tourists in trip planning. Results are consistent with the survey results gathered from domestic and international tourists of Hawaii. As future work, we aim to use other sources (e.g. TripAdvisor) to *rank* the Wikipedia pages and consider the relationship graph of Wiki-pages to improve forecasting and behavior estimation performances.

References

Artola, C., et al. 2015. Can internet searches forecast tourism inflows? *Int. Journal of Manpower* 36(1):103–116.

Athanasopoulos, G., and de Silva, A. 2012. Multivariate exponential smoothing for forecasting tourist arrivals. *Journal of Travel Research* 51(5):640652.

Bangwayo-Skeete, P., and Skeete, R. 2015. Can google data improve the forecasting performance of tourist arrivals? mixed-data sampling approach. *Tourism Management* 46:454–464.

Chan, F., et al. 2005. Modelling multivariate international tourism demand and volatility. *Tourism Management* 26:459471.

Choi, H., and Varian, H. 2012. Predicting the present with google trends. *The Economic Record* 88:2–9.

Fang, G., et al. 2015. How to extract seasonal features of sightseeing spots from twitter and wikipedia. *Bulletin of Networking, Computing, Systems, and Software* 4(1):21–26.

Gunter, U., and Onder, I. 2015. Forecasting international city tourism demand for paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management* 46:123–135.

Hecht, B., et al. 2007. Generating educational tourism narratives from wikipedia. In *Intelligent Narrative Technologies, Papers from the 2007 AAAI Fall Symposium*, 37–44.

Hickmann, K., et al. 2015. Forecasting the 20132014 influenza season using wikipedia. *PLoS Comput Biol* 11(5).

HTA Report 2002-2013. Visitor satisfaction & activity. available at: hawaiitourismauthority.org/research/reports/visitor-satisfaction/ Last accessed Aug. 2015.

Morley, C., et al. 2014. Gravity models for tourism demand: theory and use. *Annals of Tourism Research* 48:1–10.

Song, H., and Li, G. 2008. Tourism demand modeling and forecasting-a review of recent research. *Tourism Management* 29:203–220.

Song, H., and Witt, S. F. 2006. Forecasting international tourist flows to macau. *Tourism Management* 27:214224.

Xiang, Z., and Pan, B. 2011. Travel queries on cities in the united states: Implications for search engine marketing for tourist destinations. *Tourism Management* 32:8897.

Yang, X., et al. 2015. Forecasting chinese tourist volume with serach engine data. *Tourism Management* 46:386–397.