Note

Use of online travel agencies as a data source for tourism marketing

Shohei Suzuki (Department of Business Administration, Senshu University, s_suzuki@senshu-u.jp)

Abstract

This paper focuses on online travel agencies (OTAs) as a quantitative data source for tourism marketing. Most research on OTA data has focused on hotel ratings and hotel reviews, and there was insufficient discussion regarding the use of real-time updated data, such as provided plan information. Therefore, we conducted research to characterize such data. First, we constructed a database that continuously collects and accumulates data from OTAs. We verified the reliability of the data by estimating numbers of guests using the data. The results indicate that the reliability of the data depends on the scale of the analyzed city.

Keywords

online travel agency, big data, tourism statistics, tourism marketing, fixed effects model

1. Introduction

The tourism industry continues to grow globally. Therefore, many countries are making efforts to attract tourists, but decision making based on quantitative data analysis is lacking. One of the main causes of this problem is a lack of quantitative data available for analysis. In response to this lack, the Japan Tourism Agency has begun preparing tourism statistics based on common standards in Japan. However, it is difficult to perform detailed analysis of specific tourist destinations because data are often only available at a prefectural resolution. Additionally, there are only a few types of data, and only some can be used for tourism promotion throughout Japan [Okamoto et al., 2020]. Therefore, additional data preparation is required, but there is insufficient manpower for improving the accuracy of data and types of data using conventional methods.

These problems are not unique to Japan. Researchers in many countries are searching for new data collection methods based on information technology. In particular, the analysis of social networking service (SNS) data, such as Twitter and Flickr data, is being actively conducted as a part of tourism research. Such research includes the estimation of tourist sentiment by analyzing the content of posts [Jabreel et al., 2017], visualization of human flows by analyzing location information in posts [Hawelka et al., 2014], and destination image extraction by analyzing posted photos [Miah et al., 2017]. However, it is impossible for an analyst to control the amount and quality of data collected from SNS, meaning there is little certainty that the desired results will be obtained [Suzuki, 2018]. Another concern is the privacy of individuals [Yallop and Seraphin, 2020]. For these reasons, SNS data are considered by many organizations as a complementary information source, rather than a major information source for decision making.

In this study, we focused on online travel agencies (OTAs) as a data source among internet services. By collecting data from OTAs, it is possible to collect data that are directly linked to the purchasing behaviors of users continuously in a unified format. OTAs are free platforms that allow users to search for and book accommodations in the area that they wish to visit. The website “Booking.com” is a typical example. OTAs generate a huge amount of data every day. Such data include whether it is possible to stay at the target accommodation on a specific date and the contents of plans that can be reserved. However, users cannot retroactively collect information regarding plans that were provided in the past. Therefore, to use this data for analysis, a database for analysis that collects and accumulates data in real time is required. In contrast, hotel ratings and hotel reviews posted by users in the past can be easily collected. Therefore, prior research using OTAs as a data source has focused on the analysis of hotel ratings and hotel reviews by targeting Expedia and Booking.com [Gu and Ye, 2014; Xiang et al., 2017; Mariani and Borghi, 2018]. Mariani and Borghi [2018] pointed out that there is a need for discussion regarding the use of OTAs as a data source, but there has been insufficient discussion regarding information other than hotel ratings and hotel reviews, leaving significant room for additional research.

To analyze OTAs as a data source from various perspectives, we attempted to construct a database for analysis by continuously collecting OTA data. This paper describes the constructed database and outlines the characteristics of the data. Furthermore, we verify the reliability of OTA data for analyzing the situations in various areas by estimating numbers of guests based on the collected data. This research can serve as a reference for other researchers and help improve the accuracy of quantitative analysis using OTAs as a data source.

2. Database for analysis

In this study, we collected data from “Jalan.net” (hereafter referred to as “Jalan”), which is the most commonly used OTA in Japan [Recruit Lifestyle, n.d.a]. Users can collect basic information, such as the addresses of accommodations and information regarding offered plans using the Jalan application programming interface [Recruit Lifestyle, n.d.b]. We constructed a database for analysis by continuously collecting and accumulating such data. Specifically, we developed a program to collect the information of all plans that can be reserved when the stay date is four weeks (28 d) after the data collection day for all accommodations in Japan registered in Jalan. We have been collecting data daily since December of 2018 using

Table 1: Example data collected from Jalan

Area CD	Hotel ID	Collection Date	Stay Date	Plan CD	Price	Room Info
011111	111111	2019-04-01	2019-04-02	111111	10,000	Single
⋮	⋮	⋮	⋮	⋮	⋮	⋮
479999	999999	2019-04-01	2019-04-29	99999	90,000	Double

Table 2: Example accommodation information

Hotel ID	Name	Address	Post CD	Type
111111	Hotel A	Sapporo, Hokkaido	111-1111	Hotel
⋮	⋮	⋮	⋮	⋮
999999	Ryokan Z	Naha, Okinawa	999-999	Ryokan

this program. The records that are updated daily include area code, hotel ID, data collection date, stay date, plan code, price, and room information. Table 1 presents an example record. The number of updated records varies depending on the day, but it is typically within the range of 6,000,000 to 10,000,000 per day. In this database, the information of all plans that can be booked on a specified stay date is selected by simply specifying the stay date. Additionally, by referring to the accommodation information (Table 2) created separately using hotel IDs, information such as the name, address, and type of each accommodation can be used for analysis. In this study, we selected and analyzed the data for the year of 2019 from this database. Specific methods are described explained in the following section.

3. Correlation between the number of remaining plans and the number of guests

3.1 Method

In this paper, the number of plans that can be booked on the day before a specific stay date is referred to as the “number of remaining plans.” The definition of the number of remaining plans is described below. If a user attempted to book a stay for April 1 of 2019, which was a weekday, the number of plans that could be booked with OTAs on the previous day was 178,021 in Japan. Therefore, this number is the number of remaining plans for April 1 of 2019 in Japan. If users attempted to book a stay for May 3 of 2019, which was a national holiday, the number of remaining plans was only 30,844, which is significantly less than the number for April 1. These numbers indicate that the number of remaining plans changes every day and that many plans are booked on popular travel days, resulting in a small number of remaining plans. In other words, there is a negative correlation between the number of remaining plans and the number of guests.

We will now describe data limitations. The plans provided by OTAs are not provided on a per-room basis. In other words, even if two or more rooms are vacant, the plan booking those rooms may be counted as one plan. In contrast, even if only one room is available, the number of remaining plans may be two or more when multiple plans booking the same room are provided. Therefore, the number of remaining plans does not necessarily match the number of bookings. Therefore, to estimate the number of guests more accurately, it is desirable to use the number of bookings or the number of remaining rooms, but there are no OTAs that can collect this information accurately. Therefore, in this paper, we estimate the number of guests according to the number of remaining plans.

The number of guests is estimated based on the results of a statistical survey on overnight travel published by the Japan Tourism Agency [2020]. This survey provides the number of guests for each prefecture and each month and does not contain any more detailed data. Additionally, the number of remaining plans can be aggregated daily for each city. To match the granularity levels of these two variables, we create monthly data for each prefecture. Therefore, in this study, panel data from 12 months and 47 prefectures was used for analysis.

If this data format is used for analysis, results may be affected by differences between prefectures. Specifically, there are many accommodations in prefectures with many guests, meaning there are typically many remaining plans. In other words, it is likely that there is a positive correlation between the number of remaining plans and the number of guests. As described above, different hypotheses can be considered when focusing on regional differences and seasonal differences. Therefore, we conducted two different types of analysis.

First, we analyzed changes in the number of remaining plans and the number of guests caused by regional differences using the ordinary least squares (OLS) method. We used the mean value for each prefecture for both the number of guests and number of remaining plans. This method is also known as “between estimation.”

Next, we analyzed changes in the number of remaining plans and number of guests based on seasonal differences within each prefecture using “within estimation,” which is a fixed-effect model [Hsiao, 2014]. Within estimation is equivalent to OLS with deviation. We used the deviations between the mean values in each prefecture and each month for both the number of guests and number of remaining plans. This method can eliminate the effect of differences between prefectures. The use of least squares dummy variables can be considered as another method for eliminating the effects of differences between prefectures, but it was not used in this research based on the large number of variables.

3.2 Result

Table 3 lists descriptive statistics for the data used in our analysis. The mean of the number of guests is approximately 1,000,000 and the mean of the number of remaining plans is approximately 100,000. One can see that the number of remaining plans is approximately 10 % of the number of guests. Figure 1 presents a scatter plot of the data used in our analysis. From this distribution, one can see that there is a positive correlation between the number of remaining plans and the number of guests. We describe the effects of regional differences and seasonal differences below.

Table 3: Descriptive statistics of panel data

	Mean[a]	S.D.[a]	Max[a]	Min[a]
Guests	963,184	1,023,482	6,048,850	124,970
Plans	108,509	106,410	698,057	15,510

Notes: Guests/The numbers of guests by the Japan Tourism Agency; Plans/The numbers of remaining plans by OTA. [a] 47 (the numbers of prefectures) × 12 (months).

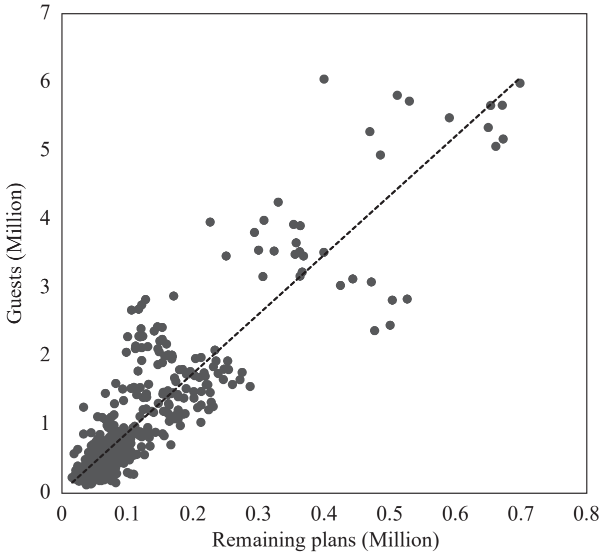

Figure 1: Correlation between the number of remaining plans and guests in each prefecture (Monthly)

For the results of OLS using mean values, the regression coefficient was 9.243 ($p < 0.01$) and the coefficient of determination was 0.900. This regression coefficient is significantly positive. As shown in Figure 2, there is a positive correlation between the number of remaining plans and the number of guests in each prefecture. In other words, when focusing on the differences between prefectures, prefectures with large numbers of remaining plans should also have large numbers of guests, which supports our hypothesis. The coefficient of determination of 0.900 indicates that the market sizes of the prefectures indicated by OTA data match the actual market sizes. Therefore, the number of guests in a prefecture can be estimated with high accuracy based on the number of remaining plans. However, this method cannot grasp increases or decreases in the number of guests caused by time series change in each prefecture.

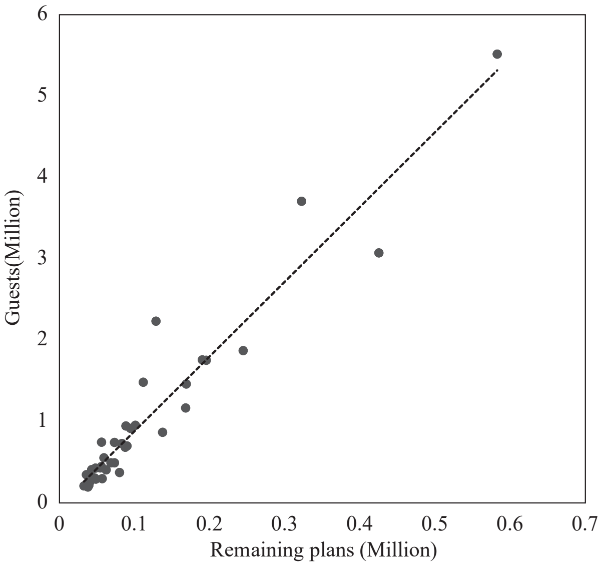

Figure 2: Correlation between the number of remaining plans and guests in each prefecture (Annual mean)

In the results of OLS using deviation, the regression coefficient was −2.925 ($p < 0.01$) and the coefficient of determination was 0.187. This regression coefficient is significantly negative, which is the opposite of the previous result. Furthermore, as shown in Figure 3, the number of guests tends to decrease as the number of remaining plans increases, supporting our

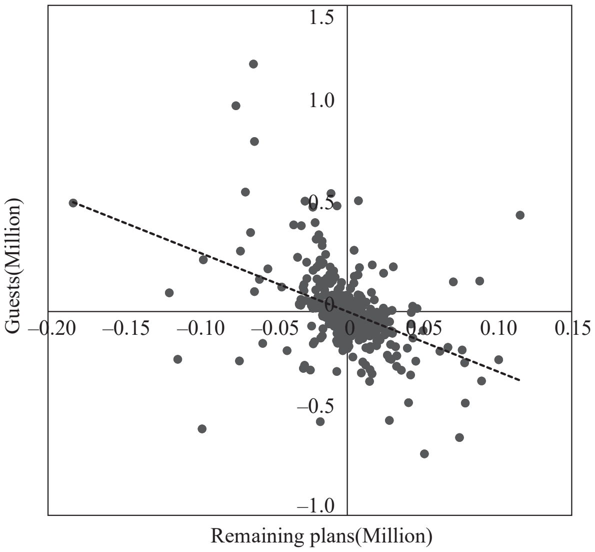

Figure 3: Correlation between the number of remaining plans and guests in each prefecture (Deviations)

hypothesis. Therefore, it is possible to evaluate an increase or decrease in the number of guests caused by time series changes by considering the relative number of remaining plans within a prefecture and excluding the differences between prefectures.

Our analysis results are summarized below. When looking at the data for each prefecture, the number of guests tends to be large in prefectures with large remaining numbers of plans. In contrast, when looking at the relative data for each month, the number of guests tends to be small in months with large remaining numbers of plans. Therefore, analysts must select an appropriate analysis method according to their goals.

3.3 Discussion

As mentioned above, the average number of guests in each prefecture can be estimated with high accuracy based on the number of remaining plans. However, the coefficient of determination when estimating the number of guests in each month is only 0.187. It appears that prefectures that do not have a significant correlation between the number of remaining plans and number of guests in each month reduced the coefficient of determination. For example, in Figure 3, the upper-right and lower-left points both represent Tokyo. This pattern differs significantly from the distribution of the other points. Because Tokyo is the largest city in Japan, there is a possibility that there is no correlation between the number of remaining plans and number of guests in large cities. Therefore, we calculated the correlation coefficient (Pearson's r) between the two variables for each prefecture and implemented a correlation test. The results are discussed below.

First, the r value for Tokyo is −0.120 ($p > 0.05$) and there is no significant correlation between the two variables. The r values of the other major cities in Japan, namely Osaka, Aichi, and Fukuoka, are −0.270, −0.290, and −0.067, respectively ($p > 0.05$ for all). In contrast, small cities exhibit strong correlations. For example, Aomori, Wakayama, Fukui, and Akita have r values of −0915, −0.876, −0.872, and −0.870, respectively ($p < 0.01$ for all).

Based on the above results, it can be concluded that when estimating the number of guests based on the number of remaining plans, there is a difference between prefectures in terms of accuracy, where small cities tend to have high accuracy and large cities tend to have low accuracy. We hypothesize that accuracy is reduced by the inherent limitations of the considered data. As mentioned previously, the number of rooms per plan may be two or more, or may be one or less. In other words, the numbers of plans and numbers of guests are not directly correlated in large hotels where the number of remaining plans does not reach zero, even if many reservations are made, or in hotels that offer various plans for a given room. One reason why the estimation accuracy for large cities is low is that there may be many large hotels in such cities.

4. Conclusion and future work

In this study, we focused on the issue of limited discussion regarding the use of real-time updated OTA data. We constructed a database through continuous data collection to remedy this issue. Furthermore, we analyzed the correlation between the number of remaining plans on the day before a given stay and the number of guests. It was determined that there is a positive correlation between the numbers of remaining plans and numbers of guests between prefectures. This result is considered to be affected by the market size of each prefecture. The results of analysis using a method to eliminate the effects of differences between prefectures revealed that there is a negative correlation between the two considered variables. This result suggests that it is possible to interpret an increase or decrease in the number of guests observed in time series changes based on the relative number of remaining plans within a prefecture.

These analysis results will be one of the main criteria for future research using OTAs as a data source. However, some subjects require further attention.

First, it is necessary to improve our method for estimating the number of guests in each month based on the number of remaining plans. To improve accuracy, it would be helpful to clarify effective variables other than the number of remaining plans. For example, knowing whether the type of room that can be reserved is single, double, or triple may help improve accuracy. Furthermore, accuracy can be further improved if the total number of rooms in an accommodation is known. This information is not provided by Jalan, but it can be collected by some OTAs, such as Rakuten Travel. Increasing the accuracy of estimation will aid tourism marketing in areas where statistical data are not available, such as small towns and villages.

We analyzed numbers of available plans, but other data from OTAs have not been analyzed. In particular, analysis related to pricing would be helpful for marketing. Based on the results of this research, it was determined that the interpretation of analysis results for data from OTAs differs between prefectures. It is also necessary to clarify differences in prices between prefectures. Furthermore, Mariani. and Borghi reported that the distribution of hotel scores differs according to hotel classes [Mariani and Borghi, 2018]. Therefore, it is necessary to clarify not only the differences between prefectures, but also differences based on the characteristics of accommodations.

References

Gu, B. and Ye, Q. (2014). First step in social media: Measuring the influence of online management responses on customer satisfaction. *Production and Operations Management*, Vol. 23, No. 4, 570-582.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, Vol. 41, No. 3, 260-271.

Hsiao, C. (2014). *Analysis of panel data*. Cambridge University Press.

Japan Tourism Agency (2020). Statistical survey on overnight travel (Retrieved March 1, 2020 from https://www.mlit.go.jp/kankocho/siryou/toukei/shukuhakutoukei.html).

Jabreel, M., Moreno, A., and Huertas, A. (2017). Semantic comparison of the emotional values communicated by destinations and tourists on social media. *Journal of Destination Marketing & Management*, Vol. 6, No. 3, 170-183.

Mariani, M. M. and Borghi, M. (2018). Effects of the Booking.com rating system: Bringing hotel class into the picture. *Tourism Management*, Vol. 66, 47-52.

Miah, S. J., Vu, H. Q., Gammack, J., and McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information & Management*, Vol. 54, No. 6, 771-785.

Okamoto, N., Ogasawara, Y., Suzuki S., and Hihara, K. (2020). Current situation of regional tourism statistics and their challenges. *The International Journal of Tourism Science*, No. 13, 61-70. (in Japanese)

Recruit Lifestyle (n.d.a). Jalan.net (Retrieved May 7, 2020 from https://www.jalan.net/en/japan_hotels_ryokan/).

Recruit Lifestyle (n.d.b). Jalan Web service. https://www.jalan.net/jw/jwp0000/jww0001.do. (last accessed 2020-05-07)

Suzuki, S. (2018). Analysis of covert and overt interest in tourist attractions with use of Twitter data: Use of social media data in destination marketing. Ph.D. Thesis, Tokyo Metropolitan University (in Japanese).

Xiang, Z., Du, Q., Ma, Y., and Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, Vol. 58, 51-65.

Yallop, A. and Seraphin, H. (2020). Big data and analytics in tourism and hospitality: Opportunities and risks. *Journal of Tourism Futures*.

(Received August 28, 2020; accepted September 15, 2020)