Estimating establishment-level tourism sales using the Regional Tourism Economic Survey and Geographical Information System[1]

Kozo Miyagawa (Keio University)
Hiroyuki Kamiyama (Nomura Research Institute)
Ryuta Shimamura (Pasco)
Fumikado Yamamoto (Japan Tourism Agency)

1. Introduction

The scale of the regional tourism economy serves as one of the most important indices for regional governments and tourism-related organizations. To determine the scale of regional tourism, organizations generally survey tourists. However, such demand-side surveys are sometimes insufficient for measuring the scale of regional tourism.[2] As a result of these challenges, in Japan, the Japan Tourism Agency (JTA) conducted the Regional Tourism Economic Survey (RTES) of tourism-related establishments in 2012, and collected data to better understand the scale of the regional tourism economy, such as the percentage of tourism sales. Since the RTES is a sample survey, it is necessary to use a complete establishment list – like Japan's "Business Register" [3] – to determine regional tourism sales. Yet, because the "Business Register" does not specifically contain tourism-related information, regional tourism sales can only be estimated using variables not related to tourism, for example, the number of establishments in the region and the establishment's total sales. As such, the estimates of tourism sales using this data would lack accuracy. In order to best utilize data from the "Business Register" in future research, we must first determine the most accurate methods for estimating regional tourism sales.

One of the objectives of our research is to determine whether the estimation accuracy of tourism sales could be improved by applying Geographic Information System (GIS). Since the "Business Register" contains address information, tourism sales by establishment can be estimated using the distance of the establishment in relation to tourist attractions,

[1] This research was supported by a Grant-in-Aid for Scientific Research (C) (25380270) from the Japan Society for the Promotion of Science (JSPS) along with financial assistance and data from Pasco, a Japanese company that provides geospatial information.

[2] In Japan, the Japan Tourism Agency conducts both the "Statistics on Inbound Tourists by Prefecture" (a survey of tourists at their destinations) and the "National Tourism Survey" (a mail survey measuring the tourism consumption of each individual). However, both surveys feature major shortcomings in their methodology. For example, the high costs of directly surveying individual tourists limit the former's thoroughness, while the latter fails to connect the collected data to the place of consumption (destination). The National Tourism Survey only collects data based on the residence of the respondent. Miyagawa (2013) further explains how the JTA conducts its surveys.

[3] The Japanese government's Ministry of Internal Affairs only began compiling the "Business Register" in 2012.

accommodations, and mass transit as explanatory variables.

The second objective of this study is to determine whether the inclusion of an establishment's tourism-related characteristics improves accuracy of the regional tourism sales estimates. Beyond estimating the tourism sales ratio, the RTES also asks for tourism-related information for each establishment. For example: Is the establishment a member of a regional tourism association? Does the establishment have a parking lot? Does the establishment have a website? Is the establishment listed in travel guides?

In this paper, we apply both GIS and tourism-related characteristics to the methods used to estimate regional tourism sales to gauge whether or not they improve their accuracy. The data used in this study is the individual establishment data of the food services and retail sectors in Shizuoka Prefecture, as provided by the RTES.

The results of this research can also help improve the estimation accuracy of tourism domestic supply by industry in the sixth table of the Tourism Satellite Account (TSA). In this way, this research can also serve an important role in the compilation of a regional TSA.

Details about the data used in this study are presented in section 2. In section 3, an analytical method for estimating regional tourism sales in proposed, and the results of these estimations are detailed in section 4. Lastly, conclusions and further topics of research are offered in section 5.

2. Data

The primary data used in this paper comes from the Japan Tourism Agency's Regional Tourism Economic Survey conducted in 2012. For this survey, the JTA mailed questionnaires to approximately 90,000 establishments classified in tourism-related industries – e.g. accommodation, restaurant, transportation, car rental, travel agencies, cultural service, and retail – across Japan's 904 tourist regions. Although the questionnaire asks a number of questions, this study focuses on four pieces of data for each establishment:

1. Its location
2. Its industrial classification
3. Its percentage of tourism sales
4. Its tourism-related characteristics

For this paper, we used data from Shizuoka Prefecture, located about 200km west of Tokyo. Its famous tourist attractions are Mt. Fuji along with seaside resorts and hot springs.

In addition to the RTES data, we used the Economic Census to obtain the addresses and

revenues of accommodation establishments. Geographic data, such as the latitude and longitude of tourism attractions, train stations, and exits of expressways are also used in this research.

3. Analytical Method

As previously mentioned, the objective of this research is to prove that geographic data and the tourism-related characteristics of the establishments improve the estimation accuracy of regional tourism sales estimates. In this analysis, we estimated the total tourism sales of all establishments based on a random sampling of data of selected establishments from the RTES. Concretely, in this research, we estimated tourism sales using the six methods detailed below:

Method 1

Method 1 applies the average tourism sales of selected establishments as the tourism sales of other establishments. This method assumes that the tourism sales of all non-selected establishments are identical to the average tourism sales of selected establishments.

1. n establishments were randomly selected from those listed in the RTES (specifically, 374 of 739 retail sales establishments and 337 of 364 food service establishments were selected)
2. After calculating whole tourism sales x of selected n establishments, estimated tourism sales for all establishments $\hat{X}$ is calculated as equation (1).

(1) $$\hat{X} = \frac{x \cdot N}{n}$$

N is the number of all establishments from the RTES.

3. By repeating this process 10,000 times to ensure its accuracy, 10,000 of $\hat{X}$ are estimated.

Method 2

Method 2 applies the average percentage of tourism sales of selected establishments as the percentage of tourism sales of other establishments, and assumes that the percentage of tourism sales is the same for all establishments.

1. n establishments are randomly selected, as in method 1.
2. After calculating the average percentage of tourism sales of the selected establishments $\bar{s}$, tourism sales of a non-selected establishment is estimated by multiplying the total sales (including non-tourism sales) by $\bar{s}$.

3. The sum of total tourism sales of selected establishments, as listed in the RTES, and the total tourism sales of non-selected establishments, as estimated in step 2, is the sum of total tourism sales for all establishments, represented by $\hat{X}$.
4. The process is repeated 10,000 times to estimate 10,000 of $\hat{X}$ to ensure the accuracy of the findings.

Method 3

Method 3 applies a logistic regression[4] whose explanatory variable is the total sales of each establishment.

1. n establishments are selected, as in methods 1 and 2.
2. The equation (2) is estimated using the data of the selected establishments.

$$\ln \frac{s_i}{1 - s_i} = \alpha + \beta Y_i \quad (2)$$

s_i and Y_i represent the i th establishment's percentage of tourism sales and total sales respectively. α and β are the parameters. As a type of logistic regression model, equation (2) assumes that the value of an establishment's total sales affects its percentage of tourism sales.

3. By assigning a value of total sales of a non-selected establishment to the estimated equation in the previous step, the percentage of tourism sales for the non-selected establishment can be estimated and the total estimated tourism sales of non-selected establishments calculated.
4. The sum of the actual total tourism sales of selected establishments, obtained from the RTES, and the total tourism sales of non-selected establishments, estimated in step 3, is the total estimated tourism sales for all establishments.
5. Again, the process is repeated 10,000 times, with 10,000 of $\hat{X}$ estimated.

Method 4

Method 4 applies a logistic regression whose explanatory variable is the number of workers of each establishment. It replaces equation (2) found in the method 3 with the equation below:

$$\ln \frac{s_i}{1 - s_i} = \alpha + \beta L_i \quad (3)$$

L_i indicates the i th establishment's number of workers. Equation (3) assumes that

[4] Logistic regression is a type of Generalized Linear Models (GLM). For more on GLM, see Dobson and Barnett (2008).

the establishment's number of workers influences its percentage of tourism sales.

Method 5

Method 5 applies a logistic regression whose explanatory variables are the location information of each establishment calculated by GIS (e.g. the distance to the establishment from tourism attractions, train stations, highway exits, and accommodations). It replaces equation (2) in method 3 with the following equation:

$$\ln \frac{s_i}{1 - s_i} = \alpha + \sum_k \beta_k D_{ki} \tag{4}$$

D_{ki} are variables related to an establishment's distance, as determined by GIS, from tourist attractions, train stations, highway exits, and accommodations. Equation (4) assumes that an establishment's location affects its percentage of tourism sales. The variables are explained in section 4.

Method 6

Method 6 applies a logistic regression whose explanatory variables are location information and the tourism characteristics of each establishment. It replaces equation (2) in method 3 with the following equation:

$$\ln \frac{s_i}{1 - s_i} = \alpha + \sum_k \beta_k D_{ki} + \sum_l \gamma_i T_{li} \tag{5}$$

T_{li} are dummy variables related to an establishment's tourism characteristics. For example, when an establishment appears in a travel guidebook, $T_{li} = 1$, and when it doesn't, $T_{li} = 0$. These variables are explained in section 4.

In this paper, by comparing the estimated tourism sales of these six methods to the actual tourism sales from the RTES, we show that the accuracy of the estimation using methods 5 and 6 are higher than those of methods 1-4. Here, data regarding establishments in the food service and retail trade sectors for Shizuoka Prefecture, as obtained from the RTES, is used.

4. Results of Analysis

4.1 Regression Model of Methods 5-6

Before proceeding, it is necessary to determine which explanatory variables should be introduced in the logistic regression models from methods 5 and 6, as explained in section 3. In this analysis, we estimated regression equations using a variety of explanatory variables and chose the best model.

We assumed that tourists mainly consume near tourist attractions, transportation hubs, and accommodation facilities. Therefore, the distances from such facilities to a food service or retail trade establishment should be one of the most important factors for determining the establishment's percentage of tourism sales. Table 1 describes the explanatory variables related to an establishment's location.

Table 1. Explanatory variables related to an establishment's location

Variable	Content
Tourist Attraction	This variable represents an establishment's proximity to tourists' attractions. Since an establishment may be located near more than one tourist attraction and the number of visitors to each tourist attraction differs, this analysis adopted the sum of $\left(\frac{\text{number of visitors of toursim attraction}}{\text{distance from an establishment}}\right)$ for each tourist attraction as the variable. The expected result is positive.
Shinkansen Station	This variable represents an establishment's proximity to station for the Shinkansen high-speed railway (Bullet Train). The distance from an establishment to the nearest Shinkansen station is used in this analysis. Since the percentage of tourism sales should be higher nearer the station, the expected result in negative.
Train Station	This variable represents an establishment's proximity to a train station operated by Japan Railways other than Shinkansen station. The distance from an establishment to the nearest train station is used this analysis. Like Shinkansen Station, the expected result is negative.
Highway Exit	This variable represents an establishment's proximity to a highway exit. Here, the distance from an establishment to the nearest highway exit is used. The expected result is negative.
Accommodation	This variable represents an establishment's proximity to accommodations. As with tourist attractions, there are often multiple accommodations near an establishment and the size of each accommodation differs. Therefore, this analysis adopted the sum of $\left(\frac{\text{revenue of accommodation}}{\text{distance from an establishment}}\right)$ for each establishment. The expected result is positive.

Table 2 explains the explanatory variables related to the tourism characteristics of an establishment.

Table 2. Explanatory variables related to the tourism characteristics of an establishment

Variable	Content
Tourism Association	This dummy variable takes a value of a 1 if the establishment is a member of a regional tourism association. The expected result is positive.
Regional Booklet	This dummy variable takes a value of 1 if the establishment is listed in a regional tourism booklet published by its regional government. The expected result is positive.
Guidebook	This dummy variable takes a value of 1 if the establishment is listed in a tourism guidebook. The expected result is positive.
Credit Card	This dummy variable takes a value of 1 if the establishment accepts credit cards. The expected result is positive.
Bus Parking	This dummy variable takes a value of 1 if the establishment has a parking area for buses. The expected result is positive.
Website 1	This dummy variable takes a value of 1 if the establishment maintains its own website. The expected result is positive.
Website 2	This dummy variable takes a value of 1 if the establishment is listed on the website of the regional tourism association. The expected result is positive.
Website 3	This dummy variable takes a value of 1 if the establishment is listed on the website of travel agencies. The expected result is positive.

Since the ultimate objective is to estimate the total tourism sales of all regional establishments, the adopted variables should be available for all establishments. GIS using geographic data and the address of an establishment from the Business Register can calculate all of the variables listed in Table 1. With the exception of "Bus Service" for the food service sector, the majority of variables listed in Table 3 can be obtained fairly easily for all establishments using publicly available data.[5] As a result, tourism sales for all establishments can be estimated.

Table 3 shows the estimated results for the adopted models. The parameters were estimated by Maxium-Likelihood (ML) method with the data for all establishments in Shizuoka Prefecture obtained from the RTES.

[5] For the retail trade sector, "Bus Parking" information can be obtained from the Census of Commerce. However, this Census does not cover establishments belonging to the food service sector, meaning that "Bus Parking" information for the food service sector remains a challenge for future research.

Table 3. Estimated results for all establishments in Shizuoka

Sector	Food Service				Retail Trade			
Method	Method 5 (Map)			Method 6 (Map+Tourism)	Method 5 (Map)			Method 6 (Map+Tourism)
Tourist Attraction	0.145	(0.000)	**	0.063 (0.000) **	0.019	(0.000)	**	0.022 (0.000) **
Shinkansen Station	0.036	(0.000)	**	0.054 (0.000) **	0.063	(0.000)	**	0.061 (0.000) **
Train Station	0.112	(0.001)	**	0.068 (0.001) **	0.078	(0.000)	**	0.045 (0.000) **
Highway Exit	-0.069	(0.001)	**	-0.017 (0.001) **	-0.208	(0.001)	**	-0.141 (0.001) **
Accommodation	0.130	(0.002)	**	0.167 (0.002) **	0.696	(0.002)	**	0.719 (0.002) **
Tourism Association				0.144 (0.007) **				0.976 (0.004) **
Regional Booklet				0.290 (0.007) **				0.411 (0.006) **
Guidebook				0.145 (0.006) **				2.045 (0.005) **
Credit Card				0.248 (0.005) **				0.681 (0.004) **
Bus Parking				1.704 (0.006) **				
Website 1				0.182 (0.006) **				
Website 2				0.440 (0.007) **				
Website 3				0.648 (0.007) **				
Degrees of freedom	668			660	743			739
Likelihood Ratio Index	0.385			0.605	0.436			0.577

Notes: 1. The figures in parentheses are Standard Errors.
2. Asterisks * and ** indicate that the corresponding coefficients are significant at the 5% and 1% level, respectively.
3. Constant terms are included, but estimated coefficients are not reported.

As shown in table 3, all of the adopted variables for all models were significant at the 1% level. Although the percentage of an establishment's tourism sales was presumed to be high when located near a Shinkansen or train station, the estimated parameters were positive; a result that differed greatly from our original expectations. Since many commuters and local consumers congregate at stations, the closer establishment's proximity to a station, the higher the percentage of non-tourism consumption.

Applying method 6 to the retail trade sector, we attempted to estimate the same formula as method 6 for the food services sector. However, all coefficients for "Bus Parking" and "Website 1," "2," and "3" were negative, meaning that the percentages of tourism sales are low for establishments with bus parking and websites. Since there is no logical explanation to explain such a result, these variables were removed when applying Method 6 to the retail trade sector.

4.2 Estimating Tourism Sales

Using methods 1 through 6 explained in section 3, we estimated tourism sales and compared the estimation accuracy of the results.

There are two important evaluative criteria for assessing the accuracy of the estimates. The first is whether or not the estimator is biased. If the expected value of $\hat{X}$ is identical to the total tourism sales X obtained from the RTES, $\hat{X}$ is considered an unbiased estimator. The other

criterion is efficiency; with unbiased estimators, the smaller the variance of $\hat{X}$, the higher its efficiency. Some of the indices for this estimation are outlined below.

First, the error of estimation is defined as:

(6) $$\hat{e}_{k,i} = \hat{X}_{k,i} - X \quad (k = 1, \cdots, 6, i = 1, \cdots, 10000)$$

$\hat{X}_{k,i}$ represents the estimated tourism sales (represented by $\hat{X}$ in the previous section), calculated as the i th calculation of 10,000 times in method k, and $\hat{e}_{k,i}$ (represented by $\hat{e}$ in the previous section) represents the error of estimation. X, which is the total tourism sales of all establishments from the RTES, is the estimation target. If $\hat{X}_{k,i}$ and X are almost equal, then $\hat{e}_{k,i}$ will be small, and the estimation accuracy can be considered high.

The average of $\hat{e}_{k,i}$ is calculated as follows:

(7) $$\overline{\hat{e}_k} = \frac{1}{N}\sum_{i=1}^{N} \hat{e}_{k,i} \quad (k = 1, \cdots, 6)$$

$N = 10{,}000$ since the calculation is conducted 10,000 times. If the estimation result is completely unbiased, $\overline{\hat{e}_k} = 0$.

To evaluate efficiency, the standard deviation of $\hat{e}_{k,i}$ is calculated as:

(8) $$S_k^e = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(\hat{e}_{i,k} - \overline{\hat{e}_k}\right)^2} \quad (i = 1, \cdots, 6)$$

Suffice it to say, S_k^e equals the standard deviation of $\hat{X}_{k,i}$. If $\overline{\hat{e}_k} = 0$ and S_k^e is smaller, then the method is considered more efficient, since the probability of obtaining an estimated value close to the target value X is high.

If we compare a case in which the absolute value of $\overline{\hat{e}_k}$ is large and S_k^e is small with a case in which the absolute value of $\overline{\hat{e}_k}$ is small and S_k^e is large, the probability of obtaining a value close to the target value X may be higher for the former case. Since the estimation results must be evaluated in terms of bias and efficiency, in this paper, the average of the absolute value of $\hat{e}_{k,i}$ is defined as $\overline{|\hat{e}_k|}$ below:

(9) $$\overline{|\hat{e}_k|} = \frac{1}{N}\sum_{i=1}^{N}|\hat{e}_{k,i}| \quad (i = 1, \cdots, 6)$$

$\overline{|\hat{e}_k|}$ represents the average range of the error of estimation for each method. It can be said that the smaller the value for $\overline{|\hat{e}_k|}$, the more favorable the method.

Table 4 and 5 reveal the estimation results for the food service and retail sectors

Table 4. Result for the Food Service sector

	Method1	Method2	Method3	Method4	Method5	Method6
$\overline{\hat{e}_k}$	-1,001	-104,097	30,373	-4,556	-10,851	5,234
$\overline{\hat{e}_k}/X$	-0.00	-0.23	0.07	-0.01	-0.02	0.01
S_k^e	92,023	46,953	95,170	75,315	48,197	38,653
$\overline{\lvert\hat{e}_k\rvert}$	76,816	104,255	81,852	61,160	38,266	30,569
$\overline{\lvert\hat{e}_k\rvert}/X$	0.17	0.23	0.18	0.14	0.09	0.07

Table 5. Result for the Retail Trade sector

	Method1	Method2	Method3	Method4	Method5	Method6
$\overline{\hat{e}_k}$	240	93,076	278	1,550	29,537	25,142
$\overline{\hat{e}_k}/X$	0.00	0.10	0.00	0.00	0.03	0.03
S_k^e	176,203	135,162	175,575	172,890	119,666	103,720
$\overline{\lvert\hat{e}_k\rvert}$	145,964	131,396	143,380	141,946	96,116	82,174
$\overline{\lvert\hat{e}_k\rvert}/X$	0.16	0.14	0.16	0.16	0.11	0.09

The first row of tables 4 and 5 indicates the average error of estimation $\overline{\hat{e}_k}$ and the second row shows the ratio of $\overline{\hat{e}_k}$ to the target value X. The results from method 1 are closest to 0 for both sectors. Such a result is reasonable, as method 1 applied the sample average of tourism sales of the selected establishments to the non-selected establishments and the sample average is, by definition, an unbiased estimator.

$\overline{\hat{e}_k}$ of method 2 was the farther from 0 for both sectors. Method 2 applied the sample average of the percentage of tourism sales of the selected establishments to the non-selected establishments and calculated tourism sales by multiplying total sales by the average percentage. Therefore, although the average percentage of tourism sales is an unbiased estimator, the estimated tourism sales can be biased when the percentage of tourism sales correlates to total sales. In fact, the correlation coefficients for percentage of tourism sales and total sales are 0.181 for food service, which is consistent with the result of $\overline{\hat{e}_k}$ and $\overline{\hat{e}_k}/X$ for method 2 in table 2.

Methods 3-6 used the result of logistic regression estimated using the ML method. Since the ML estimator is an asymptotic unbiased estimator, the difference between $\overline{\hat{e}_k}$ and 0 for these methods is not too large. $\overline{\hat{e}_k}/X$ did not exceed 0.03 except when method 3 was applied to the food service sector.

For S_k^e, as shown in the third row of the tables, method 6 shows the smallest values in both tables, suggest that the method is most favorable with regard to variance. S_5^e is also smaller than methods 1, 3, and 4.

The average of absolute value $\hat{e}_{k,i}$ is shown in the fourth row of the tables, while the ratio of $\overline{|\hat{e}_k|}$ to the target value X is shown in the fifth. For these values, methods 5 and 6 produced extremely small values compared to the other method, making them the most accurate estimates.

Figures 1 and 2 show the distribution of estimated tourism sales, $\hat{X}_{k,i}$, for the food service and retail trade industries respectively.

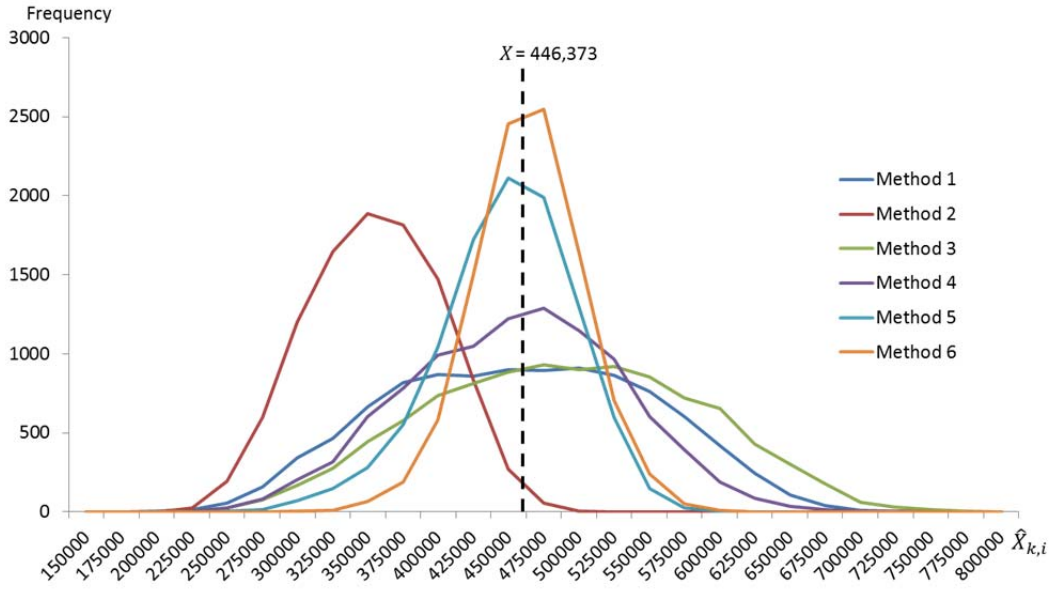

Figure 1. Distribution of $\hat{X}_{k,i}$ for Food Service

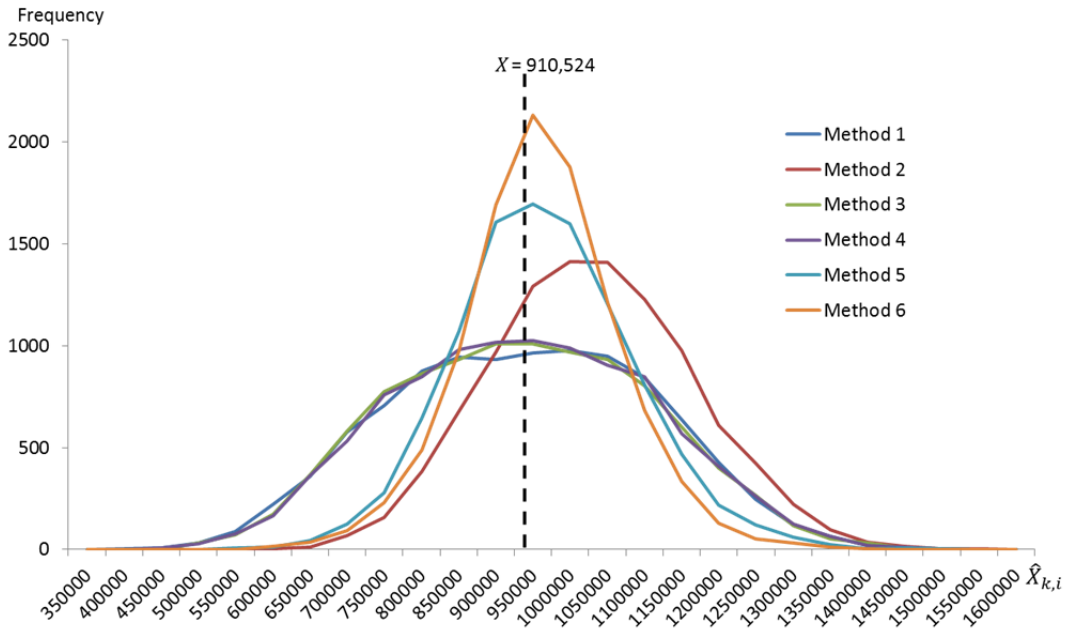

Figure 2. Distribution of $\hat{X}_{k,i}$ for Retail Trade

These figures indicate that the centers of the distributions are almost identical to the target value X, except for the case of method 3. The variances of distributions were smallest for methods 5 and 6.

Given these results, it can be concluded that methods 5 and 6 are the best methods for estimating tourism sales, meaning that geographic data and tourism characteristics can play an important role in improving the accuracy of these estimates.

5. Conclusions

Our research reveals that factoring in an establishment's location and its characteristics can improve the accuracy of existing tourism sales estimates. Since this study only utilized data from the RTES, our next challenge is to estimate regional tourism sales by introducing data from the Business Register, the registers of tourism associations, and other sources (e.g. guidebooks and websites) to the model outlined in this study.

Here, the Business Register is an especially valuable resource as it provides information about the entirety of an establishment's operations. Although the Register rarely contains information relating to a specific sector (such as tourism), our research has indicated that basic information from the Business Register can improve the estimation accuracy of specific values (in this case, regional tourism sales). From this perspective, this research is also a significant development in how the Business Register can be used in economic analysis.

Since Japan's Business Register has only been compiled for two years, it does not presently include the latitude and longitude of establishments. However, our findings suggest that this information may prove useful in the future and that the Japanese government should consider collecting this information moving forward.

Additionally, this study demonstrates the shortcomings of estimating tourism sales using only demand-side tourism statistics. The tourism characteristics of establishments can play an important role in improving the estimation accuracy of tourism sales, and these characteristics can only be considered in supply-side tourism statistics such as the RTES.

In the future, once we further improve the geographic data and the estimation models, we hope to more accurately estimate regional tourism sales, which is essential for compiling regional Tourism Satellite Accounts.

Reference

Dobson, Annette J. and Adrian Barnett (2008) *An Introduction to Generalized Linear Models*, Third Edition, Chapman&Hall.

Miyagawa (2013) “供給サイド統計調査による地域観光規模の把握に関する一考察” (in Japanese), *Bulletin of Japan Statistics Research Institute*, No.42, Japan Statistics Research Institute.

Miyagawa, Kozo, Hiroyuki Kamiyama, Yoshihito Sakuramoto and Ryuta Shimamura (2012) “Compiling a Regional Tourism Satellite Account using the Regional Tourism Economic Survey and Geographic Information System”, presented at the 11th Global Forum on Tourism Statistics. (Reykjavík, Iceland, November 14th-16th, 2012)