IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol.004, LIR281
DOI: https://doi.org/10.52731/lir.v004.281

Proposed Analytical Process for More Convenient Utilization of Open Data - Verification Using Tourist Number Data -

Soh Sakurai [*], Noriko Shibata [†],
Akira Nagamatsu [‡]

Abstract

This study proposes a statistical analysis process for the effective and convenient utilization of open data. Considering the current state of open data use in Japan and other countries, it focuses on the challenge where maximizing data utility is heavily dependent on user skills. The process is validated using easily accessible prefectural tourist data, which is rich in academic research. By applying principal component analysis and regression analysis, the study defines a specific model and proposes a process aimed at enabling more practical and straightforward applications of open data.

Keywords: open data, tourism, regression analysis, principal component analysis.

1 Introduction

In recent years, the use of open data has gained attention across various sectors of society. Many countries have emphasized the importance of open data, with governments initiating policies to promote the openness of data. For example, in the United States, the Open Government Initiative, launched in 2009, has accelerated this trend [1]. Similarly, in Japan, the "Declaration of Creation of the World's Most Advanced Digital Nation" was adopted by the Cabinet on June 14, 2013, as a new IT strategy, increasing both the openness of data and related academic research.

Progress has been made in Japan in improving access to open data, for instance, through portals like the "e-gov portal," which allows access to open data from governments and municipalities. Efforts are not only technological but also educational, such as the Ministry of Internal Affairs and Communications' project to develop open data training for local government officials.

However, Japan's initiatives have been in place for about a decade, and the situation regarding the utilization and promotion of open data is still developing. Factors or challenges making this difficult include the complexity of data collection, management, and use; human and organizational issues; and maximizing the value derived from data use [2]. Even with efforts to address these challenges, resources such as specialized knowledge in data utilization and funding are typically limited and constrained by organizational scale. To broaden the scope of use, standardization is necessary to make open data more easily usable.

In this research, as part of these efforts, we propose a statistical analysis process for more

[*] Chiba University of Commerce, Chiba, Japan
[†] Yokohama City University, Kanagawa, Japan
[‡] Graduate School of Engineering, Tohoku University, Miyagi, Japan

convenient and effective utilization of open data, using experimentally collected tourist number data by prefecture, which is relatively comprehensive and easily accessible.

2 Review of Existing Research

2.1 Research from the Perspective of Open Data Providers

Research on open data can be viewed from two perspectives: that of the data providers and that of the data users.

A significant research theme from the providers' side is the quality of open data [3] [4]. Data quality refers to "data that are fit for use by data consumers"[5]. The Office for IT General Strategy, Cabinet Office, Japan, also emphasizes the significant impact of using databases with quality issues in its "Data Quality Management Guidebook (Beta Version)" available on its website.

To manage data quality comprehensively, the concept of Total Data Quality Management was proposed in the 1990s [6]. This concept not only includes data quality from the provider's perspective, such as the reliability and accuracy of data, but also considers Information Quality, which is essential from the user's perspective. Information Quality refers to “fitness for use by Information consumers” [6]. In other words, providing data that meets the needs of users is also a crucial aspect of quality. However, as open data users and their purposes are diverse, particularly because open data is not collected for specific research purposes like survey data, issues often arise when data needed for hypothesis testing are scattered across multiple open datasets.

Maintaining and improving data and information quality requires robust data governance. With various stakeholders involved in open data, data governance is particularly critical concerning the quality of data for academic research [7]. Open data governance refers to the policies and processes concerning the management, regulation, and use of open data, including policies addressing privacy and security concerns [8]. Even aggregated open data pose risks of privacy and security breaches, as noted by the Cybersecurity and Infrastructure Security Agency (CISA), which has published documents discussing the security challenges and business impacts of aggregating large amounts of data and the corresponding countermeasures [9].

While open data is not necessarily big data, it can be considered a specific example of big data. There is a study on security and privacy issues that occur throughout the entire lifecycle of big data [10]. This study focuses on security and reliability issues at each stage from data collection to disposal, the importance of protecting personal information, and the infringement of individual privacy. Additionally, they analyze the current state of and research related to international standards for big data security and privacy protection.

2.2 Research from the Perspective of Open Data Users

Another area of focus is research from the perspective of open data users, which includes numerous studies on statistical methods, including data mining. Open data is published and utilized across various fields, and its use is expanding in the social sciences, natural sciences, and as well as humanities [11].

The collection of examples of open data use has been increasing yearly. In December 2009, the U.S. government issued the "Open Government Directive," mandating increased data transparency and accessibility. This has led to the publication and utilization of data across a wide range of areas including environment, health, energy, and education, enhancing the user environment. The U.S. government has also been providing technical support, such as API provision assistance, early on to promote the use of data.

In Japan, the "Basic Act on the Advancement of Utilizing Public and Private Sector Data" (Act No. 103 of 2016) mandates that both national and local governments engage with open data. This is expected to solve various issues through citizen participation and public-private collaboration, stimulate the economy, and enhance and streamline government administration. In December 2017, to promote the publication and use of open data, the government compiled "Recommended Data Sets (renamed 'Local Government Standard Open Data Sets' in December 2023)," outlining the data to be published by the government and the rules and formats to adhere to [12].

Government-led efforts to enhance the open data utilization environment are progressing, and portal sites like e-Gov and e-Stat are seeing increased actual usage, making open data more accessible. Local governments are leading the increase in data utilization contests.

The development of open data in the tourism sector is particularly advanced compared to other fields. The tourism industry emphasizes data-driven services. For instance, Germany has established the Open Data Tourism Alliance as part of its open data strategy, reflecting the progression of national-level open data projects [13]. This has also advanced the standardization of tourist information and the development of knowledge graphs. In Europe, the utilization of open data is being promoted through data stories, use cases, and various studies, particularly increasing the demand for applications and online communities related to travel planning [14]. Furthermore, the use of AI is transforming the tourism industry by improving customer service, enhancing operational efficiency, providing personalized travel experiences, and supporting sustainability initiatives [15].

In this way, open data in the tourism sector is more developed than in other sectors in terms of data volume, ease of access, use cases, and academic research. The digitalization and publication of data mutually enhance each other, contributing to the promotion of new business models and innovations.

In Japan, especially in the tourism sector, open data utilization is one of the more active areas. Local governments and tourist associations publish information on tourist spots, events, accommodation, and transportation access, and are advancing the provision of multilingual information in anticipation of increasing foreign tourists.

However, there are variations in the quality and update frequency of government data and the publication of data by local governments in Japan, and the extent of utilization varies by region. The use of data by private companies and the ordinary citizens is still relatively limited compared to other countries. Additionally, in Japan, research utilizing tourism-related open data often employs relatively sophisticated statistical methods, posing problems in terms of practical usability. There is not much research on methods to simplify the use of open data for a wider range of users.

Therefore, this research proposes an analytical process using tourist number data to make the utilization of open data more convenient.

3 Analytical Process and Model

3.1 The Analytical Process When Using Open Data

When utilizing variables for statistical analysis from open data, there are generally two approaches. One approach focuses solely on the variable of interest for analysis. In this case, calculating basic statistical measures like mean, standard deviation, or frequency distribution and applying visual techniques such as pie charts or line graphs would be appropriate. The other approach involves examining the relationship between the variable of interest and other variables that may have a causal or covariant relationship with it. For this, multivariate analysis techniques or data mining methods are used. This research proposes the latter analytical process.

Figure 1 illustrates the analytical process proposed in this study. The steps include: 1) identifying business challenges that need to be resolved, 2) formulating research questions to determine what information is necessary to address these challenges, 3) developing hypotheses and statistical verification methods, 4) collecting data, 5) considering whether adjustments to the analysis methods are necessary, 6) testing the hypotheses, and 7) interpreting the analysis results to derive implications for business challenges.

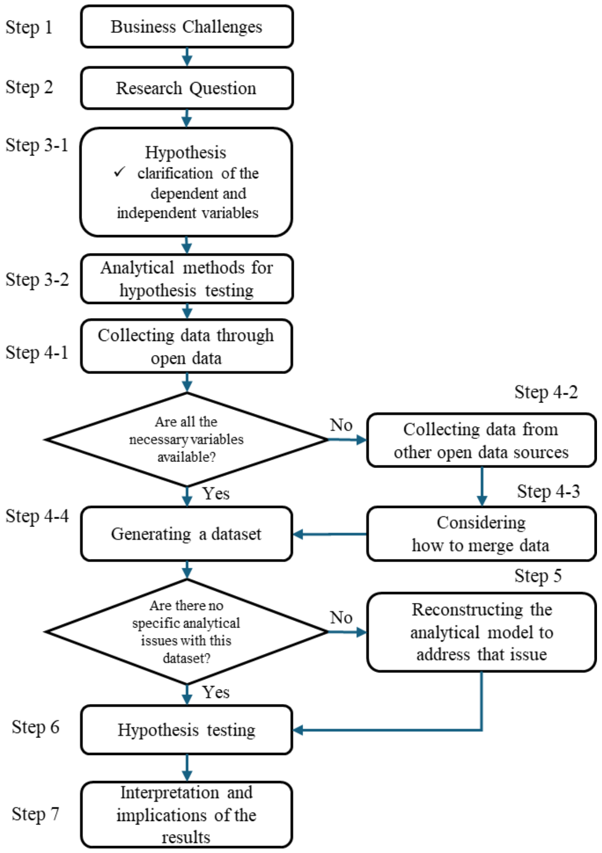
Step 1
Business Challenges
Step 2
Research Question
Step 3-1
Hypothesis
✓ clarification of the dependent and independent variables
Step 3-2
Analytical methods for hypothesis testing
Step 4-1
Collecting data through open data
Are all the necessary variables available?
No
Step 4-2
Collecting data from other open data sources
Yes
Step 4-3
Considering how to merge data
Step 4-4
Generating a dataset
Step 5
Are there no specific analytical issues with this dataset?
No
Reconstructing the analytical model to address that issue
Yes
Step 6
Hypothesis testing
Step 7
Interpretation and implications of the results

Figure 1: Analytical Process Using Open Data

A unique aspect of open data is that the dependent variable and independent variables may be collected from different sources, as accounted for in Step 4-2 of the process. When generating a dataset from multiple data sources, consider the methods for merging data as shown in Step 4-3. Identify the key variables for merging and check the aggregation level. If the aggregation levels differ, it is necessary to decide which level to align with. Align with the coarser level of aggregation. For example, if Data A is aggregated on a quarterly basis and Data B on a monthly basis, then re-aggregate the latter to a quarterly basis before merging it with Data A.

Unlike experimental data, open data is not measured in a controlled environment for independent variables. This may necessitate considering issues unique to the data set created for analysis, such as strong correlations between independent variables, and modifying the statistical model accordingly. Incorporating Step 5 to address this issue is another distinctive feature of this study. This research uses open data from the tourism sector. The next section will formalize the statistical model for Step 5 in this research.

3.2 A General Model for Analyzing Tourist Number Open Data

Consider the total number of tourists in an area such as a prefecture as the dependent variable y, and assume that the total number of tourists is a linear function of tourism resources and environmental factors. That is, using a linear multiple regression model. Let y_t be the dependent variable at time t, x_i (where $i = 1,2,\cdots,p$) be the available explanatory variables, α be the constant term, β_i be the regression coefficient for the i-th explanatory variable, and ε_t be the residual. The multiple regression model is as follows:

$$y_t = \alpha + \sum_{i}^{p} \beta_i x_i + \varepsilon_t \cdots (1)$$

Tourists often visit several attractions to gain more satisfaction from a single trip, which might suggest a pattern of visiting multiple tourism resources [16] [17] [18]. Understanding these touring patterns holds practical significance, such as for recommendations on a tourism department's website or issuing coupons that meet user needs. This indicates that explanatory variables in Equation (1) may be correlated, potentially leading to multicollinearity. There are two ways to handle this: one is to use only one of the highly correlated variables, which may obscure the specific effects of the variables that were omitted. Practitioners, such as staff in the tourism department, might wish to understand the impact of all variables, which leads to the second method: compressing highly correlated variables to a few variables. This study employs principal component analysis to perform regression analysis on compressed explanatory variables [19]. Principal component analysis is widely available in standard statistical software like SPSS.

Suppose the first I variables x_i (where $i = 1,2,\cdots,I$) are compressed to a few principal components using principal component analysis. The derivation of principal components in this study uses correlation coefficients, which is automatically performed in software, but the extraction of principal components involves standardizing each variable to have a mean of 0 and variance of 1. The analysis was conducted using SPSS ver.27.

$$x_i' = \frac{x_i - \bar{x}_i}{s_{x_i}}$$

Where s_{x_i} is the unbiased estimator of the standard deviation of x_i. Let z_j (where $j = 1,2,\cdots,J$) be the j-th principal component, then the calculation formula for that principal component score is:

$$z_j = \sum_{i=1}^{I} \gamma_{ji} x_i' \cdots (2)$$

Here, $\sum_i \gamma_{ji}^2 \leq 1$, and each z_j has a mean of 0 and a variance of 1, being independent from each other.

The dependent variable y_t ranges over $t = 1,2,\cdots,T$, the explanatory variables z_j over $j = 1,2,\cdots,J$, and the other non-compressed explanatory variables x_i over $i = I+1, I+2,\cdots,p$. We redefine equation (1) as follows, which forms the basis of our model:

$$y_t = \alpha + \sum_{j=1}^{J} \beta_j z_j + \sum_{i=I+1}^{p} \eta_i x_i + \varepsilon_t \cdots (3)$$

In this equation, β_j and η_i are regression coefficients. Interpreting η_i is straightforward, but interpreting β_j requires caution because it represents the regression coefficient of the j-th principal component, which affects the interpretation of the original variables $x'_i(i = 1,2,\cdots,I)$ used to extract the principal component.

Expanding the second term on the right side of equation (3) using equation (2), we get:

$$\sum\nolimits_j^J \beta_j z_j = \sum\nolimits_j^J \beta_j \sum\nolimits_i^I \gamma_{ji} x'_i = \sum\nolimits_i^I \sum\nolimits_j^J \beta_j \gamma_{ji} x'_i$$

Thus, the impact on the dependent variable when the i-th variable x'_i changes by one unit is given by $\sum_j^J \beta_j \gamma_{ji}$. Since x'_i is standardized, if you want to assess the impact of the original variable x_i before standardization, you should divide by s_{x_i}.

4 Implementation

To succinctly summarize Steps 1 to 3-1 envisioned in this study: the business challenges in Step 1 is "increasing the number of tourists," the research question in Step 2 is "why do tourists visit that area," and the hypothesis in Step 3-1 is "the number of tourists is influenced by tourism resources and environmental factors." These are standard assumptions but hold significant practical importance. In Step 3-2, the hypothesis is addressed by considering the number of tourists as a function of tourism resources and environmental factors, employing a linear regression model.

4.1 Open Data – Tourism Statistics from the Japan Tourism Agency Website

This section corresponds to Steps 4-1 to 4-4 from Figure 1.

4.1.1 Data Usage Period

In this study, open data published on the "Standardized Tourism Statistics" page of the Japan Tourism Agency's website was downloaded to generate the analysis dataset. Although data from 2010 onwards is available, this study used open data from 2014 to 2019. Since the Basic Act on the Advancement of Utilizing Public and Private Sector Data, which mandates the engagement with open data by national and local public bodies, was enacted in 2016, data from 2016 onwards is preferable for accuracy. However, due to the peculiarities of the data post-2020 caused by the COVID-19 pandemic, this period was excluded from the model estimation dataset. Estimations are possible for the four years from 2016 to 2019, but it would not be appropriate to verify predictive power using data beyond 2020.

The focus on open data as part of a national strategy was notably reinforced with the revised "Declaration on the Creation of the World's Most Advanced IT Nation" in June 2015, first announced in June 2013[20]. Therefore, the groundwork for the measurement and collection of open data was deemed to be well-established by 2014, which was selected as the starting year for the dataset, using data through 2019. To validate the accuracy of the analysis model, the data was split into two segments: data from 2014 to 2017 for model estimation, and data from 2018 to 2019 for testing predictive power.

4.1.2 Characteristics of the Utilized Open Data and Selection of Dependent and Independent Variables

The "Standardized Tourism Statistics" are published annually in Excel files, quartered. Each file represents one year, containing data for four quarters. Each Excel file has 7 sheets, excluding the index (Table 1).

Table 1: Sheets in the Excel File

Sheet Number	Content
1	Number of tourists by prefecture, unit price of tourism spending, total tourism spending (Japanese, tourism purpose)
2	Number of tourists by prefecture, unit price of tourism spending, total tourism spending (Japanese, business purpose)
3	Number of tourists by prefecture, unit price of tourism spending, total tourism spending (foreign visitors)
4	Number of tourism sites and events by prefecture
5	Number of tourists by tourism site and event by prefecture
6	Survey points for tourism site parameters by prefecture
7	Results of tourism site parameter surveys by prefecture

For the dependent variable total number of tourists, data on the number of Japanese tourists for tourism purposes from sheet 1 was selected. Data on the number of foreign visitors from sheet 3 could also be suitable as a dependent variable, but it was excluded in this instance.

The characteristics required for the independent variables to test the hypothesized relationships are: 1) they should represent the richness or operational status of tourism resources in that prefecture, and 2) they should be considered environmental factors. Data meeting the first criterion could be identified in sheets 5. There were no sheets collecting data corresponding to the second criterion. Although it would be better to perform Step 4-2 from Figure 1, this step was not executed in this study for simplification.

The data presented in Sheet 5 represents the number of tourists measured in the previous year [21]. For example, the data for the second quarter of 2019 on Sheet 5 is from the second quarter of 2018. Although it does not directly represent the operational status of each tourist site for that year, it serves as a proxy variable. Therefore, this study decided to use the data listed in Sheet 5 as independent variables.

The specific content of the dependent and independent variables used from Sheets 1 and 5 is explained below.

4.1.3 Specific Details of Dependent and Independent Variables

Sheet 1 under Table 1 lists four types of data concerning the number of Japanese tourists visiting for tourism purposes. These are categorized into tourists from within the prefecture and those from outside, further divided into overnight and day-trip visitors. Thus, there are four types of tourist classifications: "overnight visitors from outside the prefecture," "overnight visitors from within the prefecture," "day-trip visitors from outside the prefecture," and "day-trip visitors from within the prefecture." Of these, "overnight visitors from outside the prefecture" was used as the dependent variable.

Sheet 5 includes seven categories of tourist sites where visitor numbers are recorded: "Natural," "Historical & Cultural," "Hot Springs & Health," "Sports & Recreation," "Urban Tourism," "Others," and "Festivals & Events." The specific contents of these categories are detailed in Table 2[21].

Table 2: Detailed Contents of Visitor Statistics

Category	Content
Natural	Mountains, plateaus, lakes, rivers, seas, underwater, islands, other natural sites (including green tourism)
Historical & Cultural	Historical sites, castles, shrines and temples, gardens, historic towns, old roads, museums, art galleries, memorial museums, zoological and botanical gardens, aquariums, industrial tourism (e.g., wineries, brewery tours), and other historical buildings
Hot Springs & Health	Hot spring areas (treating the entire accommodation and hot spring facilities within an area named "XX Hot Spring" as one site), other hot spring and health facilities not regulated by the Hot Spring Law
Sports & Recreation	Sports and recreation facilities (excluding sports spectating), ski resorts, campgrounds, fishing spots, beaches, marinas and yacht harbors, parks, amusement parks, theme parks, and other sports and recreation sites
Urban Tourism	Commercial facilities, districts and shopping streets, food and gourmet, other urban tourism (including direct sales outlets for agricultural and seafood products, product halls)
Others	Other unclassified tourism sites (roadside stations, parking areas, etc.)
Festivals & Events	Local festivals, cherry blossom viewing, New Year's visits, firework displays, local performing arts, local customs, expos, concerts, sports spectating, film festivals, conventions and international conferences, and other unclassified festivals and events

This study selected six out of these categories, excluding "Others," as independent variables. Generally, to maximize satisfaction from a trip, tourists visit multiple tourist spots [16] [17] [18], suggesting a potential for high multicollinearity among these six independent variables. These were compressed by principal component analysis and used as independent variables.

4.2 Re-formulation of the Analysis Model

This section corresponds to Step 5.

The independent variables used in this study are considered as tourism resources likely to be toured, and thus all are compressed for use. Therefore, there are no variables corresponding to the third term on the right-hand side of the basic model formula (3).

In regression analysis and other inferential statistics, it is essential to carefully define the underlying population. The population is defined in terms of elements, sampling units, scope, and time dimensions [22] [23]. In this case, the elements are people visiting the prefecture for tourism, the sampling unit is individuals, and the time is from the first quarter of 2014 to the fourth quarter of 2019. However, the geographic scope of the population is defined as each prefecture, assuming that each prefecture belongs to a different population.

Japan has 47 prefectures. Tourism statistics open data is published for 46 prefectures, excluding Osaka, which does not participate. It covers almost all of Japan, so it would be possible to consider the geographical range of the population as the entire country. However, practically speaking, each prefecture has vastly different characteristics in terms of the amount and appeal of tourism resources, population size, area, and industrial base. Although there is a national organization like the Japan Tourism Agency, specific tourism strategies and tactics are formulated at the local tourism department level, making it impractical to assume that each prefecture belongs to the same population. Therefore, analysis using formula (3) is conducted on a prefectural basis, estimating the model several times over, once for each prefecture.

Based on the above, formula (3) is redefined as follows:

$$y_t^{(k)} = \alpha^{(k)} + \sum\nolimits_{j=1}^{J^{(k)}} \beta_j^{(k)}{}_j z_j^{(k)} + \varepsilon_t^{(k)} \cdots (4)$$

Where k is the prefectural identifier. If the estimated values of the regression coefficients $\beta_j^{(k)}$ are significant, it indicates support for the hypothesized relationships in Step 3-1.

5 Analysis Results

In this study, open data from the first quarter of 2014 to the fourth quarter of 2019, covering 24 quarters, was downloaded to create the dataset. Of this, data from 16 quarters up to the fourth quarter of 2017 were used for model estimation, and the remaining 8 quarters were used for testing predictive power.

During the model estimation period, out of 47 prefectures, 39 had no missing values, and 5 had data entries for more than 50% (12 to fewer than 24 quarters) and were considered valid for analysis (Table 3). The prefectures that were not included in the analysis were Ishikawa, Okinawa, and Osaka, which did not participate.

The prediction power testing used two years of data, which is minimal due to quarterly data, hence up to 8 quarters. For this testing, 32 prefectures with a 100% data completeness rate were analyzed. Newly excluded were 12 prefectures including Hokkaido, Tochigi, Tokyo, Shizuoka, Mie, Kyoto, Hyogo, Wakayama, Tottori, Shimane, Kochi, and Miyazaki.

Table 3: Checking Missing Data

Data Completeness	Model Estimation Period (2014–2017)	Predictive Power Testing Period (2018–2019)
100%	39	32
75% to less than 100%	4	0
50% to less than 75%	1	4
Greater than 0% to less than 50%	2	2
0%	1	9

Note: The numbers in the table represent the count of prefectures.

5.1 Results of the Principal Component Analysis

The number of principal components extracted was decided based on the eigenvalue of the correlation matrix of independent variables being above 1. Since the analysis was done by prefecture, the number of principal components extracted varies by prefecture. Eight prefectures extracted up to three principal components, 29 had two, and 7 had one.

Similarly, the loadings of each variable on the principal components also vary, so the interpretation of each principal component differs by prefecture. For instance, Niigata, Nagano, and Okayama prefectures all have the same number of principal components, but the loadings vary considerably, leading to different interpretations (Table 4).

If interpreted using a relatively high threshold of |0.6| for the loadings, in Niigata Prefecture, the first principal component is made up of "Natural," "Historical & Cultural," and "Urban Tourism," and the second component includes "Sports & Recreation," "Hot Springs & Health," and "Festivals & Events." In contrast, in Nagano Prefecture, "Natural," "Hot Springs & Health," and "Urban Tourism" make up the first principal component, and "Sports & Recreation" the second. Similarly, in Okayama Prefecture, "Hot Springs & Health" makes up the second component, with the remaining five variables comprising the first.

Table 4: Principal Component Matrix by Prefecture

Prefecture	Niigata		Nagano		Okayama	
Variable	First Principal Component	Second Principal Component	First Principal Component	Second Principal Component	First Principal Component	Second Principal Component
Natural	0.975	-0.134	0.904	0.139	0.938	0.180
History & Culture	0.974	0.143	0.384	-0.698	0.766	-0.577
Hot Springs & Health	0.114	0.855	0.900	0.310	0.545	0.792
Sports & Recreation	-0.354	0.914	-0.341	0.757	0.878	-0.277
Urban Tourism	0.970	-0.144	0.967	0.017	0.888	0.127
Festivals & Events	0.590	0.605	-0.641	-0.166	0.749	-0.038

5.2 Results of Model Estimation

Table 5 summarizes the estimated results using formula (4). The prefectural identifier k is assigned from north to south across Japan. The dependent variable $y_t^{(k)}$ represents "overnight guests from outside the prefecture," referred to hereafter as the "non-resident overnight guest model." The table header's R^2 is the coefficient of determination, adj R^2 is the adjusted coefficient of determination, and DW is the Durbin-Watson ratio.

The average coefficient of determination for the non-resident overnight guest model across the 44 analyzed prefectures was 0.505. Table 6 is a frequency distribution table for the coefficient of determination. The coefficient of determination was above 0.8 for 9 prefectures, constituting approximately 20.5%, and between 0.6 and 0.8 for 13 prefectures, about 29.5%. The cumulative frequency for coefficients above 0.6 was exactly half, 22 prefectures or 50.0%. There was considerable variation in the coefficient of determination across prefectures, indicating that while the results are not outstanding, but they are still very good.

Table 5: Coefficients of Determination for the Non-Resident Overnight Guest Model

k	prefecture	R^2	$adj\ R^2$	DW	k	prefecture	R^2	$adj\ R^2$	DW
1	Hokkaido	0.784	0.750	2.657	25	Shiga	0.803	0.753	1.622
2	Aomori	0.831	0.819	1.836	26	Kyoto	0.224	0.031	2.847
3	Iwate	0.803	0.773	2.039	27	Oosaka	N/A	N/A	N/A
4	Miyagi	0.646	0.591	3.089	28	Hyogo	0.295	0.178	2.637
5	Akita	0.812	0.799	2.885	29	Nara	0.812	0.765	1.513
6	Yamagata	0.644	0.619	2.396	30	Wakayama	0.668	0.617	2.104
7	Fukushima	0.581	0.476	1.770	31	Tottori	0.711	0.667	1.724
8	Ibaraki	0.225	0.032	1.413	32	Shimane	0.678	0.629	1.485
9	Tochigi	0.529	0.457	2.169	33	Okayama	0.858	0.836	2.442
10	Gunma	0.777	0.743	1.129	34	Hiroshima	0.236	0.118	1.856
11	Saitama	0.653	0.566	2.130	35	Yamaguchi	0.615	0.556	3.009
12	Chiba	0.041	-0.107	1.033	36	Tokushima	0.307	0.200	1.537
13	Tokyo	0.240	0.123	2.553	37	Kagawa	0.579	0.474	0.909
14	Kanagawa	0.014	-0.138	2.083	38	Aichi	0.222	0.102	0.799
15	Niigata	0.848	0.825	1.246	39	Kochi	0.643	0.563	1.854
16	Toyama	0.601	0.540	2.979	40	Fukuoka	0.474	0.394	1.333
17	Ishikawa	N/A	N/A	N/A	41	Saga	0.020	-0.050	1.738
18	Fukui	0.442	0.219	2.496	42	Nagasaki	0.357	0.258	2.209
19	Yamanashi	0.917	0.912	1.348	43	Kumamoto	0.710	0.665	1.703
20	Nagano	0.830	0.804	1.843	44	Ooita	0.176	-0.031	1.531
21	Gifu	0.182	0.123	2.908	45	Miyazaki	0.026	-0.124	1.900
22	Shizuoka	0.540	0.438	1.447	46	Kagoshima	0.146	0.075	2.901
23	Aichi	0.066	-0.078	2.227	47	Okinawa	N/A	N/A	N/A
24	Mie	0.667	0.616	1.502					

Table 6: Frequency Distribution Table of Coefficients of Determination by Model

Coefficient	Frequency	Cumulative	Relative	Cumulative
Range			Frequency	
[.8,1.0)	9	9	20.5%	20.5%
[.6, .8)	13	22	29.5%	50.0%
[.4, .6)	6	28	13.6%	63.6%
(0, .4)	16	44	36.4%	100.0%

Table 7 includes the prefectures that rank in the top ten based on the coefficient of determination. Yamanashi Prefecture achieved the best fitting results.

Additionally, the prefectures of Aomori, Akita, Iwate in the Tohoku region, and Hokkaido are included in the top ten. These prefectures are in the northern part of Japan. The number of principal components and the estimated values of parameters vary considerably among the prefectures, which means that comparing them could potentially capture the unique characteristics of tourism in each prefecture.

Table 7: Regression Coefficients of Non-Resident Overnight Guest Model (Top 10 Prefectures)

Prefecture	R^2	$adj\ R^2$	$\alpha_{\square}^{(k)}$	$\beta_1^{(k)}$	$\beta_2^{(k)}$	$\beta_3^{(k)}$	DW
Yamanashi	0.917	0.912	1188.00***	379.41***			1.348
Okayama	0.858	0.836	359.06***	47.22***	25.5***		2.442
Niigata	0.848	0.825	874.4***	70.75**	236.14***		1.246
Aomori	0.831	0.819	324.63***	110.79***			1.836
Nagano	0.830	0.804	2129.72***	568.42***	247.13***		1.843
Akita	0.812	0.799	249.81***	110.66***			2.885
Nara	0.812	0.765	314.66***	53.08***	-3.23	28.1***	1.513
Iwate	0.803	0.773	401.21***	74.06***	10.1		2.039
Shiga	0.803	0.753	495.47***	87.37***	30.74**	10.42	1.622
Hokkaido	0.784	0.750	905.46***	189.97***	56.81*		2.657

Note: Statistical significance levels are indicated as ***p<0.01, **p<0.05, *p<0.1.

5.3 Predictive Power

Figure 2 presents a scatter plot with the coefficient of determination from model estimation data (2014–2017) on the horizontal axis and the predictive power testing data (2018–2019) on the vertical axis. The 45-degree line illustrates that points close to this line reflect a similar fit in both estimation and predictive datasets. Points above this line suggest a better fit in the predictive dataset, while those below indicate a poorer fit. Notably, four prefectures with negative multiple correlations in the predictive data—Kanagawa, Shimane, Ehime, and Saga—were excluded from this analysis. Thus, 28 out of 32 prefectures are plotted in the scatter diagram.

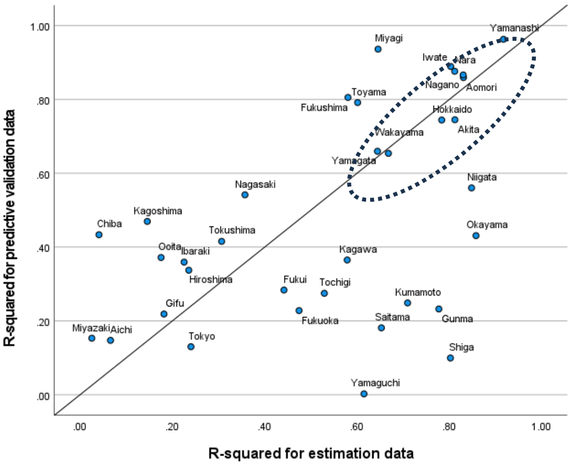

Figure 2: scatter plot with the coefficient of determination

While prefectures such as Shiga, Yamaguchi, Gunma, Saitama, Okayama, and Niigata appear notably below the line, indicating less predictive accuracy, prefectures like Yamanashi, Nara, Aomori, Iwate, Nagano, Akita, and Hokkaido, which had high coefficients of determination in the estimation data, showed consistent fitting trends in the predictive data. These prefectures are enclosed in an ellipse in the figure, demonstrating the utility of the proposed methodology in this research.

6 Summary and Challenges

This research aimed at utilizing open data, specifically in the field of tourism, to propose an analysis process that is highly convenient. The regression analysis, which avoided the strong correlation of explanatory variables through principal component analysis, can be performed by anyone with some training in analysis. Although there were variations in fit among the prefectures, the results showed good fits. Demonstrating that good results can be obtained with low monetary and effort costs is a contribution of the analysis process we proposed.

However, there are several challenges as well. The data used in this study are time-series data, but they were treated as cross-sectional data for regression analysis. After conducting the Durbin-Watson test, only three prefectures—Ehime, Miyagi, and Gifu—out of the 44 were determined to have either positive or negative serial correlation, which was not a significant problem. However, while 18 prefectures showed no serial correlation, 23 prefectures were indeterminate, indicating that the problem was not entirely absent. Considering methods that account for serial correlation in future models is one of the challenges ahead.

Furthermore, while priority was given to simplicity using principal component analysis, there are other methods to avoid multicollinearity. For instance, ridge regression [24], LASSO regression [25], PLS regression [26], or machine learning could be considered as alternatives.

The challenges discussed pertain to analysis, but it is also crucial to address the limitations of the data. There was variation in fit across prefectures. Generally, better fits were observed in regions of northern or eastern Japan, such as the Tohoku region, and lower tendencies in the south, such as Kyushu region (see. Table 5&7). Whether this is due to regional differences, issues in the data collection process, or other reasons is currently unclear.

Moreover, the granularity of the data, being aggregated at the quarterly and prefectural levels, is coarse. This might have somewhat lowered the adequacy of the analysis results. Therefore, it is necessary to validate the process proposed in this research using more finely grained open data.

References

[1] WHITE HOUSE, “Promoting Transparency in Government,” 8 Dec. 2009; https://obamawhitehouse.archives.gov/blog/2009/12/08/promoting-transparency-government.

[2] S. Miyagawa, M. Endo, M. Urata, and T. Yasuda, “Development of data utilization methods for objective analysis in municipal operations: Effectiveness and visualization through information linkage between local events and operations,” Studies in Science and Technology, vol. 12, no. 2, 2023, pp. 131-136.

[3] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, “From data quality to big data quality, ” Journal of Database Management, vol. 26, no. 1, 2015, pp. 60-82.

[4] S. Sadiq, and M. Indulska, "Open data: Quality over quantity,” International Journal of Information Management, vol. 37, no. 3, 2017, pp. 150-154.

[5] R.Y. Wang and D.M. Strong, “Beyond accuracy: What data quality means to data consumers,” Journal of Management Information Systems, vol. 12, no. 4, 1996, pp. 533.

[6] R.Y. Wang, “A product perspective on total data quality management,” Communications of the ACM, vol. 41, no. 2, 1998, pp. 58-65.

[7] T. Koltay, “Quality of Open Research Data: Values, Convergences and Governance,” Information, vol. 11, no. 175, 2020; DOI:10.3390/info11040175.

[8] M. Archie, S. Gershon, A. Katcoff, and A. Zeng, “Who’s Watching?,” MIT OpenCourseWare, 2018, Spring; https://courses.csail.mit.edu/6.857/2018/project/Archie-Gershon-Katchoff-Zeng-Netflix.pdf.

[9] US-CERT, “Protecting Aggregated Data,” 2005; https://www.cisa.gov/sites/default/files/publications/Data-Agg-120605.pdf

[10] J. Koo, G.G. Kang, and Y.G. Kim, “Security and Privacy in Big Data Life Cycle: A Survey and Open Challenges,” Sustainability, vol. 12, no. 24, 2020, 10571; DOI: 10.3390/su122410571.

[11] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp, Digital Humanities, The MIT Press, 2012.

[12] Digital Agency, “About the Municipal Standard Open Data Set (Former Standard Data Set),” March 31, 2023. https://www.digital.go.jp/resources/open_data/municipal-standard-data-set-test

[13] German National Tourist Board, “Open Data Destination Germany,” 17 Feb. 2020; https://open-data-germany.org/en/how-the-largest-digital-infrastructure-project-in-tourism-is-taking-shape/.

[14] European Union, “Open Data in tourism,” 24 Oct. 2018; https://data.europa.eu/en/publications/datastories/open-data-tourism.

[15] García-Madurga, Miguel-Ángel, and Ana-Julia Grilló-Méndez. "Artificial Intelligence in the tourism industry: An overview of reviews." Administrative Sciences 13, no. 8, 172, 2023; doi: 10.3390/admsci13080172.

[16] S.I. Stewart and C.A. Vogt, “Multi-destination trip patterns,” Annals of Tourism Research, vol. 24, no. 2, 1997, pp.458-461; DOI:10.1016/S0160-7383(97)80017-5.

[17] C. Tideswell and B. Faulkner, “Multidestination Travel Patterns of International Visitors to Queensland,” Journal of Travel Research, vol. 37, no. 4, 1999, pp. 364-374; DOI:10.1177/004728759903700406.

[18] C.C. Lue, J.L. Crompton, and D.R. Fesenmaier, “Conceptualization of multi-destination pleasure trips,” Annals of Tourism Research, vol. 20, no. 2, 1993, pp.289-301; DOI: 10.1016/0160-7383.

[19] T.H. Wonnacott and R.J. Wonnacott, Regression, John Wiley & Sons, Inc., 1981, Translated by Y. Tabata and T. Ohta as "Kaikibunseki to sono ouyou" (Regression Analysis and Its Applications), Gendai Sugaku-sha, Japan, 1998.

[20] Honda, Masami, "Declaration to be the World's Most Advanced IT Nation and E-government Policy," IPSJ-SIG Security Psychology and Trust, vol. 2015, no. 10, 2015, pp. 1-7. [in Japanese].

[21] Ministry of Land, Infrastructure, Transport and Tourism, Japan Tourism Agency, "Common Standards for Tourism Statistics Survey Guidelines, Revised March 2013," Kanko Irikomikyaku Tokei ni kansuru Kyotsu Kijun Chosa Yoryo, Heisei 25-nen 3-gatsu kaitei [in Japanese], 2013; https://www.mlit.go.jp/kankocho/content/810003353.pdf

[22] Malhotra, Naresh K. Marketing Research: An Applied Orientation. 4th ed., Prentice Hall, 2004. Translated by Kazuo Kobayashi, supervised by the Japan Marketing Research Association. Japanese title: "Marketing Risachi no Riron to Jissai – Riron Hen." Doyukan, 2006.

[23] Weisberg, Sanford. Applied Linear Regression. 4th ed., John Wiley & Sons, Inc., 2014. Japanese edition supervised by Etsuo Miyaoka, translated as "Deta Kaiseki no tame no Senkei Kaiki [Gencho Daiyonhan]", Kyoritsu Shuppan, 2024.

[24] A.E. Hoerl and R.W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” Technometrics, vol. 42, no. 1, 2000, pp.80-86; DOI:/10.2307/1271436.

[25] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1 1996, pp. 267-288; https://www.jstor.org/stable/2346178

[26] H. Wold, “Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach”, Journal of Applied Probability, vol. 12, Issue S1, 1975, pp. 117-142.