Big data analytics - towards the enrichment of content tourism for revitalization of Japanese rural area

Ali Bakdur[a], Fumito Masui and Michal Ptaszynski

Department of Computer Science, Kitami Institute of Technology, 090-8507 Kitami, Japan

Abstract. Japan's domestic travel and tourism industry expenditure has been declining gradually since 1998 (from 33.5 in 1998 to 21.6 trillion JPY in 2016). Our research purpose is to construct a data analysis model to transform the collected data to a meaningful graphical format by using big data analytics techniques to discover anomalies and sustainable development possibilities for economy and tourism of Japan's rural areas, with a particular focus on the prefecture of Hokkaido, subprefecture of Okhotsk. To strengthen the reliability of this model we apply popular Monte Carlo simulation combined with Bayesian statistic and implement it on an Apache Spark platform to acquire results within the span of the study. Through this research, we focus on observing and analyzing interests, expectations and tendencies of Japanese people living in rural areas. From such collected information, we can obtain reasons for the decline of this sector's impact on Japan's economy. Measuring public awareness has become more efficient since the content generator role has been passed on to ordinary people. Therefore, the analysis of Big Data with the use of data science techniques has become important to comprehend human behavior from multiple points of view, including the scientific, economic, political, historical and sociological.

1 Introduction

Following the commercial success of world's first generation jet airliners in the second half of the 1950's, passenger transport has begun to play a major role in the development of economy and globalization. Over the past six decades, tourism has experienced continued expansion and diversification to become one of the largest and fastest-growing economic sectors in the world. International tourist arrivals had increased from 25 million globally in 1950 to 278 million in 1980, 674 million in 2000, and 1186 million in 2015. Moreover, the number of global domestic tourists is estimated to attain 5 to 6 billion people as of 2015 [1].

There are several technical definitions of tourism but at this point, we will apply the definition of tourism considering the economy as a parameter, namely:

> *"Tourism can be defined as the science, art, and business of attracting and transporting visitors, accommodating them and graciously catering to their needs and wants."* [2]

This economic approach to the definition of tourism is also appropriate in the case of Japan on which we focused.

[a] Corresponding author : alibakdur@ialab.cs.kitami-it.ac.jp

Japan is a unitary island country with a population of 126.8 million people. Although it's comparatively large population, Japan has a rapid population aging problem and a negative population growth rate. It is one of the most urbanized countries in the OECD, as 57% of its population live in predominantly urban areas [3]. Although it has almost 25 years of slow economic growth [4] (The notorious 'Japanese asset price bubble', an economic bubble in Japan which burst by the end of 1991, can be given as the reason for the country's economic decline), with 520 trillion JPY (4.8 trillion US Dollars) GDP nominal value in 2017 [5], Japan is still the 3rd largest economy in the world. Japan is also the first very highly developed country in Asia.

In addition to the above information regarding Japan, travel and tourism industry generated a total impact of 38 trillion JPY (343 billion US Dollars) of Japan's GDP in 2016. It makes Japan the 3rd country in the world regarding the total contribution of travel and tourism to the GDP after the United States and China. In view of Japan's island nature, 87.9% of the total contribution of the tourism industry to the GDP consists of domestic expenditure. Based on its direct, indirect, and induced GDP impact, travel and tourism generated 7.4% of Japan's GDP in 2016. This is nearly half the impact of the automotive manufacturing sector at 16.5%. Travel and tourism sustained a total of 4.5 million direct, indirect and induced jobs in Japan in 2016 [6].

It is obvious that Japan has a well-developed domestic travel and tourism industry. Its total contribution to Japanese GDP is making it one of the most influential sectors. However, both the country's GDP and the participation of the tourism industry in it had been declining gradually since 1998. Although Japanese domestic tourism industry has been in decline comparing to other countries, it is still on a very high level in 2016, but depending on the historical data it has lost great economical value since 1998 (from 33.5 in 1998 to 21.7 trillion JPY in 2016) [7].

A correct comprehension of the situation requires detailed knowledge and data analysis, which depends on scientific, economic, political, historical and sociological aspects. By the start of the digital era around 2002, the quality and quantity of available data have started to increase dramatically. Data have never before been as valuable as today. Not only data but also efficient processing of it has gained great importance. Nowadays many organizations [1, 3-9] have started to share their data with the public, which allowed us to collect required data samples for our research.

2 Domestic tourism overview of Japan

2.1 General overview of Japan

Simultaneously with the decline of Japanese domestic travel and tourism expenditure, the participation rate in domestic tourism activities in Japan is also in decline between 1986 and 2006. Participation in sightseeing has dropped from 65.9 % to 49.3 % of the overall population [8]. Our understanding from the statistical data is that today in Japan fewer people than in the past are traveling for leisure. Perhaps the reason for this is that people feel more comfortable in their usual habitat. Nevertheless comparing only historical statistical data values from past and present in a frequentist way may not give the correct comprehension of this phenomenon. The consideration of living habits changes throughout the time, the development of popular technologies such as the Internet, communication, home entertainment systems etc. is essential for a better interpretation and inference of the real state of Japanese daily routine and understanding of their insight on leisure.

As explained in the previous paragraph the necessity of interpreting different terms of time with several different kinds of parameters brought us to the Bayesian inference approach. From our point of view, the usage of Bayesian statistic, which we implemented in the technical part of this study to prove our concept, has given a better conceptual quality of probability.

2.2 Okhotsk subprefecture of Japan

To give an adequate understanding of human experience in the current situation in Japan, we are using Okhotsk subprefecture's statistical data as samples throughout this study. The Okhotsk subprefecture of Hokkaido, second largest island of Japan, consists of 3 cities: Abashiri, Kitami, and Monbetsu. It gets its name from the Sea of Okhotsk. The climate is much colder than in other parts of Hokkaido during the winter. The area is also famous for drift ice sightseeing. Agriculture is an important sector. Compared with other regions in Japan, the density of the population in Okhotsk subprefecture is low.

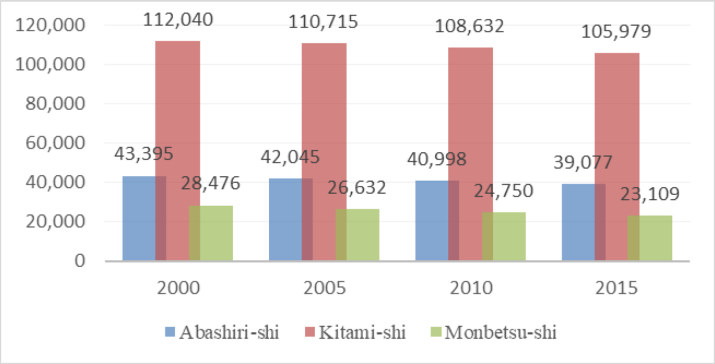

Figure 1. Population Census in Okhotsk (2000 -2005 – 2010- 2015) [9]

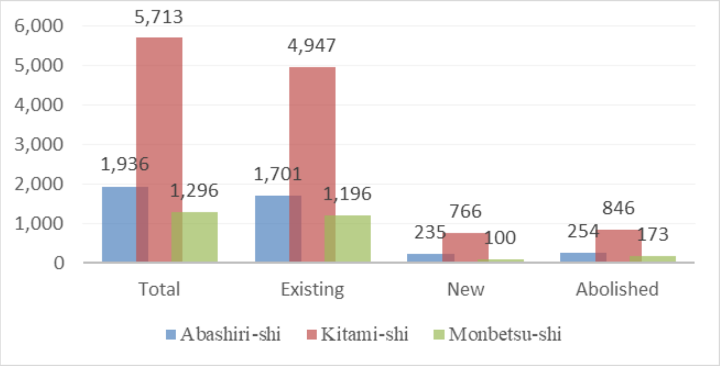

Figure 2. Establishment's Existence in Okhotsk (2014) [9]

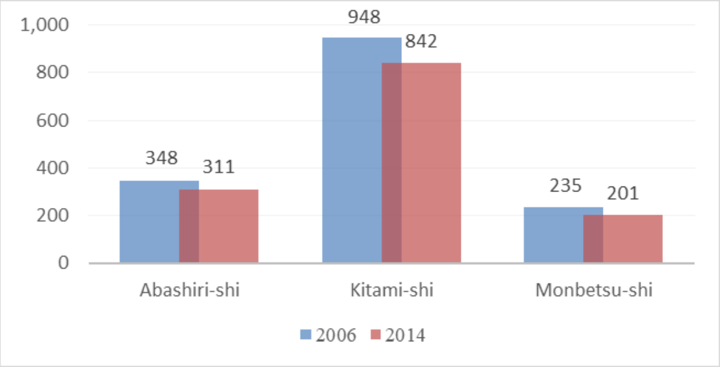

Figure 3. Accommodation, Food Service Businesses in Okhotsk (2006-2014) [9]

In this analysis, the character of Okhotsk statistical data, which is shown respectively in Figure 1, 2 and 3, indicates similarities with Japan's general characteristics. Okhotsk also has a negative population growth rate, but in contrary to the rest of the country all business establishments including tourism-related businesses are disappearing.

3 The purpose of applying the Bayesian approach to this study

Tourism information science combines several kinds of technology and knowledge, such as data science, or Big Data. In addition to a better understanding of the mechanisms behind various kinds of sightseeing spots, the data, from historical statistics and/or current streaming, will be processed in this research using data science approach. A statistical model will be produced by Monte Carlo simulation combined with Bayesian statistics and the result will be displayed with probability distribution [10].

Distributed computing with Apache Spark is essential for the performance issue. The principal working steps of this model will be revealed after conducting a preliminary research focused on tourism and demographic indicators in chronological statistic and prebuilt survey data in Okhotsk subprefecture of Japan. This observation is required to build a reliable model of local community characteristics and expectations.

The next step will be formulating a mathematical structure model with the detected factors and parameters relying on the previous studies. The analysis of data related to continuous events (social media networks) will be a part of the next phase for detecting present-day tendencies of the community.

We aim at collaborating with the local authorities in order to provide a better tomorrow for the community and to improve their quality of life-based on scientific proof earned through the completion of this study.

3.1 Bayesian statistics

The Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides users with the tools to update their beliefs in the evidence of new data. If the occurrence of any incident depends on a variety of conditions then Bayesian interference combined with Monte Carlo method would be appropriate as an observation technique.

An important part of Bayesian inference is the establishment of parameters and models. The main mechanism is feeding the analysis model with repeated random sampling inputs to get numerical output. We assume that a repeating sampling input, depending on modeling complexity, may require high processing power and time, which could be achievable with Apache Spark components.

3.2 Monte Carlo method

Monte Carlo methods can be used to solve any problem having a probabilistic interpretation. A very powerful class of Monte Carlo techniques is the so-called Markov chain Monte Carlo (MCMC) algorithms. Monte Carlo methods are often necessary for the implementation of optimal Bayesian estimators. Monte Carlo methods are mainly used in three distinct problem classes [11]:

- Optimization,
- Numerical integration,
- Generating draws from a probability distribution.

3.3 Apache Spark

Apache Spark is an in-memory framework that is an alternative to MapReduce. Spark has a higher data processing engine for large datasets and in-memory computing. It can run programs up to 100 times faster than Hadoop MapReduce. The Spark core is designed to scale up from one to thousands of nodes. Spark powers a stack of libraries including SQL and Data Frames, MLlib for machine learning, GraphX, and Spark Streaming. It is possible to combine these libraries within the same application [12]. We used IBM Big Data University Virtual Lab Environment for faster processing [13].

3.4 PyMC

PyMC is a Python package for Bayesian statistical modeling and Probabilistic Machine Learning, which focuses on advanced Markov chain Monte Carlo and various fitting algorithms. Its flexibility and extensibility make it applicable to a large array of problems [14].

4 Case study and solution steps

Before proceeding to the case study assignment we should mention that an important element of statistical knowledge and eventually probabilistic program development are random variables. Random variables are basically divided into two sub-categories:

- Observed Random Variables,
- Unobserved Random Variables

A data scientist who disposes of a dataset sample can use observed random variables that can come in handy in likelihood distributions. On the other hand, unobserved random variables have no sample dataset values and are defined by prior distributions.

If we define random variables more specifically in terms of probability programming and probability distribution aspects, we need to mention at least two types:

- Discrete Random Variables

Discrete random variables find usage alongside the probability mass function when the output of the function data value is finite numbers. Examples include discrete uniform distribution, Bernoulli distribution, binomial distribution, Poisson distribution, and geometric distribution, etc.

- Continuous Random Variables

Continuous random variables are defined by a probability density function and the output of the data value can be defined as infinite numbers. Examples include normal distribution, uniform distribution, beta distribution and gamma distribution, etc. In the following steps, we will explain the usage of this information in more detail.

4.1 Problem statement

Japanese domestic travel and tourism industry has been declining gradually since 1998. What are the expectations for its future?

4.2 Data

We obtained Japanese domestic travel and tourism expenditure data for the years in between 1995 and 2015 (local currency in trillion JPY), which we then used for this analysis [7].

Table 1. Travel and Tourism expenditure/year data in trillion JPY by year (1995- 2015)

Year	Value	Year	Value	Year	Value
1995	26	2002	31	2009	24
1996	29	2003	29	2010	22
1997	32	2004	28	2011	21
1998	34	2005	27	2012	21
1999	32	2006	28	2013	20
2000	32	2007	26	2014	22
2001	32	2008	26	2015	22

As seen in Table 1, we have some prior states dataset for intuitive perspective. Since we have a sample dataset it makes our case suitable for defining observed random variables.

Using the bivariate numerical data in Table 1, we create the following regression line graphic with a negative linear relationship in Figure 4.

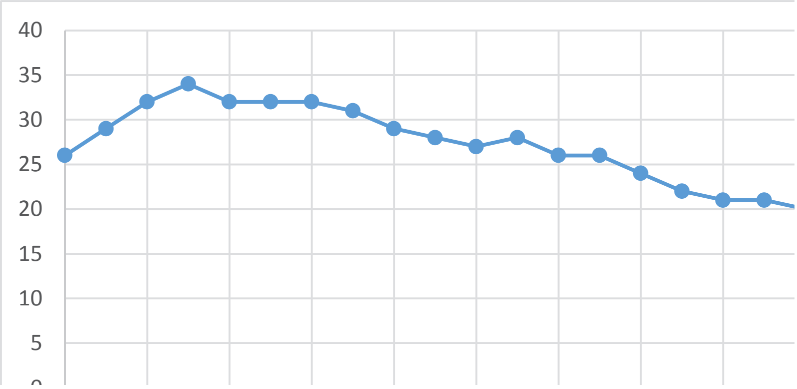

Figure 4. Travel and Tourism Expenditure (trillion JPY) /Yearly regression line graph.

4.3 Model

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long-term frequencies [15]. In our case, we have a dataset with prior beliefs which gives us observed random variable opportunity. Our dataset values are also discrete and their value is an integer. This makes it possible for us to use the probability mass function for modeling. For this reason, we used Poisson distribution for probability mass function which is the best fit for our purpose.

For Poisson distribution (1), we can say that:

- x events that occur at times.
- λ is event occurrence rate.
- e is the constant number equal to 2.71828… (Euler's Number)

$$Poi(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0, 1, 2 \ldots \tag{1}$$

From the equation, the probability is dependent on λ. The value of λ determines the shape of the probability distribution. Two very different λ s will produce two very different distributions. We can plot a Poisson distribution with two different values of λ [13]. Observation of Figure 4, can help us to determine a turning point in the time period of event occurrence if there is any. In other words, if there is any unusual behavior on the observed random variables regression line graph the mean of which is our distribution governed by two different values of λ one before the turning point and one after the turning point.

```
import pymc as pm

def tourism_model(data):

    alpha = 1.0 / data.mean()
    lambda_1 = pm.Exponential("lambda_1", alpha)
    lambda_2 = pm.Exponential("lambda_2", alpha)
tau = pm.DiscreteUniform("tau", lower=0,
       upper=len(data))
observation = pm.Poisson("obs", lambda_,
                  value=data.values, observed=True)
 model = pm.Model([observation, lambda_1, lambda_2, tau])
```

Figure 5. Sample Python code block responsible for the building of a probabilistic model with our data.

While the concept is complicated, to make it simple we initially attempted to solve this problem using Python programing language and its probability programming specified package- PyMC. PyMC allows building a probabilistic model in a short time. With the given sample Python code block as seen in Figure 5, we can define the explained model in a short time. τ (tau) represents the turning point which we are looking for.

4.4 Inference

After creating a model definition we can start to use the MCMC sampling algorithm.

- In the first step, we are introducing our observed data set to the model.
- In the second step, we are assigning the MCMC (Markov chain Monte Carlo) fitting algorithm for the sampling of our model. With MCMC algorithms, we are sampling our model 50000 times with random values.
- At the end of the sampling process, we got the required set of λ1, λ2, and τ (tau) data.

Next, a code inside the model creates a graph set for the interpretation of this problem. The key purposes of our research to provide a more user-friendly output for people in our community that do not have a technical background.

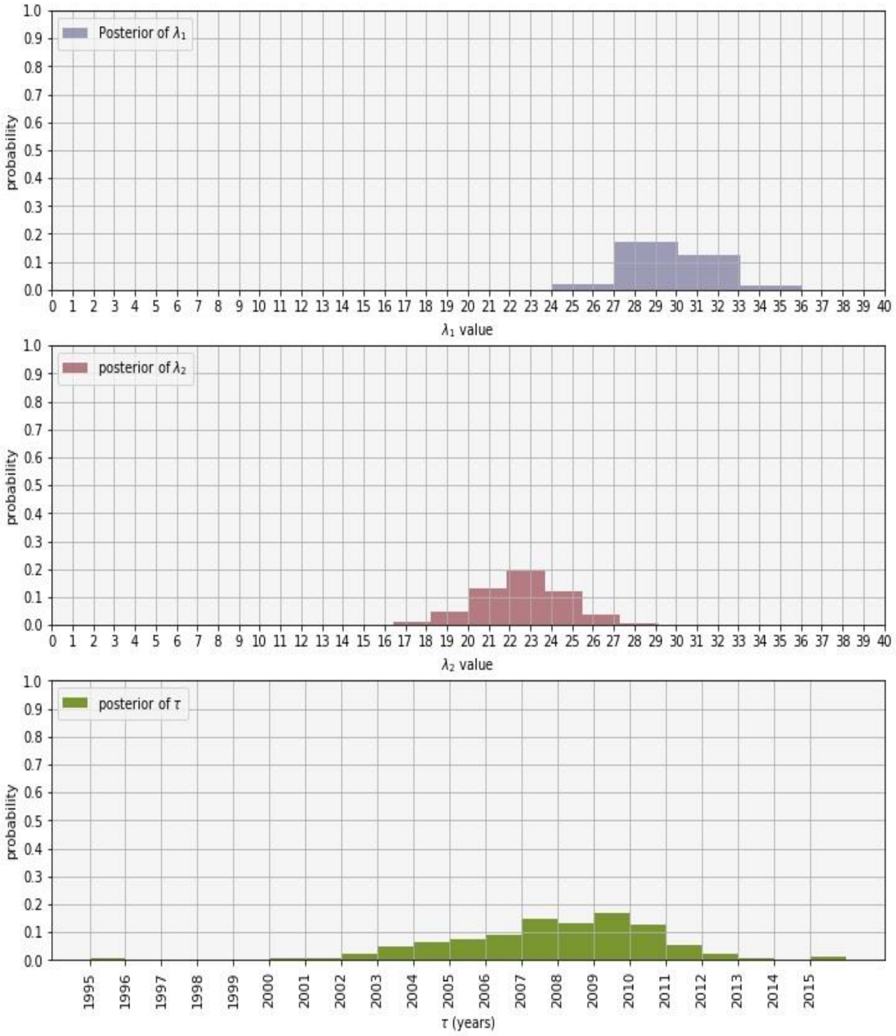

Figure 6. Posterior distributions of λ1, λ2, and τ.

4.5 Result

In Figure 6, the first graph λ1 shows the distribution before τ, while the second graph shows λ2 after τ. It should be noted that τ graphs have not had any certain turning point because of our prior data showing a continuous decline of the domestic travel and tourism expenditure. Since there was no change, λ1 and λ2 showed similar possibilities. We can conclude that posterior distribution of domestic travel and tourism spending will change in between two groups of data first realistic 22 up to 23 trillion JPY or optimistic 28 up to 30 trillion JPY for the following years.

Indeed, the expenditure for domestic travel and tourism in 2016 was 21.7 trillion yen. Another important point is that from the beginning of 2015, which was relatively a very good year for tourism in Japan, this trend has started to change in a positive way.

5 Conclusion

The result of this study will create an impact so as to increase the efficiency of the everyday life of the community living in the Okhotsk area of Japan and with the same approach based on Bayesian statistics the study could help understand better both prefectural (Hokkaido) and country level (Japan) human experience.

The value distribution of our surroundings is changing at every moment, and the impact it creates requires it to be measured as quickly as possible for fluent comprehension in modern society. Never before has data been as valuable as today, but processing the data in time is of crucial importance everywhere in the world. Big Data enforced various processing solutions, including data cleaning, and analysis to derive value from it. Nevertheless, the vast majority of the people needs for Big Data to be presented in a user-friendly manner so that they could understand it clearly and make their decisions.

Our main interest will be to support the Okhotsk sub-prefecture of Japan and the people who live there through the data collected and processed in this research.

References

1. UNWTO (United Nations World Tourism Organization), Tourism Highlights 2016 Edition, http://www.e-unwto.org/doi/pdf/10.18111/9789284418145
2. McIntosh, R. and Goeldner, C. (1977) “Tourism, 2nd edition.” Wiley, New York.
3. OECD (Organization for Economic Co-operation and Development), Regional Outlook 2016: Productive Regions for Inclusive Societies, Japan, http://www.oecd-ilibrary.org/urban-rural-and-regional-development/oecd-regional-outlook-2016/japan_9789264260245-32-en
4. OECD (Organization for Economic Co-operation and Development), Japan: Boosting Growth and Well-being in an Ageing Society, http://www.oecd-ilibrary.org/economics/japan-boosting-growth-and-well-being-in-an-ageing-society_9789264256507-en
5. IMF (International Monetary Fund), World Economic Outlook (April 2017), http://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx
6. WTTC (World Travel & Tourism Council), Japan Benchmarking Report 2017, https://www.wttc.org/-/media/files/reports/benchmark-reports/country-reports-2017/japan.pdf
7. WTTC (World Travel & Tourism Council), WTTC Data Gateway, Japan Domestic Tourism Spending, https://www.wttc.org/datagateway/
8. Historical Statistics of Japan, Japan Statistic Bureau, http://www.stat.go.jp/english/data/chouki/index.htm
9. MIC (Ministry of Internal affairs and Communications), Japan Statistic Bureau, http://www.stat.go.jp/english/data/index.htm
10. S.L. Scott, A.W. Blocker, F.V. Bonassi, H.A. Chipman, E.I. George and R.E. McCulloch, International Journal of Management Science and Engineering Management, **11**(2), 78-88 (2016)

11. D.P. Kroese, T. Brereton, T. Taimre and Z.I. Botev, Wiley Interdisciplinary Reviews: Computational Statistics, **6**(6), 386-392 (2014)
12. Apache Foundation, Apache Spark, http://spark.apache.org/
13. IBM Big Data University Virtual Lab Environment, https://my.datascientistworkbench.com/
14. PyMC3, https://pymc-devs.github.io/pymc3/index.html
15. Kevin P. Murphy, *Machine Learning A Probabilistic Perspective* (Chapter 2, 2012)