Generating reliable tourist accommodation statistics: Bootstrapping regression model for overdispersed long-tailed data

Nguyen Van Truong
University of Transport and Communication, Vietnam

Tetsuo Shimizu
Tokyo Metropolitan University, Tokyo, Japan

Sunkyung Choi
Tokyo Institute of Technology, Japan

Abstract
***Purpose**: Few studies have applied count data analysis to tourist accommodation data. This study was undertaken to investigate the characteristics and to seek for the most fitting models for population total estimation in relation to tourist accommodation data.*
***Methods**: Based on the data of 10,503 hotels, obtained from by a nationwide Japanese survey, the bootstrap resampling method was applied for re-randomisation of the data. Training and test sets were derived by randomly splitting each of the bootstrap samples. Six count models were fitted to the training set and validated with the test set. Bootstrap distributions for parameters of significance were used for model evaluation.*
***Results**: The outcome variable (number of guests), was found to be heterogenous, over dispersed and long-tailed, with excessive zero counts. The hurdle negative binomial and zero-inflated negative binomial models outperformed the other models. The accuracy (se) of the estimation of total guests with training sets that ranged from 5% to 85%, was from 3.7 to 0.4 respectively. Results appear rather overestimated.*
***Implications**: Findings indicated that the integration of the bootstrap resampling method and count regression provide a statistical tool for generating reliable tourist accommodation statistics. The use of bootstrap would help to detect and correct the bias of the estimation.*

Keywords: *tourism statistics, bootstrap, count regression, heterogeneity, over dispersed data, zero-inflated data*

JEL Classification: *C4, L8, C24, Z3*

Biographical note: Nguyen Van Truong is lecturer at University of Transport and Communication, Vietnam (ngvtruong@utc.edu.vn). Tetsuo Shimizu is professor at Tokyo Metropolitan University, and member of the National Committee on Tourism Statistics Development in Japan (t-sim@tmu.ac.jp). Takeshi Kurihara is associate professor at the Department of International Tourism Management of Toyo University, Japan (kurihara039@toyo.jp). Sunkyung Choi is a specially-appointed associate professor at Tokyo Institute of Technology, Japan (choi@jterc.or.jp). Corresponding author: Nguyen Van Truong (ngvtruong@utc.edu.vn)

1 INTRODUCTION

Count data regression, which is regression for discrete, non-negative integer data, has a long history in academic literature. Count data analysis was first introduced by Bortkiewicz (1898), in his work "The Law of Small Numbers"; in one example, he showed that the number of Prussian army soldiers who died after being kicked by a horse followed a Poisson (1837) distribution, where the conditional mean and variance were equal. Unobserved heterogeneity can lead to a longer right tail and/or under- or overdispersion; extensions of the Poisson distribution, such as negative binomial (NB), Poisson-inverse Gaussian, Sichel, and Delaporte distributions, wherein additional parameters are added to a single-parameter distribution, may be more suitable in such cases. Additionally, depending on the data generation process (DGP), excessive zero counts may arise. Thus, the regression model most suitable for a given DGP, e.g. zero-inflated, hurdle, truncated, or censored, depends on the problem.

It is well-known, in literature, that statistical data and statistical methods are important in every practical and academic fields (Hand, 2008). In economic development in general, statistical data and statistical methods are considered as pivotal tools to support policy decision making. Sanga (2011) even stressed the role of statistical data and methods as essential basic for poverty reduction strategies. Tourism

ISSN: 2529-1947. UDC: 658.8+338.48+339.1+640(05)
http://doi.org/10.5281/zenodo.3835845
Published online: 30 May 2020
www.jthsm.gr

have recognised as one of the largest economic industry in the world. According to WTTC (2019), tourism industry contributed approximate 10.3% of global GDP and one in ten jobs around the world as total impact. The contribution is projected to continue increasing in the next ten years. It is expected that, many countries have been developing their tourism statistical database for aiding the planning, management, and policy making. However, there is a lack of information on how the statistical database was developed and how good quality the data are. The utilisation of poor-quality statistical data, and the in-appropriate method may be an obstacle to the development goal or even leads to the failure of policy implementation in practice; and it is a reason of unsatisfactory works for publication in academia (Fonton & Hounkonnou, 2011; Zepp, 2011).

As few studies recognised using count regression in tourist accommodation data (see in literature review section for more details), the overall objective of the current study was to attempt to find an appropriate statistical method that support to develop a high-quality accommodation statistical data. The specific objectives of the study were as follows: firstly, to investigate the characteristics of count data pertaining to tourist accommodation facilities, and determine the most suitable non-econometric count model for estimating the total number of guests staying at all hotels in Japan; secondly, to examine the performance of count models for estimating guest numbers based on various-sized sub-samples; and finally, to emphasise the advantages of the bootstrap resampling method for model selection, evaluation, and validation, in the context of tourist accommodation statistics. The remainder of the paper is organised as follows. In the next section, the literature review, wherein the focal points will be count data regression and bootstrap technique, will be presented. The study data and the methodology used for the analysis in the following section. The key findings of the study, including the best-fitting model will be presented in the analysis result section. The performance of that model is then demonstrated using various sizes of test sets. Finally, the further discussions, implications and conclusions of the study will be presented.

2 LITERATURE REVIEW

The application of count data regression in both non-econometrics and econometrics, including single and mixed distributions integrated into one- or two-part regression models, has been widely documented in a variety of research fields. In the political field, King (1988) employed statistical models to show the number of United States (US) House representatives who switched political parties each year over the period 1802–1876; the event counting process followed the Poisson law. Research has shown that among ordinary least square (OLS), logarithmic OLS, and Poisson regression, the latter is superior, in terms of efficiency, consistency, and bias, for analysing event count data. Count regression techniques have also been applied in biological research fields. For example, Alves et al. (2013) suggested that "standing crop line transect counts" was superior to other methods, in terms of precision, accuracy, and efficiency, for estimating the density of red deer in Lousã, Central Portugal. In the medical and healthcare fields, Deb and Trivedi (1997) used econometric count regression when analysing data of the US National Medical Expenditure Survey of 1987–1988; their analysis, which applied certain statistical selection criteria, argued that the mixture model, i.e. a hurdle model that extended the standard NB, is preferable for describing unobserved heterogeneity. In another study, Deb and Trivedi (2002) demonstrated the superior performance of econometric count regression for distinguishing groups of "ill" and "healthy" patients according to their need to see a doctor (high or low; data provided by the RAND Health Insurance Experiment, USA). In social, psychological, and economic research fields, many scholars, including Duarte and Escario, (2006), Hausman et al., (1984), and Solis-Trapala and Farewell, (2005) successfully applied generalised econometric count modelling to account for covariates in observations of various phenomena.

In tourism, count models have been used to some extent. Many researchers have applied standard count regression models, with various extensions, to analyse tourism data generated by on-site surveys. Martínez-Espiñeira and Amoako-Tuffour (2008) utilised Poisson, NB, and truncated models to analyse overdispersed data pertaining to tourist trips to Gros Morne National Park in Newfoundland, New Zealand. A comparison of typical count models for analysing data on recreational fishing trips in Pantanal, Brazil was conducted by Shrestha et al. (2002). Another notable example comes from Grogger and Carson (1991), who analysed count data generated from a survey of 1,063 Alaskan households, where the data were truncated by removing the samples of all households that did not take at least one fishing trip; the data were well-fitted to their truncated NB model.

The common characteristic of count data generated by on-site surveys is that the interviewees are present at the site, thus leading to truncation and endogenous stratification. The modelling in such cases requires considerable care to avoid overestimates. This phenomenon was mentioned by Yen and Adamowicz (1993) in their analysis of count data pertaining to tourism demand in Alberta, Canada in 1981. The issue of truncated count data can be addressed by improving the sampling or correcting for truncation. For instance, household and off-site survey methods can improve the truncation issue; however, household and off-site surveys are more expensive than on-site ones, and the DGP of off-site surveys may produce more zero counts, which must be addressed by specific analysis techniques such as hurdle and zero-inflated models. Thus, researchers must choose the appropriate survey type and analysis method, based on their understanding of the trade-offs between approaches. To date, few studies in the literature have applied count data analysis to tourist accommodation data.

The bootstrap resampling method (B. Efron, 1979) is well-known with great advantages. The first is that, bootstrap is considered as an intuitive and practical in applications since it "is a data-based simulation method for statistical inference" (Bradley Efron & Tibshirani, 1994:5) for confidence interval. The second, bootstrap provides more accurate estimation of confidence interval than other standard interval which obtained from sample variance with assumption that sample is normally distributed. In reality, in many cases wherein the normality is violated, the standard estimation of the confidence interval, which relying on the sample variance, becomes impossible; this is when bootstrap comes in and

appears outperformed other standard methods (DiCiccio & Efron, 1996). Finally, bootstrap is powerful in detecting bias of estimation and support to correct the biased estimation (Carpenter & Bithell, 2000).

Although, the bootstrap has been used by many researchers in numerous fields of study, this technique has been applied in few studies related to tourism. Palmer Pol et al. (2006) pointed out that based on 1790 academic articles, which sourced from 12 relevant tourism journals covering the period 1998-2002, there was no article utilised bootstrap methods. The finding indicated the big gap in tourism research at the time, and that might be a fertile area for applications of bootstrap in tourism. Later time, there were several studies found in tourism that utilised bootstrap methods. For example, Pol et al. (2006) first applied the bootstrap method to 2001 survey data for the Balearic Islands, to evaluate the fundamental variables associated with tourist expenditure. Chou (2013) examined the relationship between spending by tourists and economic growth in 10 countries via panel data analysis. Assaf et al. (2010) and Assaf and Agbola (2011) evaluated the efficiency of a total of 78 hotels in Taiwan and Australia, in the periods 2004-2008 and 2004-2007, respectively, using meta-frontier and data envelopment analysis (DEA) approaches, respectively. Chen and Fomby (1999) and Gergaud et al. (2018) compared the performance of different models of Hawaiian tourism and assessed the impact of terrorism events on wine tourism in France, respectively; both studies employed time series models. All of the above-mentioned studies utilised a bootstrap technique to generate mean, variance, skewness, and interquartile range data for the parameters of interest. Although bootstrap has been acknowledged few in tourism studies, it is, as far as our knowledge, not yet recognised in tourist accommodation data. This study attempts to investigate the characteristics of the tourist accommodation data and seek for an appropriate statistical model to support the development of high-quality tourist accommodation statistics.

3 METHODOLOGY AND DATA

The current study aims to identify the most appropriate statistical model that can estimate reliable tourist accommodation statistics. In order to develop the national accommodation statistical database, the Japan Tourism Agency (JTA), under the direction of the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) of Japan, conducted a national survey focusing on the operating accommodation facilities. Clustered sampling was utilised relying on the sampling frame of 49,850 accommodation facilities, i.e. hotels, registered in Japan nationwide (JTA, MLIT, 2016). The clusters were set up as prefectural administrative territories. The accommodation survey was expected to obtain 35% of total hotels in the sampling frame, wherein hotels for survey were randomly selected at each cluster. The survey questionnaires; beside the items for trip purpose, guest categories, etc., included four major items to ask for the total guests stayed in one month, the number of rooms, capacity of the hotels, and the number of employees working for the hotels. The survey obtained 10,530 respondents in total. After removing the missing values of the four major items, the data achieved 10,503 tourist accommodation facilities (i.e. hotels) throughout Japan.

The outcome variable was the number of guests (Guests.Persons) staying at each property in December 2016. Data also included three other variables, namely the number of rooms, guest capacity, and number of employees. During the DGP, many zero counts were observed, with each indicating that no guests stayed at given hotels during the 1-month study period. Table 1 provides a statistical summary of the data; 6.3% of the hotels had no guests. The highest number of guests at a single hotel was almost 100,000; the variance (3,230^2 = 104.33 105) differed significantly from the mean (1,681.29), suggesting that the data were zero-inflated, overdispersed, and long-tailed. Figure 1, truncated at value of 5,000, shows a histogram of the response variable

Table 1. Definition and statistical summary of variables of interest

Variable name	Explanations	Mean	Standard deviation	Min–Max
Guests.Persons	Total number of guests staying at the accommodation properties in December 2016	1,681.29	3,230.00	0–98,281
Rooms	Total number of rooms in the accommodation properties	65.76	103.95	1–3,560
Capacity	Capacity of the accommodation properties, defined as the total number of guests that can be accommodated simultaneously	150.99	228.40	2–6,424
Employees	Number of employees at the accommodation properties	36.04	74.11	1–1,964

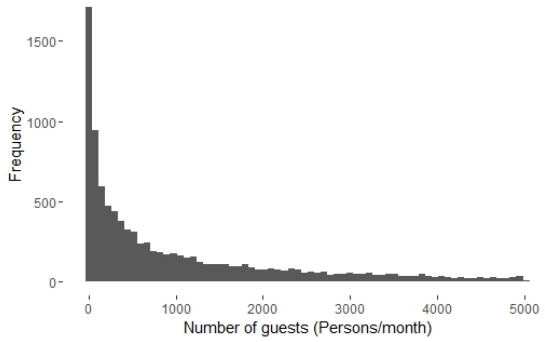
The first step in the analysis is to determine which explanatory variable is the most important for obtaining a good model fit. To achieve this, we used Pearson's correlation coefficient between the explanatory and outcome variable [r=cov(x,y)/√[var(x)var(y)]. The results, shown in Figure 2, indicated that the outcome variable was most strongly correlated with Rooms (r = 0.89; 95% confidence interval [CI], 0.886–0.894), followed by Capacity (r = 0.88; 95% CI, 0.873–0.882) and Employees (r = 0.68; 95% CI, 0.673–0.693). Furthermore, Rooms, Capacity, and Employees were strongly correlated with each other, suggesting a high possibility of multicollinearity if they were included together in the same model. Thus, to avoid the negative effects of multicollinearity on model precision and bias, only Rooms was retained as an explanatory variable; the other two variables were excluded. The correlation between the outcome variable (Guests.Persons) and explanatory variable (Rooms) was linear (Figure 2); furthermore, in the count regression, the link function is in logarithmic form, in that the mean of the response variable is expressed according to the explanatory variable on an exponential scale. The explanatory variable was transformed into a logarithmic scale to maintain the linear correlation between the variables. In addition, in the scatter plot Guests.Persons ~ Rooms (the top-left scatter plot panel in Figure 2), the variance of outcome variable, Guests.Persons, tended to vary in wider range as Rooms increased, suggesting the existence of heterogeneity. The treatment of heterogeneity was also taken into account in the specification of count model.

The procedure for bootstrap resampling and regression analyses was as follows.

Step 1: To identify the best-fitting model, the bootstrap method (B. Efron, 1979) was used to create a random subsample of data with a size equal to that of the original sample. The training set and test set, which were 85% and

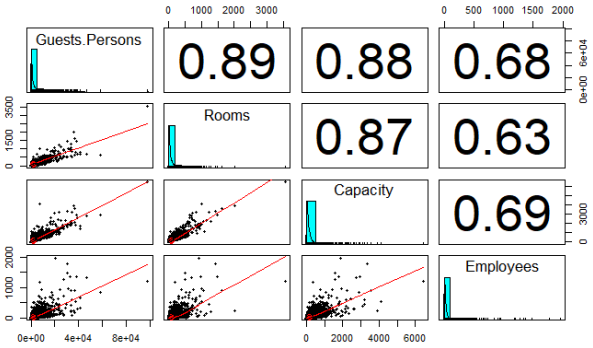
15% the size of the original sample, respectively, were derived by randomly splitting the subsample.

Figure 1. Distribution of guests (truncated at 5000) staying at the surveyed accommodation properties, December 2016



Step 2: In total, six models, which represented various combinations of three types of generalised count regression model (one-part, zero-inflated, and hurdle), with one of two distributions (Poisson or NB), were fitted to the training set. The model parameters were derived by a log-likelihood estimator. As described in the preceding section, the data were zero-excessive, long-tailed, and overdispersed. The Poisson distribution, having a single parameter, may not be sufficiently flexible to describe real-world data. The NB distribution, a well-known gamma–Poisson mixture, contains two parameters, shape and scale, that have additional flexibility for describing discrete, real-world data. Nevertheless, Poisson distribution models were also taken into account, to ensure that the most suitable model was identified.

Figure 2. Pearson's correlation coefficients for the outcome variable and explanatory variables



Integrating the explanatory variable (Rooms) into Poisson and NB distributions allows the mean value of guests at hotels to vary at different levels of Rooms, which partially controls for heterogeneity. In practice, the accommodation data in this study showed a long-tailed distribution, existence of heterogeneity, and the zero counts were excessive. Thus, we employed two-part models, modified from basic count models, consisting of hurdle and zero-inflated parts; such models were first introduced by Mullahy (1986) to analyse the National Survey of Personal Health Practices and Consequences, Wave II (USA) (Mullahy, 1986) and Australian Health Survey 1977–1978 (Mullahy, 1997) data; the ability of the model to capture both underdispersion and overdispersion was demonstrated in their works. Furthermore, the heterogeneity was also modelled by allowing the dispersion parameter varied with the explanatory variable.

The hurdle model considers the probability, f1(0), of zeros; and (1− f1(0)) *f2(y) / (1 − f2(0)), which is associated with truncation, denotes the probability of positive counts. The hurdle model allows different processes for specifying zero and the truncated part. The zero-inflated model uses a separate component for calculating the probability of a zero count, in the same fashion as the hurdle model. In this study, in both the hurdle and zero-inflated models, we applied a logarithmic link function to the part of the model concerned with zero counts, and Poisson and NB distributions to the part concerned with positive counts.

Estimates of the model parameters and the criteria for model evaluation, e.g. the Akaike information criterion (AIC=-2lnL+2k; Akaike, 1973), Bayesian information criterion (BIC=-2lnL+(ln *n*)k) and log-likelihood criterion were obtained. The raw residual (RR) were used as basic goodness-of-fit criteria. The heterogeneity of the accommodation data was explored with studentised and standardised residual analyses. The analysis was implemented with R language. R package glmmTMB of Brooks et al. (2017) was utilised for count regression.

Step 3: Using the model obtained in Step 2, the difference between the total value of estimated- ($\hat{y}_i$) and observed- ($\bar{y}_i$) guests in the test set ($\frac{\{\sum \hat{y}1-\sum \bar{y}1\}}{\sum \bar{y}1}$) can be employed as an indicator of model performance.

We repeated Steps 1–3 with B bootstrap iterations to yield the distribution of the variables of interest and their 95% CIs. Some researchers have suggested a minimum number of iterations when using the bootstrap method to construct the CI. Hall (1986) argued that B should depend on the sample size and precision of the CIs. Simar & Wilson (2007) took B to be 2,000 in their two-stage, semi-parametric regression model of 322 US banks.

In this study, we used 10,000 bootstrap iterations to fit each model, and for model validation to ensure reliable simulation results. Model selection was based on comparison of the AIC, Bayesian information criterion (BIC), log-likelihood values among the six models. The RR analyses were performed to determine the goodness of fit of the models. Data heterogeneity was detected using a studentised residual values, derived by dividing the RR by residual standard deviation, and standardised residuals, derived by dividing the RR by its standard deviation. Given that the AIC, BIC, and log-likelihood are all Gaussian, a two-sample t-test was preferable for comparison of their mean values.

4 RESULTS AND ANALYSIS

A single parameter distribution, i.e. Poisson distribution, was less suitable for the dataset compared with a two-parameter distribution, i.e. NB. Three models with a Poisson distribution, namely Poisson, zero-inflated Poisson, and hurdle Poisson models, showed a poor fit; their AIC, BIC, and RMS values were much higher than those of models with an NB distribution, while their log-likelihood values were much lower (Table 2 and Figure 3). The t-test was used to

compare model evaluation criteria among NB, zero-inflated NB (ZINB), and hurdle NB (HNB) models; the NB model showed a poorer fit compared with the ZINB and HNB models. The HNB model was slightly better than the ZINB model in terms of the AIC value, which was 13.1 lower (two-sample t-test, p = 0.086), as was the BIC (two-sample t-test, p = 0.086); meanwhile, the log-likelihood was 6.54 higher (two-sample t-test, p = 0.086).

Table 2. Evaluation criteria of three models with Poisson distributions

	Poisson	Zero-inflated Poisson (ZIP)	Hurdle Poisson (HP)
AIC (95% CI) 10^6	4.57 (4.33 to 4.85)	4.31 (4.07 to 4.58)	4.31 (4.07 to 4.57)
BIC (95% CI) 10^6	4.57 (4.33 to 4.85)	4.31 (4.07 to 4.58)	4.31 (4.07 to 4.57)
Log-likelihood (95% CI) 10^6	−2.29 (−2.42 to −2.17)	−2.16 (−2.29 to −2.03)	−2.15 (−2.28 to −2.03)

AIC: Akaike information criterion; BIC: Bayesian information criterion; CI: confidence interval.

Figure 3. Model evaluation criteria of three models with a negative binomial (NB) distributionnegbin: negative binomial (NB); hurdle.negbin: hurdle NB (HNB); zeroinfl.negbin: zero-inflated NB (ZINB); AIC: Akaike information criterion; BIC: Bayesian information criterion

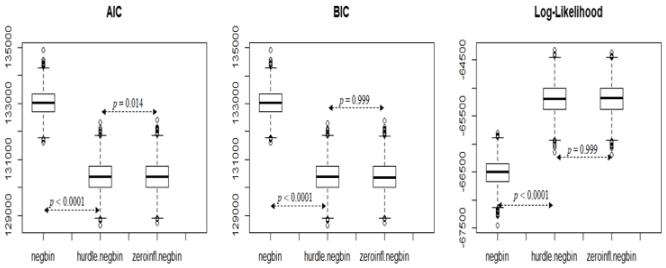
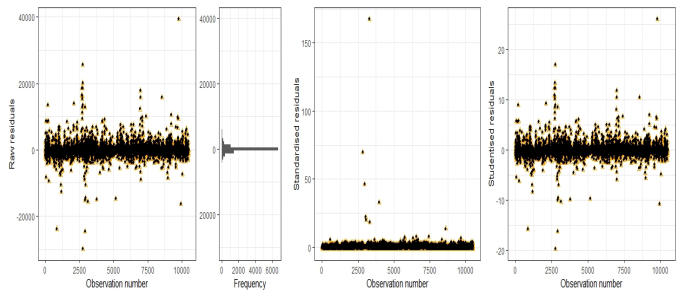


Figure 4. Residual distribution of the HNB and ZINB model. Black-circle points: Residuals of HNB; Orange-triangle points: Residuals of ZINB



The standardised residual analysis results, shown in Figure 4, illustrate that many hotels had a markedly higher or lower number of guests than average (i.e. relative to other hotels with the same number of rooms). We detected 721 (6.8%), 377 (3.6%), and 18 (0.2%) observations with RRs that were 1.5, 2, and 5 times higher, respectively, than the standard deviation. Especially, there were 6 observations with RRs that were 20 to 170 times higher than the standard deviation. Studentised residual analysis (Figure 4) indicated that there were approximately 604 (5.7%), 345 (3.3%), and 55 (0.5%) observations with RRs that were 1.5, 2.0, and 5.0 times higher, respectively, than the standard deviation values. Some cases (16) had exceptionally high RRs, i.e. 10 times higher than the standard deviations. These results clearly confirm the significant heterogeneity in such count data. The difference between residuals of HNB and ZINB were not statistically significant, and the results of heterogeneity detection based on RRs of HNB and ZINB were consistent.

Table 3. Coefficients of the HNB and ZINB model

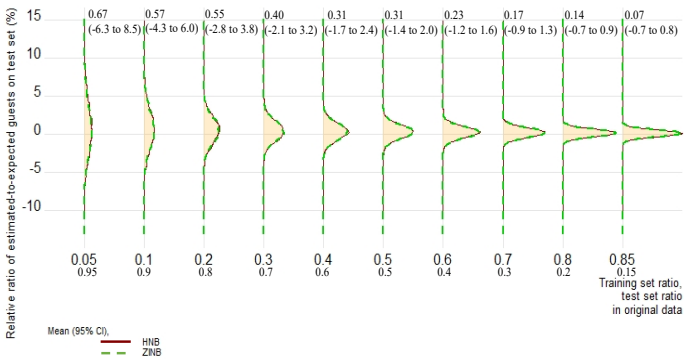
	Part 1: Conditional model		Part 2: Zero-inflated model		Dispersion model	
	Intercept	log (Rooms)	Intercept	log (Rooms)	Intercept	log (Rooms)
HNB	2.85*** (2.77 to 2.95)	1.08*** (1.06 to 1.09)	0.55*** (0.34 to 0.76)	-1.27*** (-1.38 to -1.18)	-1.44*** (-1.66 to -1.26)	0.57*** (0.52 to 0.62)
ZINB	2.85*** (2.77 to 2.95)	1.08*** (1.06 to 1.09)	-0.13 (-0.46 to 1.07)	-1.06*** (-1.18 to -0.95)	-1.44*** (-1.66 to -1.26)	0.57*** (0.52 to 0.62)

*, **, ***: Statistically significant at $p < 0.05$, 0.01 and 0.0001, respectively.
HNB: hurdle negative binomial; *ZINB*: zero-inflated negative binomial; *CI*: confidence interval.
(95% CI): The figures in the parentheses indicate 95% confidence interval.

To test the performance of the HNB and ZINB models in estimating the total guest population with various-sized samples, 10 scenarios were validated with 10 different sizes of test sets corresponding to 10 different sizes of training set obtained from randomly splitting the original sample. The two models were fitted to the training sets, which varied from 5% (525 observations) to 85% (8,925 observations); the estimates of total guest population were yielded by applying the fitted models to test sets that contained the remaining samples of the original data. Generally, increasing the sample size of the training set led to more accurate estimates of the total number of guests of the test set.

For a training set containing 85% of the original sample data, the estimated total number of guests, in average, differed by 0.07% (95% CI, −0.7% to 0.8%) from the observed total. When the training set contained only 5% (525 hotels) of the data from the original sample, the validation demonstrated good performance (0.67% difference between estimated and observed number of guests; 95% CI, −6.3% to 8.5%). Both HNB and ZINB illustrated the same capability in validation process (Figure 5). It was also found that, based on the estimates of coefficients of HNB and ZINB presented as in Table 3, the dispersion of tourist accommodation data was significant and it increased as the scale (e.g. the number of rooms) of the hotels increased.

Figure 5. Performance of the HNB and ZINB models with respect to validation of the test set



5 DISCUSSION AND IMPLICATIONS

Although several studies have applied count data analysis to tourism (Grogger & Carson, 1991; Martínez-Espiñeira & Amoako-Tuffour, 2008; Shrestha et al., 2002; Yen & Adamowicz, 1993), few have specifically examined tourist accommodation data using this method. This study provided

some insight into the complexities of count data analysis when applied to tourism accommodation data. It was found that the accommodation count data were over-dispersed, zero-excessive, and long-tailed. High heterogeneity was recognised in the data. The two-parameter distribution, i.e. NB integrated into a two-part hurdle (HNB) and zero-inflated (ZINB) models outperformed the other count models. The HNB and ZINB slightly overestimated the number of guests. Calibration refers to how accurately a model estimates the size of a population, as measured by the ratio between the estimated and observed population.

The total number of guests, in average, was found to be overestimated by 0.07% (95% CI, −0.7% to 0.8%) with the test set that used 85% of the original sample, indicating excellent validation performance. Regarding the sample size, previous studies using count data analysis relied on the sampling of several hundreds to thousands of samples. For instance, Pohlmeier & Ulrich (1995), Deb & Trivedi (1997), Bulmer (1974), Solis-Trapala & Farewell (2005), and Taylor (1967) analysed sample sizes of 5,096, 4,406, 924, 651, and 623, respectively. Especially, Arbous & Sichel (1954) analysed absenteeism data based on only 248 observations. In this study, we analysed the data of 10,503 hotels throughout Japan, to ensure reliability and accuracy of the analysis.

The HNB and ZINB models could be important for the further development of tourist accommodation statistics. In some countries, such as France (Insee, 2017), Spain (INE, 2017), the United Kingdom (Visitbritain, 2018), Japan (JTA, MLIT, 2016), Thailand (NSO-Thailand, 2016), and Vietnam (GSO-VN, 2011), the total number of guests staying in tourist accommodation facilities are estimated using a linear estimator. i.e. the mean of guests stayed obtained from sample data multiplied by the "expansion rate", which is defined as the ratio between the total population size and the sample size.

Accommodation data are typically stratified, where different strata represent accommodation or region types (e.g. provinces or, for lower strata, cities and/or districts). Estimates of the total number of guests per stratum are commonly required. The linear estimator can be applied in the same manner to each stratum; this method is easy to implement but may be less precise, or even infeasible for lower strata if data samples for these strata are not available or are missing. Regression can be used to overcome these issues, given its ability to control for heterogeneity to some extend and interpolate the total count for a given stratum based on the available data.

Heterogeneity was noted in the accommodation data used. The reasons for the heterogeneity have yet to be elucidated in this study; however, there are several plausible explanations. For example, the data were obtained from various areas of Japan, in which tourism demand may differ according to seasonality. Even within the same region, hotels vary in terms of design and business strategy, which may lead to a degree of heterogeneity. For instance, December may be a high-demand period in some areas, whereas in other areas having hotels of the same size, occupancy may be low, or they may even be empty.

As alluded to above, the variety in accommodation types may also lead to heterogeneity. For example, a traditional type of accommodation known as "Ryokan" has large rooms that can accommodate many guests simultaneously, as can the accommodation facilities of sports or training centres, such as large huts in mountainous areas designed for hikers and mountain climbers; these can in fact host several dozen to hundreds of people in one room and thus differ considerably from regular hotels, in which rooms are commonly designed for single-, double-, or triple-occupancy. Additionally, some hotels provide substantial discount packages, whereas others do not. Interestingly, two-part count models can distinguish zero demand from non-zero demand for lodging units, as well as uncover errors due to heterogeneity. Including more explanatory variables and/or stratifying the data can minimise heterogeneity.

The bootstrap resampling method can play a key role in model selection and validation, and in detecting overestimation. The AIC, BIC, and log-likelihood values (Figure 3) indicated that one may make a wrong selection between the HNB and the ZINB model if their selection criteria were compared based on single-time modelling. Furthermore, it can be difficult for the HNB and ZINB models to detect under- or overestimation with single-time validation. Fitting and validating model with arbitrary B bootstrap sub-samples (in this study, B = 10,000), representing the vast of possible samples of the population, provides B models.

With each model, one set of model coefficients, model selection and model validation criteria were derived. The six models generated herein were compared in terms of the AIC, BIC, and log-likelihood values by t-test (Figure 3), and the results tended to support the HNB and ZINB models (Figure 5); moreover, overestimation was well-recognised by the bootstrap method, suggesting the estimate of population total, i.e. total guests, should be corrected. Bradley Efron & Tibshirani (1994:138) recommended a method to correct the parameters of interest, which a bias-corrected estimator is determined by subtracting the estimate of population total by the estimate of bias. For example, let $\hat{G}$ denotes for the estimate of total guests, $\widehat{(bias)}$ denotes for the estimate of bias (e.g. 0.5%), then the bias-corrected estimate of total guests would be $\bar{G} = \hat{G}(1-\widehat{(bias)}) = 0.995\hat{G}$.

Non-econometric HNB and ZINB models should prove useful for deriving reliable accommodation statistics. We recommend the integrating bootstrap resampling and count regression for reliable tourist accommodation statistics which strongly support decision making process in tourism planning, management, monitoring and evaluating the policy implementations.

The findings of this study, on one hand, could be used directly to estimate the total guests stayed by relying on a random sample drawn from the finite population of accommodation facilities, which is widely known as design-based estimation method. On the other hand, the estimates of parameters derived in the analysis are useful for feeding the model-based estimation method which the inferential framework recommended by Fisher & Russell (1922). The model-based approach may appear as the exclusive choice for administrative territories where there is no respondent. The results of this study, although derived from the tourist accommodation survey as a case study in Japan, but other countries may be able to adapt since the analysis procedure could be easily generalised universally.

6 CONCLUSIONS AND FURTHER RESEARCH

In conclusion, the results of this analysis of a large dataset show that tourist accommodation count data are highly over-dispersed, zero-excessive, and long-tailed. Heterogeneity is common in tourist accommodation data. The HNB and ZINB models appear to be appropriate models for such data. In particular, the HNB and ZINB models, which were fitted to a training set that contained 85% of the original data, resulted little overestimate, but high accuracy; The estimated-to-expected relative ratio of the total number of guests was 0.07% (95% CI, −0.65% to 0.77%). Even when reducing the training set to include only 5% of the data of the original sample, the overestimate, in average, was only 0.67% (95% CI, −6.30% to 8.50%). The bootstrap technique is particularly useful for detecting overestimates. This suggests that overestimation could be resolved by adjusting the estimated value by the overestimate ratio. Thus, overestimation is not considered to be a major problem for tourist accommodation data.

Some countries (GSO-VN, 2011; INE, 2017; Insee, 2017; JTA, MLIT, 2016; NSO-Thailand, 2016; Visitbritain, 2018), as stated previously, have been using a linear estimator to estimate the population total (i.e. total guests), Australia (ABS, 2016) have been utilising time series analysis to estimate the total guests based on tourist accommodation survey. The estimation of total guests of a month of the later year will be inferred based on the information of the same month of the previous year with updated information, for example the annual percentage change of guests, the seasonal adjustment, etc. The limitation of this study is that the performance of the count regression has not been compared with the other methods have been applying in such countries in various circumstances (e.g. small and large population, alternatives of sampling schemes). Further research needs to be carried out to elaborate those issues.

ACKNOWLEDGEMENTS

The study was completed in the framework of tourist accommodation statistical database joint-research. We acknowledge the data provision of Japan Tourism Agency (JTA) and the financial support of Nippon Foundation.

REFERENCES

ABS. (2016, November 25). *The Survey of Tourist Accommodation* (STA). http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8635.0Explanatory%20Notes12015-16?OpenDocument. Accessed 3/12/2019, at 12:35.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 1973, 268–281.

Alves, J., Alves da Silva, A., Soares, A.M.V.M., & Fonseca, C. (2013). Pellet group count methods to estimate red deer densities: Precision, potential accuracy and efficiency. *Mammalian Biology - Zeitschrift Für Säugetierkunde*, 78(2), 134–141. https://doi.org/10.1016/j.mambio.2012.08.003

Arbous, A.G., & Sichel, H.S. (1954). New techniques for the analysis of absenteeism data. *Biometrika*, 41(1/2), 77–90. https://doi.org/10.2307/2333007

Assaf, A., Barros, C.P., & Josiassen, A. (2010). Hotel efficiency: A bootstrapped metafrontier approach. *International Journal of Hospitality Management*, 29(3), 468–475. https://doi.org/10.1016/j.ijhm.2009.10.020

Assaf, A.G., & Agbola, F.W. (2011). Modelling the performance of Australian hotels: A DEA double bootstrap approach. *Tourism Economics*, 17(1), 73–89. https://doi.org/10.5367/te.2011.0027

Bortkiewicz, L. von. (1898). *Das Gesetz der Kleinen Zahlen*. Leipzig: B.G. Teubner.

Brooks, M.E., Kristensen, K., Benthem, K.J. van, Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Mächler, M., & Bolker, B.M. (2017). Modeling zero-inflated count data with glmmTMB. *BioRxiv*, 132753. https://doi.org/10.1101/132753

Bulmer, M.G. (1974). On fitting the poisson lognormal distribution to species-abundance data. *Biometrics*, 30(1), 101–110. https://doi.org/10.2307/2529621

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19(9), 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<114

Chen, R., & Fomby, T.B. (1999). Forecasting with stable seasonal pattern models with an application to Hawaiian tourism data. *Journal of Business & Economic Statistics*, 17(4), 497–504. https://doi.org/10.2307/1392408

Chou, M.C. (2013). Does tourism development promote economic growth in transition countries? A panel data analysis. *Economic Modelling*, 33, 226–232. https://doi.org/10.1016/j.econmod.2013.04.024

Deb, P., & Trivedi, P.K. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12(3), 313–336. https://doi.org/10.1002/(SICI)1099-1255(199705)12:3<313::AID-JAE440>3.0.CO;2-G

Deb, P., & Trivedi, P.K. (2002). The structure of demand for health care: Latent class versus two-part models. *Journal of Health Economics*, 21(4), 601–625. https://doi.org/10.1016/S0167-6296(02)00008-5

DiCiccio, T.J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–212.

Duarte, R., & Escario, J.J. (2006). Alcohol abuse and truancy among Spanish adolescents: A count-data approach. *Economics of Education Review*, 25(2), 179–187. https://doi.org/10.1016/j.econedurev.2005.01.007

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. Boca Raton: Chapman and Hall/CRC Press.

Fisher, R.A., & Russell, E.J. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Fonton, N.H., & Hounkonnou, N. (2011). *Promoting, Fostering and Development of Statistics in Developing Countries*. Berlin: Springer-Verlag.

Gergaud, O., Livat, F., & Song, H. (2018). Terrorism and wine tourism: The case of museum attendance. *Journal of Wine Economics*, 13(4), 375–383. https://doi.org/10.1017/jwe.2018.41

Grogger, J.T., & Carson, R.T. (1991). Models for truncated counts. *Journal of Applied Econometrics*, 6(3), 225–238. https://doi.org/10.1002/jae.3950060302

GSO-VN. (2011). *Tourist Expenditure Survey, Viet Nam 2011*. https://www.gso.gov.vn/default.aspx?tabid=763&ItemID=13625. Accessed 5/11/2019, at 11:42.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics*, 14(4), 1453–1462. https://doi.org/10.1214/aos/1176350169

Hand, D.J. (2008). *Statistics: A Very Short Introduction (Vol. 196)*. Oxford: Oxford University Press.

Hausman, J.A., Hall, B.H., & Griliches, Z. (1984). *Econometric Models for Count Data with an Application to the Patents-R&D Relationship* (Working Paper No. 17). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/t0017

INE. (2017, February). Hotels: Occupancy survey, price index and profitability indicators and methodology. http://www.ine.es/en/daco/daco42/ocuphotel/meto_eoh_en.pdf. Accessed 16/12/2019, at 10:17.

Insee. (2017). Statistical Processing—Hotel Attendance Survey 2017, Insee. https://www.insee.fr/fr/metadonnees/source/operation/s1457/processus-statistique. Accessed 17/11/2019, at 13:32.

JTA, MLIT. (2016). *Accommodation Statistics Survey*. Tokyo: Japan Tourism Agency. http://www.mlit.go.jp/kankocho/siryou/toukei/shukuhakutoukei.html. Accessed 5/11/2019, at 11:50

King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential poisson regression model. *American Journal of Political Science*, 32(3), 838–863. https://doi.org/10.2307/2111248

Martínez-Espiñeira, R., & Amoako-Tuffour, J. (2008). Recreation demand analysis under truncation, overdispersion, and endogenous stratification: An application to Gros Morne National Park. *Journal of Environmental Management*, 88(4), 1320–1332. https://doi.org/10.1016/j.jenvman.2007.07.006

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12(3), 337–350. https://doi.org/10.1002/(SICI)1099-1255(199705)12:3<337::AID-JAE438>3.0.CO;2-G

NSO-Thailand. (2016). Hotel and guest house survey 2016. http://www.nso.go.th/sites/2014en/Pages/survey/Economics/Tourism-and-Sports.aspx. Accessed 12/11/2019, at 16:21.

Palmer Pol, A., Pascual, M.B., & Vázquez, P.C. (2006). Robust estimators and bootstrap confidence intervals applied to tourism spending. *Tourism Management*, 27(1), 42–50. https://doi.org/10.1016/j.tourman.2004.06.016

Pohlmeier, W., & Ulrich, V. (1995). An econometric model of the two-part decision-making process in the demand for health care. *The Journal of Human Resources*, 30(2), 339–361. https://doi.org/10.2307/146123

Poisson, S.D. (1837). Recherches sur la probabilité des jugements en matière criminelle et en matière civile precédées des règles générales du calcul des probabilités. Paris: Bachelier.

Sanga, D. (2011). Role of statistics: Developing country perspective. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1262–1263). Vienna: Springer Verlag. https://doi.org/10.1007/978-3-642-04898-2_63

Shrestha, R.K., Seidl, A.F., & Moraes, A.S. (2002). Value of recreational fishing in the Brazilian Pantanal: A travel cost analysis using count data models. *Ecological Economics*, 42(1), 289–299. https://doi.org/10.1016/S0921-8009(02)00106-4

Simar, L., & Wilson, P.W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136(1), 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009

Solis-Trapala, I.L., & Farewell, V.T. (2005). Regression analysis of overdispersed correlated count data with subject specific covariates. *Statistics in Medicine*, 24(16), 2557–2575. https://doi.org/10.1002/sim.2121

Taylor, P.J. (1967). Individual variations in sickness absence. *Occupational and Environmental Medicine*, 24(3), 169–177. https://doi.org/10.1136/oem.24.3.169

Visitbritain. (2018). Accommodation Occupancy: Latest results. VisitBritain. https://www.visitbritain.org/accommodation-occupancy-latest-results. Accessed 10/11/2019, at 16:37.

WTTC. (2019). *Economic Impact*. London: World Travel & Tourism Council. https://wttc.org/Research/Economic-Impact. Accessed 16/11/2019, at 13:52.

Yen, S.T., & Adamowicz, W.L. (1993). Statistical properties of welfare measures from count-data models of recreation demand. *Applied Economic Perspectives and Policy*, 15(2), 203–215. https://doi.org/10.2307/1349443

Zepp, R. (2011). Selection of appropriate statistical methods in developing countries. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1294–1295). Vienna: Springer Verlag. https://doi.org/10.1007/978-3-642-04898-2_68

SUBMITTED: MAR. 2019
REVISION SUBMITTED: OCT. 2019
2nd REVISION SUBMITTED: JAN. 2020
ACCEPTED: MAR. 2020
REFEREED ANONYMOUSLY

PUBLISHED ONLINE: 30 MAY 2020