Enhancing Tourist Forecasting in Thailand's National Parks with Zero-Inflated Models

SAMACH SATHITVUDH[1], PIYADA WONGWIWAT[2*], WIKANDA PHAPHAN[3,4]
[1]Department of Statistics, School of Computer, Data & Information Sciences
University of Wisconsin–Madison, Madison, WI 53705,
USA
[2]Department of Computer Science and Data Innovation, Faculty of Science and Technology,
Suan Sunandha Rajabhat University, Bangkok 10300,
THAILAND
[3]Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok 10800,
THAILAND
[4]Research Group in Statistical Learning and Inference,
King Mongkut's University of Technology North Bangkok, Bangkok 10800,
THAILAND
*Corresponding author

Abstract: This article focuses on predicting the number of tourists visiting Thailand's national parks using count data models. Given the discrete and overdispersed nature of the tourist count data, traditional Poisson regression models were extended to include Negative Binomial (NB) and Zero-Inflated models. Using data from 2016 to 2022, we evaluated four model types: Poisson, Negative Binomial, Zero-Inflated Poisson, and Zero-Inflated Negative Binomial (ZINB). Model performance was assessed using the Akaike Information Criterion (AIC), log-likelihood values, and the Vuong test. Findings reveal that the ZINB model best fits the data, addressing both overdispersion and excess zeros, resulting in more accurate predictions. This model is thus recommended for similar count data applications in tourism and environmental studies. Future work may focus on optimizing the model by reducing complexity and improving outlier handling.

Key-Words: Count data, Regression model, Poisson, Negative Binomial, Zero-inflated model, Overdispersion, Tourists

Received: July 29, 2024. Revised: February 9, 2025. Accepted: March 16, 2025. Published: April 15, 2025.

1 Introduction

The idea of national parks establishment had emerged after World War II in which there was a tremendous increase in population leading to jungles and forest destruction. The issue of wild animal extinction had been raised by that time until the government realized and determined to tackle the problem seriously in order to conserve the natural environment, especially for rare species of plants as well as breeds of wild animals. Having national parks established also significantly drives economic development due to the beauty of nature which leads to various types of business and service.

Due to the rapid growth of tourism in many regions in Thailand. A large number of tourists, both local and international, are captivated to visit the national parks since they offer numerous activities and services that most people will enjoy. That being said, in order to effectively welcome and provide appropriate facilities for visitors under a limited budget, knowing an approximate number of tourists visiting a specific national park in a specific time frame helps plan the strategies, financial management, and many more for those who visit the park. Statistical modeling plays a crucial role in predicting the interesting outcome, number of tourists/visitors in this case. Thus, it is also vital to select the best model among various model candidacies that explains and predicts the count data based on the data we possess in our hands.

In this article, the author has fetched the open data from the governmental data center. This involves the number of counts of national park visitors between 2016 and 2022 categorized by years, months, and

the national parks which are the primary and obvious factors that affect the number of tourists. These three variables are considered to be independent variables while it is clear that the number of tourists visiting the national parks is response variable. There are 146 national parks included in the data set in which they are encoded as dummy variables.

Since the response variable is related to the count data which is discrete and non-negative integers, the most direct approach to study the relationship is to explore the models that are able to model the integer-valued data. This primarily includes the Poisson (POI) and Negative Binomial (NB) regressions which require an assumption that the response variable follows the Poisson and Negative Binomial distributions respectively.

The main objective is to compare the effectiveness of the count modeling for the number of tourists visiting national parks. We have applied four different-sophisticated regression models. This ranges from the Poisson (POI) Regression Model, Negative Binomial (NB) Regression Model, and Zero-inflated Poisson (ZIP) Regression Model to Zero-inflated Negative Binomial (ZINB) Regression Model. Ultimately, they are compared using three distinct model evaluation methods including the Log-likelihood ratio test, Akaike Information Criterion, and Vuong Test.

2 Methodology

2.1 Poisson (POI) Regression Model

The model was developed by [1], [2]. It is a building block for the count regression model where the Poisson distribution has been assumed by the model, [3]. Let y_i be independent Poisson random variables with parameter λ_i. The probability mass function for y_i given x_i is defined as

$$f(y_i|\mathbf{X}_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{1}$$

Note that $\mathbb{E}[y|x_i] = \mathbb{V}[y_i|x_i] = \mu_i$ where μ_i can be computed by $\exp(\mathbf{X}^\intercal\boldsymbol{\beta})$. In general, the POI model can expressed as the log-likelihood function. Under the independence assumption, we have

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(-\exp^{\mathbf{X}_i^\intercal\boldsymbol{\beta}} - y_i\mathbf{X}_i^\intercal\boldsymbol{\beta} - \ln y_i!\right) \tag{2}$$

The main applications for this specific count modeling are ubiquitous and have been exposed for a long history in various fields such as [4], [5], [6], [7].

2.2 Negative Binomial (NB) Regression Model

The model was invented by [8], which was derived from a mixture between Poisson and Gamma distribution due to the restriction of the definition of the density function for negative binomial. This model compromises the assumption of having both expectation and variance coincide. Thus, the model alleviates modeling the Poisson-distributed dependent variable whose variance is greater than its mean. The probability mass function is defined below:

$$f(y_i|\mathbf{X}_i) = \frac{\Gamma\left(y_i + \frac{1}{\theta}\right)}{\Gamma(\frac{1}{\theta})y_i!}\left(\frac{1}{1+\theta\lambda_i}\right)^{\frac{1}{\theta}}\left(\frac{\theta\lambda_i}{1+\theta\lambda_i}\right)^{y_i} \tag{3}$$

where λ_i is the mean and θ is the overdispersion parameter defined by

$$\mathbb{E}[y_i|\mathbf{X}_i] = \lambda_i \text{ and } \mathbb{V}[y_i|\mathbf{X}_i] = \lambda_i(1+\theta\lambda_i)$$

Note that when $\theta = 0$, the above coincides with the Poisson distribution. According to the independence assumption, this model can be written as a log-likelihood function as

$$\begin{aligned}\ln L(\boldsymbol{\beta}) = &\sum_{i=1}^{n}\ln\left[\Gamma\left(y_i+\frac{1}{\theta}\right)\right] - \sum_{i=1}^{n}\ln y_i! \\ &- \sum_{i=1}^{n}\ln\left[\Gamma\left(\frac{1}{\theta}\right)\right] + \sum_{i=1}^{n} y_i\ln\theta + \sum_{i=1}^{n} y_i\mathbf{X}_i^\intercal\boldsymbol{\beta} \\ &- \sum_{i=1}^{n}\left(y_i+\frac{1}{\theta}\right)\ln\left(1+\theta e^{\mathbf{X}_i^\intercal\boldsymbol{\beta}}\right)\end{aligned} \tag{4}$$

The main applications for the NB regression model are also of interest to numerous fields and have been studied by many researchers such as [9], [10], [11].

2.3 Overdispersion

Overdispersion refers to the phenomenon where the empirical observations have higher variance than the theoretical model. Otherwise, it is called underdispersion phenomenon. Overdispersion usually occurs when modeling simple parametric model where the variance cannot be independently adjusted, [12]

Once overdispersion is mentioned, this relates to the zero-inflation circumstance. Hence, we consider the zero-inflated models which are commonly used in count data analysis with an excessive number of zero counts. This type of model embraces various types of situations of observing many zeros. Two of them are studied in this paper.

2.4 Zero-inflated Poisson (ZIP) Regression Model

The ZIP model combines Poisson and a degenerate distribution at zero, [13]. The independent random

variable y_i breaks into two different conditions with different outcomes

$$y_i \sim \begin{cases} k_{i0} & \text{, with probability } \pi_i \\ \text{Poisson}(\lambda_i) & \text{, with probability } 1-\pi_i \end{cases} \quad (5)$$

Then the probability mass function is defined as

$$y_i \sim \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i} & ; y_i = 0 \\ (1-\pi_i)\frac{e^{-\lambda_i}\lambda^{y_i}}{y_i} & ; y_i > 0 \end{cases} \quad (6)$$

where $\lambda_i > 0$ is the rate parameter, and $0 \leq \pi_i \leq 1$ denote the probability of i-th observation. Given $\mathbf{X}_i$, we obtain the expectation and variance as

$$\begin{aligned} \mathbb{E}[y_i|\mathbf{X}_i] &= \lambda_i(1-\pi_i) \\ \mathbb{V}[y_i|\mathbf{X}_i] &= \lambda_i(1-\pi_i)(1+\pi_i\lambda_i) \end{aligned}$$

When modeling the Zero-inflated data, [9], describe that the parameters λ_i and π_i satisfy the following

$$\begin{aligned} \ln(\lambda_i) &= \mathbf{B}_i^\intercal \boldsymbol{\beta} \\ \text{logit}(\pi_i) &= \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{G}_i^\intercal \boldsymbol{\gamma} \end{aligned} \quad (7)$$

where we denote $\mathbf{B}_i^\intercal$ and $\mathbf{G}_i^\intercal and$ as the vectors of independent variables in covariate matrices $\mathbf{B}$ and $\mathbf{G}$ respectively. We can also express the ZIP regression model as a log-likelihood function when π_i is not a function of λ_i as follows:

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\sum_{i:y_i=0} \ln\left(e^{\mathbf{G}_i^\intercal\boldsymbol{\gamma}} + e^{-e^{\mathbf{X}_i^\intercal\boldsymbol{\beta}}}\right) \\ &- \sum_{i:y_i=0} \ln\left(1+e^{\mathbb{G}_i^\intercal\boldsymbol{\gamma}}\right) - \sum_{i:y_i>0} \ln(y_i!) \\ &+ \sum_{i:y_i>0} \left(y_i\mathbf{X}_i^\intercal - e^{\mathbf{X}_i^\intercal\boldsymbol{\beta}}\right) \end{aligned} \quad (8)$$

2.5 Zero-inflated Negative Binomial (ZINB) Regression Model

Analogous to the NB regression model, the ZINB was developed by [14], who considers that there is an excessive number of zeros while there exists an overdispersion phenomenon. Thus, it is suitable for fitting the response variable y_i whose variance is greater than its expectation. Suppose $y_i \sim \text{ZINB}(\lambda_i, \theta, \pi_i)$. Then the probability mass function is defined as

$$f(y_i|\mathbf{X}_i) = \pi_i + (1-\pi_i)\left(\frac{1}{1+\theta\lambda_i}\right)^{\frac{1}{\theta}} \quad (9)$$

when $y_i = 0$ and

$$\begin{aligned} f(y_i|\mathbf{X}_i) = &(1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)y_i!}\left(\frac{1}{1+\theta\lambda_i}\right)^{\frac{1}{\theta}} \\ &\times\left(\frac{\theta\lambda_i}{1+\theta\lambda_i}\right)^{y_i} \end{aligned} \quad (10)$$

when $y_i > 0$. Given $\mathbf{X}_i$, the expectation and variance are computed by

$$\begin{aligned} \mathbb{E}[y_i|\mathbf{X}_i] &= \lambda_i(1-\pi_i) \\ \mathbb{V}[y_i|\mathbf{X}_i] &= \lambda_i(1-\pi_i)(1+\pi_i\lambda_i+\theta\lambda_i) \end{aligned}$$

To implement the ZINB regression model, we express it as a log-likelihood function for both cases, $y_i = 0$ and $y_i > 0$ as follows: For $y_i = 0$,

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\sum_{i=1}^{n} \ln\left(e^{\mathbb{G}_i^\intercal\boldsymbol{\gamma}} + \left(\frac{1}{1+\theta e^{\mathbf{X}_i^\intercal\boldsymbol{\beta}}}\right)^{\frac{1}{\theta}}\right) \\ &- \sum_{i=1}^{n} \ln\left(1+e^{\mathbb{G}_i^\intercal\boldsymbol{\gamma}}\right) \end{aligned} \quad (11)$$

and for $y_i > 0$,

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &-\sum_{i=1}^{n} \ln(1+e^{\mathbb{G}_i^\intercal\boldsymbol{\gamma}}) \\ &+ \sum_{i=1}^{n} \ln\left[\Gamma\left(y_i+\frac{1}{\theta}\right)\right] - \sum_{i=1}^{n}\left[\Gamma\left(\frac{1}{\theta}\right)\right] \\ &- \sum_{i=1}^{n} \ln y_i! + \sum_{i=1}^{n} y_i \ln\theta + \sum_{i=1}^{n} y_i\mathbf{X}_i^\intercal\boldsymbol{\beta} \\ &- \sum_{i=1}^{n}\left(y_i+\frac{1}{\theta}\right)\ln\left(1+\theta e^{\mathbf{X}_i^\intercal\boldsymbol{\beta}}\right) \end{aligned} \quad (12)$$

2.6 Model Comparison Criteria

In this paper, the author has proposed three different evaluation methods to compare the effectiveness of the models. This involves the Akaike Information Criterion (AIC), the Likelihood-ratio (LR) test, and the Vuong test for a non-nested model. The mathematical (and statistical) formulations are explained below:

2.6.1 Akaike's Information Criterion (AIC)

The AIC is an estimator of the prediction error as well as the quality of the statistical models for a given data set, [15], invented in 1973. This criterion mainly focuses on how much information is lost by a model which is widely used in information theory

and statistical inference. More formally, let q be the number of estimated parameters in the model and $L(\cdot)$ be the maximum likelihood function. The AIC value is computed by [16].

$$\text{AIC} = 2q - 2\ln L(\cdot) \tag{13}$$

The rule of thumb for the AIC value is that the lower the value, which implies that the model is able to maintain more information, the higher the quality of the model.

2.6.2 Likelihood-ratio Test (LRT)

When comparing two candidate statistical models, the LRT (Wilk's test) plays a crucial role in comparing the goodness of fit between these models. Two hypotheses are posed by the complexity of the model in which the null hypothesis constrains the model to a simpler version whereas the alternative hypothesis is specified under the more complex model. More formally,

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta$$

where the statistical models with a parameter space Θ. The null hypothesis is interpreted as the parameter θ lies in a more constrained subset Θ_0 of Θ. By rejecting the null hypothesis, it means that the more complex model explains the ground truth of the data more effectively.

The log-likelihood ratio test statistics under the null hypothesis H_0 is given by [17].

$$\Lambda_{LR} = -2\left[\ln \sup_{\theta\in\Theta_0} L(\theta) - \ln \sup_{\theta\in\Theta} L(\theta)\right] \tag{14}$$

Since the numerator is constrained, it cannot exceed the denominator. Therefore, the ratio is bounded in $[0, 1]$.

2.6.3 Vuong Test for Non-nested Model

Analogous to the LRT, the Vuong test [18] uses the Kullback–Leibler information criterion to construct the likelihood-ratio-based test for two candidate models. It focuses on identifying whether these two models are approximately equally close to the ground truth of the data. More formally, the hypotheses are specified as:

H_0 : two models are equally comparable
H_1 : the alternative model is closer
to the ground truth data

Under the strictly non-nested models, the Vuong test statistic is defined by

$$V = \frac{\sum_{i=1}^{n} m_i}{\sqrt{n} s d_m} \tag{15}$$

where $m_i = \ln\left[\frac{f_1(y_i|\mathbf{X}_i)}{f_2(y_i|\mathbf{X}_i)}\right]$ and f_i's are the likelihood functions of model 1 and 2 respectively. Furthermore, the Vuong test for non-nested models is well known for comparing the zero-inflated count model to its non-zero-inflated one although there is some argument claiming that the test has not been yet globally valid due to the nestedness.

3 Empirical Results

As the main objective is to evaluate the models' efficiency for predicting the number of tourists visiting the national parks. In this section, the empirical findings are presented in four subsections; the descriptive analysis, the models, the model comparison, and the prediction.

3.1 Descriptive Analysis

Under 12,264 total observations, the number of tourists visiting national parks in Thailand ranges from 0 to 675,818. Overall, from 2016 to 2022, there are 16.01%, 17.94%, 18.36%, 17.97%, 12.30%, 6.83%, and 10.59% of the number of tourists visiting the parks respectively. Table 1 illustrates the percentage of tourists visiting the national parks in different aspects; by month (descending order). Table 2 describes the top 10 most visited national parks in proportion. Lastly, Table 3 illustrates the Cramer's V coefficient for each covariate X_1, X_2 and X_3 which correspond to year, national park (dummy), and month respectively whereas the response variable Y is the number of tourists.

Table 1. Proportion of tourists visiting the national parks for each month

Month	Proportion
December	14.63%
January	11.99%
April	10.87%
February	9.84%
March	9.83%
November	8.47%
October	8.23%
July	6.73%
May	5.57%
August	5.50%
September	4.48%
June	3.87%

According to the Cramer's V coefficient table, it shows a high association between the response variable (the number of tourists) and each of the predictors (year, dummy of park, and month). Meanwhile, there exists a small multicollinearity between the year and the dummy variable of the park.

Table 2. Top 10 national parks visited

Park name	Proportion
Khao Yai	9.48%
Hat Noppharat Thara (Phi Phi)	7.75%
Khao Leam Ya - Mo Ko Samet	5.98%
Phang Nga Bay	5.75%
Doi Inthanon	4.63%
Khao Khitchakut	4.22%
Erawan	3.41%
Namtok Phlio	3.25%
Mu Ko Similan	3.23%
Khao Sok	2.07%

Table 3. Cramer's V coefficient

Variable	X_1	X_2	X_3	Y
X_1	1	0.211	0	0.788
X_2	0.211	1	0	0.739
X_3	0	0	1	0.74

3.2 Modeling Results

In this subsection, the resulting models with parameter estimates $\boldsymbol{\beta}$ are provided. For the non-zero-inflated model such as POI and NB, the proposed model is the following:

$$\begin{aligned}\ln(\lambda_i) =&\beta_0 + \beta_{1_j}\text{Year}_j + \beta_{2_k}\text{National}_k \\ &+ \beta_{3_l}\text{Month}_l\end{aligned} \tag{16}$$

where $j \in \{1, ..., 6\}, k = \{1, ..., 145\}, l = \{1, ..., 11\}$ and β's for each model are shown in Table 4, Table 5, Table 6, and Table 7 On the other hand for both zero-inflated regression models, two separate models are considered. One is similar to 16 with different estimated parameters. The other part which considers inflation is formulated as

$$\text{logit}(\pi_i) = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_{1_l}\text{Month}_l \tag{17}$$

where $l = \{1, ..., 11\}$.

Therefore, Table 4, Table 5, Table 6, and Table 7 illustrate the estimation of parameters (non-zero/count parts) from all four proposed models, POI, NB, ZIP, and ZINB. Note that the models have assigned one level for each variable to be baselines. This includes `year7`, `national146` and `month12`.

Since the models ZIP and ZINB contain zero part, Table 8 displays the estimation of parameters from the zero-inflated models specifically the zero part. Note that the zero parts are obtained from `month` variable where `month12` is the baseline.

3.3 Model Comparison

After obtaining proper models for predicting the number of tourists visiting the national parks, one must investigate systematically which one is the most effective model using AIC and the log-likelihood.

Table 4. Estimated coefficients and p-values from Poisson and Negative Binomial Regression Models

Parameters	POI	NB Regression	ZIP	ZINB
Intercept	10.9163	11.5455	10.8723	11.4013
β_{1_1}	0.5051	0.1823	0.4792	0.0898
β_{1_2}	0.3337	-0.3922*	0.3263	-0.3831
β_{1_3}	0.4932	0.2376	0.4784	0.1724
β_{1_4}	0.5281	0.2265	0.5027	0.1334
β_{1_5}	0.149	-0.1506	0.3997	0.1689
β_{1_6}	-0.4384	-0.645	-0.3232	-0.2636
β_{2_1}	-3.6356	-3.6148	-3.6133	-3.5931
β_{2_2}	-1.5123	-1.5054	-1.5203	-1.5725
β_{2_3}	-6.3596	-6.3277	-6.2063	-6.2021
β_{2_4}	-2.8388	-2.8954	-2.8319	-2.8849
β_{2_5}	-6.014	-6.1893	-5.9815	-6.1414
β_{2_6}	-4.075	-4.0876	-4.0648	-4.0703
β_{2_7}	-5.1485	-5.3946	-5.1276	-5.3473
β_{2_8}	-4.1226	-4.3498	-4.0501	-4.2157
β_{2_9}	-2.648	-2.9448	-2.5834	-2.7713
$\beta_{2_{10}}$	-2.9489	-3.0525	-2.9378	-3.0193
$\beta_{2_{11}}$	0.2123	0.0312*	0.2097	0.1297*
$\beta_{2_{12}}$	-1.2335	-1.1687	-1.2337	-1.2011
$\beta_{2_{13}}$	-4.2523	-4.3359	-4.2659	-4.3874
$\beta_{2_{14}}$	-5.3464	-5.3662	-5.2667	-5.2145
$\beta_{2_{15}}$	-1.6085	-1.5711	-1.5845	-1.5446
$\beta_{2_{16}}$	-3.689	-3.7553	-3.686	-3.7405
$\beta_{2_{17}}$	-0.9474	-0.9469	-0.913	-0.9056
$\beta_{2_{18}}$	-2.5116	-2.5956	-2.5036	-2.5711
$\beta_{2_{19}}$	-0.5021	-0.4822	-0.5082	-0.5065
$\beta_{2_{20}}$	-1.3182	-1.2619	-1.3171	-1.3026
$\beta_{2_{21}}$	-4.1913	-4.1906	-4.1938	-4.274
$\beta_{2_{22}}$	-2.1941	-2.2305	-2.186	-2.2184
$\beta_{2_{23}}$	-2.1994	-2.343	-2.1649	-2.2369
$\beta_{2_{24}}$	-2.34	-2.4411	-2.3233	-2.4203
$\beta_{2_{25}}$	0.5608	0.6485	0.5527	0.5798
$\beta_{2_{26}}$	1.0487	1.1936	1.0374	1.1007
$\beta_{2_{27}}$	-1.5662	-1.5345	-1.5663	-1.5858
$\beta_{2_{28}}$	-5.7596	-5.9397	-5.7516	-5.8757
$\beta_{2_{29}}$	-0.8882	-0.8358	-0.8802	-0.8626
$\beta_{2_{30}}$	-2.9628	-2.9911	-2.9547	-3.0207
$\beta_{2_{31}}$	-0.8795	-0.9984	-0.8635	-0.9649
$\beta_{2_{32}}$	-2.3089	-2.2296	-2.3025	-2.2567
$\beta_{2_{33}}$	-2.7534	-2.6806	-2.7543	-2.7577
$\beta_{2_{34}}$	-3.5755	-3.5292	-3.5661	-3.5707
$\beta_{2_{35}}$	-4.3964	-4.3631	-4.3488	-4.2744
$\beta_{2_{36}}$	-5.0273	-5.2552	-5.006	-5.1817
$\beta_{2_{37}}$	-1.693	-1.8783	-1.677	-1.8036
$\beta_{2_{38}}$	-2.4642	-2.6838	-2.4611	-2.6471
$\beta_{2_{39}}$	-3.2905	-3.3395	-3.2643	-3.3388

* refers to insignificant variables under the $\alpha = 0.05$.

Table 5. Estimated coefficients and p-values from Poisson and Negative Binomial Regression Models

Parameters	POI	NB Regression	ZIP	ZINB
$\beta_{2_{40}}$	-5.634	-5.8245	-5.5806	-5.6766
$\beta_{2_{41}}$	-0.7536	-0.7922	-0.7506	-0.7707
$\beta_{2_{42}}$	-2.9302	-3.044	-2.7719	-2.8682
$\beta_{2_{43}}$	0.3015	0.0904*	0.3301	0.1974*
$\beta_{2_{44}}$	-4.0315	-3.8451	-3.9997	-3.8922
$\beta_{2_{45}}$	-0.9231	-1.1299	-0.834	-0.9462
$\beta_{2_{46}}$	-3.0117	-3.2928	-2.9871	-3.2074
$\beta_{2_{47}}$	-0.616	-0.5916	-0.6135	-0.6316
$\beta_{2_{48}}$	-3.4807	-3.5571	-3.4646	-3.5141
$\beta_{2_{49}}$	-5.3999	-5.4049	-5.371	-5.3446
$\beta_{2_{50}}$	-3.5475	-3.6469	-3.5139	-3.5544
$\beta_{2_{51}}$	-1.7587	-1.97	-1.7648	-1.9652
$\beta_{2_{52}}$	-2.7892	-2.9071	-2.7505	-2.8475
$\beta_{2_{53}}$	-3.7288	-4.0211	-3.7067	-3.9186
$\beta_{2_{54}}$	-1.828	-1.9124	-1.8225	-1.9145
$\beta_{2_{55}}$	-3.8403	-3.7775	-3.8208	-3.758
$\beta_{2_{56}}$	-2.2303	-2.2116	-2.2101	-2.2168
$\beta_{2_{57}}$	-2.2137	-2.3094	-2.2012	-2.2889
$\beta_{2_{58}}$	-2.4077	-2.1966	-2.3692	-2.1768
$\beta_{2_{59}}$	-1.5681	-1.5498	-1.5762	-1.6045
$\beta_{2_{60}}$	-1.3091	-1.3452	-1.3116	-1.3446
$\beta_{2_{61}}$	-1.7292	-1.6785	-1.7268	-1.6577
$\beta_{2_{62}}$	-4.3338	-4.6923	-4.3308	-4.6103
$\beta_{2_{63}}$	-4.155	-3.9866	-4.0888	-3.8886
$\beta_{2_{64}}$	-3.3334	-3.3571	-3.3415	-3.4105
$\beta_{2_{65}}$	-0.445	-0.4587	-0.4309	-0.42
$\beta_{2_{66}}$	-3.7092	-3.8598	-3.6585	-3.7724
$\beta_{2_{67}}$	-2.906	-2.7623	-2.8451	-2.7117
$\beta_{2_{68}}$	-1.7928	-1.7121	-1.7438	-1.6366
$\beta_{2_{69}}$	-0.0492	-0.028*	-0.0518	-0.063*
$\beta_{2_{70}}$	-1.8035	-1.9979	-1.7322	-1.8285
$\beta_{2_{71}}$	-4.3903	-4.6529	-4.3427	-4.4467
$\beta_{2_{72}}$	-1.967	-2.0176	-1.9474	-1.97
$\beta_{2_{73}}$	-2.4782	-2.4067	-2.4519	-2.3597
$\beta_{2_{74}}$	-3.719	-3.7926	-3.7251	-3.8418
$\beta_{2_{75}}$	-1.8541	-1.8546	-1.8294	-1.7943
$\beta_{2_{76}}$	-3.1686	-3.2255	-3.156	-3.204
$\beta_{2_{77}}$	-2.5294	-2.3291	-2.5141	-2.3803
$\beta_{2_{78}}$	-2.8203	-3.0804	-2.7913	-2.9664
$\beta_{2_{79}}$	-3.7993	-3.7279	-3.7746	-3.7197
$\beta_{2_{80}}$	-3.0355	-2.8622	-2.967	-2.8334
$\beta_{2_{81}}$	-2.1532	-2.0541	-2.1239	-1.9915
$\beta_{2_{82}}$	-1.6524	-1.2564	-1.6371	-1.3357
$\beta_{2_{83}}$	-2.7148	-2.8941	-2.6778	-2.7863
$\beta_{2_{84}}$	-0.9558	-1.0062	-0.9533	-0.9951
$\beta_{2_{85}}$	-3.7678	-3.8894	-3.744	-3.8514

* refers to insignificant variables under the $\alpha = 0.05$

Table 6. Estimated coefficients and p-values from Poisson and Negative Binomial Regression Models

Parameters	POI	NB Regression	ZIP	ZINB
$\beta_{2_{86}}$	-1.876	-2.2138	-1.8314	-2.0555
$\beta_{2_{87}}$	-4.3008	-4.3423	-4.2847	-4.3042
$\beta_{2_{88}}$	-1.7685	-1.6223	-1.7484	-1.6218
$\beta_{2_{89}}$	-1.7047	-1.8463	-1.6871	-1.8231
$\beta_{2_{90}}$	-2.4965	-2.6016	-2.4839	-2.5152
$\beta_{2_{91}}$	-3.351	-3.2292	-3.3467	-3.3173
$\beta_{2_{92}}$	-4.3748	-4.2605	-4.3546	-4.3011
$\beta_{2_{93}}$	-5.8085	-5.8761	-5.7761	-5.8308
$\beta_{2_{94}}$	-4.9231	-5.1277	-4.907	-5.0072
$\beta_{2_{95}}$	-1.1777	-1.477	-1.1696	-1.4016
$\beta_{2_{96}}$	-2.0439	-1.8154	-1.991	-1.8325
$\beta_{2_{97}}$	-1.6403	-1.6089	-1.6257	-1.6273
$\beta_{2_{98}}$	-3.0901	-2.8258	-3.0744	-2.926
$\beta_{2_{99}}$	-5.0023	-4.9711	-4.9698	-4.9393
$\beta_{2_{100}}$	-3.9065	-4.1005	-3.887	-4.0489
$\beta_{2_{101}}$	-3.2564	-3.2883	-3.2404	-3.2394
$\beta_{2_{102}}$	-0.6797	-0.8726	-0.6814	-0.8388
$\beta_{2_{103}}$	-4.176	-4.5431	-4.1149	-4.3922
$\beta_{2_{104}}$	-3.9575	-4.0567	-3.9414	-3.9885
$\beta_{2_{105}}$	-2.6941	-2.766	-2.6563	-2.6358
$\beta_{2_{106}}$	-6.4544	-6.8285	-6.1393	-6.3237
$\beta_{2_{107}}$	-2.8335	-3.0682	-2.8073	-2.9676
$\beta_{2_{108}}$	-3.7215	-3.5679	-3.7167	-3.7109
$\beta_{2_{109}}$	-4.0744	-4.5299	-4.0159	-4.3432
$\beta_{2_{110}}$	-4.382	-4.5828	-4.3802	-4.5966
$\beta_{2_{111}}$	-2.3205	-2.3249	-2.3116	-2.3666
$\beta_{2_{112}}$	-4.8025	-4.9321	-4.7964	-4.9141
$\beta_{2_{113}}$	-2.1013	-2.2387	-2.0514	-2.1204
$\beta_{2_{114}}$	-3.2035	-3.2783	-3.1373	-3.1619
$\beta_{2_{115}}$	-2.7677	-2.8976	-2.722	-2.8005
$\beta_{2_{116}}$	-4.8764	-5.1718	-4.6146	-4.7658
$\beta_{2_{117}}$	-1.7063	-1.7681	-1.6712	-1.6809
$\beta_{2_{118}}$	-3.4747	-3.4174	-3.45	-3.3861
$\beta_{2_{119}}$	-2.8421	-2.8565	-2.8346	-2.8945
$\beta_{2_{120}}$	-3.4556	-3.6562	-3.3482	-3.4845
$\beta_{2_{121}}$	-1.6415	-1.9588	-1.5998	-1.8385
$\beta_{2_{122}}$	-3.4217	-3.5403	-3.396	-3.4745
$\beta_{2_{123}}$	-1.9808	-1.9127	-1.9728	-1.93
$\beta_{2_{124}}$	-3.7355	-3.8988	-3.6765	-3.7687
$\beta_{2_{125}}$	-4.5104	-4.5153	-4.448	-4.4124
$\beta_{2_{126}}$	-6.0085	-6.3699	-5.8318	-6.0099
$\beta_{2_{127}}$	-2.624	-2.6952	-2.6302	-2.7056
$\beta_{2_{128}}$	-0.8554	-0.9051	-0.8499	-0.8873
$\beta_{2_{129}}$	-2.6997	-2.5939	-2.7022	-2.67
$\beta_{2_{130}}$	-1.8335	-1.7151	-1.8305	-1.7836
$\beta_{2_{131}}$	-6.8521	-7.0412	-6.7211	-6.8419

* refers to insignificant variables under the $\alpha = 0.05$.

Table 9 shows the empirical results for AIC and log-likelihood values for each particular model. This shows that the Zero-inflated Negative Binomial regression model appears to be the most appropriate

Table 7. Estimated coefficients and p-values from Poisson and Negative Binomial Regression Models

Parameters	POI	NB Regression	ZIP	ZINB
$\beta_{2_{132}}$	-1.1731	-1.3899	-1.165	-1.2959
$\beta_{2_{133}}$	-0.0562	-0.4331	0.187	0.079
$\beta_{2_{134}}$	-2.4229	-2.7151	-2.1797	-2.2597
$\beta_{2_{135}}$	-1.6155	-1.4737	-1.5386	-1.4021
$\beta_{2_{136}}$	-1.2651	-1.5629	-1.249	-1.4471
$\beta_{2_{137}}$	-2.2718	-2.1736	-2.2557	-2.2756
$\beta_{2_{138}}$	-1.7203	-1.891	-1.7209	-1.8988
$\beta_{2_{139}}$	0.8171	0.8515	0.8146	0.8625
$\beta_{2_{140}}$	-2.6257	-2.5668	-2.6152	-2.6123
$\beta_{2_{141}}$	-2.9554	-3.1157	-2.9393	-3.09
$\beta_{2_{142}}$	-2.4289	-2.5304	-2.3843	-2.473
$\beta_{2_{143}}$	-2.5904	-2.7628	-2.5657	-2.6846
$\beta_{2_{144}}$	0.5195	0.5428	0.5225	0.5864
$\beta_{2_{145}}$	-2.3617	-2.1842	-2.3282	-2.1368
β_{3_1}	-0.1989	-0.3351	-0.1997	-0.2942
β_{3_2}	-0.3966	-0.7611	-0.3965	-0.7426
β_{3_3}	-0.3974	-0.8536	-0.3921	-0.7943
β_{3_4}	-0.2968	-0.4785	-0.1244	-0.1841
β_{3_5}	-0.9661	-1.6039	-0.7637	-1.1474
β_{3_6}	-1.3331	-1.9616	-1.067	-1.4652
β_{3_7}	-0.7765	-1.2745	-0.6657	-0.9671
β_{3_8}	-0.9777	-1.4986	-0.888	-1.2309
β_{3_9}	-1.1832	-1.5288	-1.1077	-1.343
$\beta_{3_{10}}$	-0.5745	-0.8417	-0.5625	-0.7731
$\beta_{3_{11}}$	-0.5459	-0.6362	-0.5379	-0.6665

* refers to insignificant variables under the $\alpha = 0.05$.

Table 8. Estimated coefficients from Zero-inflated Poisson and Negative Binomial Regression Models

Parameters	ZIP	ZINB
Intercept	-5.8488	-5.8256
γ_{3_1}	-1.0718*	-1.1056*
γ_{3_2}	0.8762*	0.8168*
γ_{3_3}	2.2435	2.2128
γ_{3_4}	4.2443	4.2207
γ_{3_5}	4.7473	4.7235
γ_{3_6}	4.9031	4.8786
γ_{3_7}	4.3198	4.2943
γ_{3_8}	4.0657	4.0382
γ_{3_9}	3.8778	3.8499
$\gamma_{3_{10}}$	2.8829	2.8517
$\gamma_{3_{11}}$	2.2435	2.2077

* refers to insignificant variables under the $\alpha = 0.05$.

regression model for the number of tourists visiting the national parks due to the smallest AIC and the largest log-likelihood. Furthermore, it is proper to determine overdispersion for the response variable and make sure that our ZINB regression model best

Table 9. AIC and Log-likelihood for POI, NB, ZIP, and ZINB Regression Models

Regression models	AIC	Log-likelihood
POI	58650892	-29325283
NB	212652.9	-106162.4
ZIP	47533436	-23766543
ZINB	202669	-101158.5

explains the variation and the data by comparing POI with the NB regression model. Restating the hypotheses test for overdispersion below, the likelihood ratio test is performed.

$$H_0 : \text{POI appropriately fits the data}$$
$$H_1 : \text{NB appropriately fits the data}$$

The result shows that the test statistic is 58438241 and the p-value is sufficiently small to reject the null hypothesis implying that there exists an overdispersion so that the negative binomial family performs better at explaining the phenomenon of a number of tourists visiting the national parks in Thailand.

3.4 Prediction

According to the previous finding, the ZINB regression model is determined to be the most appropriate model for predicting the number of tourists. The entire data set was modeled and the predicted values of the number of tourists were obtained. The Figure 1 shows the line plot comparing the behavior of the actual and predicted values. As

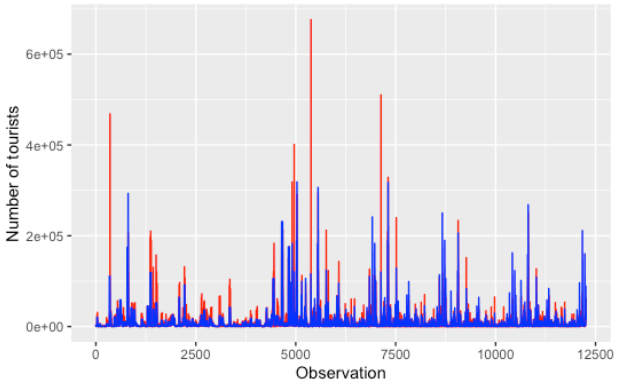

Fig. 1: Prediction of number of tourists (blue) against the actual number of tourists visiting the national parks (red)

shown in the graph, the prediction of the number of tourists (blue) obtained from the ZINB regression model is significantly accurate as the actual values

(red) and the prediction behave similarly although there are a few observations that the prediction is trivially lower than expected. This might be considered later due to the issue of outliers. The box plot (Figure 2) also illustrates another aspect of how the ZINB model performs in predicting the number of tourists. The box plot confirms our claim that the ZINB model performs delightfully as for the most part, the plots are almost similar to each other which provides a satisfactory sign that the model is appropriate.

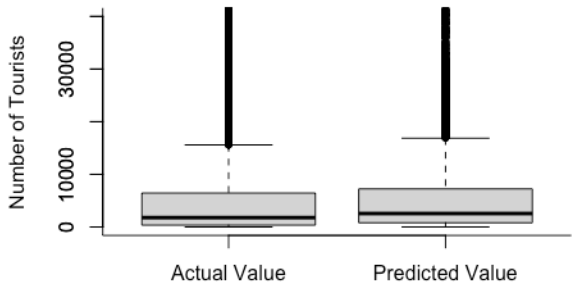

Fig. 2: Box plot of the prediction against the actual values for the number of tourists

Throughout, the RMSE of the ZINB model is $16,616.12$ whereas the RMSE of the NB model is $25,528.13$. This also confirms our claim that the zero-inflated negative binomial regression model outshines all of the proposed count models in terms of AIC, Log-likelihood, and the RMSE itself.

4 Conclusion

This paper evaluates the performance of four regression models, two of which are ordinary count models: Poisson and Negative Binomial regression models. There are two special models on top of the two models that embrace the notion of zero-inflated count models. The models were developed and tested to predict the number of tourists visiting the national park in Thailand between 2016 and 2022.

According to the findings, the Zero-inflated Negative Binomial (ZINB) regression model is the most appropriate model for predicting the number of tourists in this study and is superior to all of the other model candidates. It has been systematically shown that the data is over-dispersed, and requires the negative binomial (NB) regression family to tackle the modeling part. Furthermore, when performing evaluation methods by calculating the AIC and the log-likelihood, it shows that the ZINB wins against the ordinary NB model. For this reason, the prediction of the number of tourists was displayed and proved that this model outperforms the others which is also confirmed by comparing their RMSE.

Among all these situations, there is some room for improvements that future research can take and develop. One is that there exist some limitations in predicting a number of tourists when the actual number is significantly large and the model does not perform impressively. Second, the model might be considered complex since it contains a huge number of variables. It is recommended that grouping dummy variables of national parks based on their location might be a practical and acceptable approach to make the model less complicated and easier to understand. Moreover, we can extend our work to other models of count data like, [19], [20], [21], [22].

Acknowledgments:
The authors are grateful to anonymous referees for the valuable comments, which have significantly improved this article. This research was funded by Suansunandha Rajabhat University.

References:

[1] G. King, "A seemingly unrelated Poisson regression model," *Sociological Methods and Research*, vol. 17, 1989, pp. 235–255.

[2] R. Winkelmann and K. F. Zimmermann, "Count data models for demographic data," *Mathematical Population Studies*, vol. 4, no. 3, 1994, pp. 205–221.

[3] J. D. Elhai, P. S. Calhoun, and J. D. Ford, "Statistical procedures for analyzing mental health services data," *Psychiatry Research*, vol. 160, no. 2, 2008, pp. 129–136.

[4] W. Gardner, E. P. Mulvey, and E. C. Shaw, "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models," *Psychological Bulletin*, vol. 118, no. 3, 1995, pp. 392–404.

[5] P. L. H. Yu, J. S. K. Chan, and W. K. Fung, "Statistical exploration from SARS," *The American Statistician*, vol. 60, no. 1, 2006, pp. 81–91.

[6] J. Grogger, "The deterrent effect of capital punishment: An analysis of daily homicide counts," *Journal of the American Statistical Association*, vol. 85, 1990, pp. 295–303.

[7] W. Phaphan, N. Sangnuch, and J. Piladaeng, "Comparison of the Effectiveness of Regression Models for the Number of Road Accident

Injuries," *Science & Technology Asia*, vol. 28, no. 4, Oct.–Dec. 2023, pp. 54–66. Available: https://tci-thaijo.org/index.php/SciTechAsia

[8] A. C. Cameron and P. K. Trivedi, "Econometric models based on count data: Comparisons and applications of some estimators and tests," *Journal of Applied Econometrics*, vol. 1, no. 1, 1986, pp. 29–53.

[9] P. M. Westgate et al., "Marginal modeling in community randomized trials with rare events: Utilization of the negative binomial regression model," *Clinical Trials*, vol. 19, no. 2, 2022, pp. 162–171.

[10] A. Byers, H. Allore, T. Gill, and P. Peduzzi, "Application of negative binomial modeling for discrete outcomes," *Journal of Clinical Epidemiology*, vol. 56, 2003, pp. 559–564.

[11] H. Liu, R. A. Davidson, D. V. Rosowsky, and J. R. Stedinger, "Negative binomial regression of electric power outages in hurricanes," *Natural Hazards Review*, vol. 11, 2005, pp. 258–267.

[12] J. K. Lindsey, P. M. E. Altham, Analysis of the Human Sex Ratio by Using Overdispersion Models, Journal of the Royal Statistical Society Series C: Applied Statistics, Volume 47, Issue 1, March 1998, Pages 149–157, https://doi.org/10.1111/1467-9876.00103

[13] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, 1992, pp. 1–14.

[14] W. H. Greene, "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models," NYU Working Paper No. EC-94-10, 1994.

[15] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Processing Magazine*, 2004, pp. 36–47.

[16] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, 1974, pp. 716–723.

[17] K. R. Koch, *Parameter Estimation and Hypothesis Testing in Linear Models*. New York: Springer, 1998.

[18] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, no. 2, 1989, pp. 307–333.

[19] A. M. M. Badr, T. Hassan, T. Shams El Din, and F. A. M. Ali, "Zero Truncated Poisson - Pareto Distribution: Application and Estimation Methods," *WSEAS Transactions on Mathematics*, vol. 22, pp. 133–138, 2023. DOI: 10.37394/23206.2023.22.16.

[20] I. Vagelas, "Analysis of Over-Dispersed Count Data: Application to Obligate Parasite Pasteuria Penetrans," *WSEAS Transactions on Environment and Development*, vol. 18, pp. 333–339, 2022. DOI: 10.37394/232015.2022.18.33.

[21] S. Aryuyuen, I. Thaimsorn, and U. Tonggumnead, "Bayesian Inference for a New Negative Binomial-Samade Model for Time Series Data Counts with Its Properties and Applications," *WSEAS Transactions on Mathematics*, vol. 22, pp. 586–600, 2023. DOI: 10.37394/23206.2023.22.65.

[22] W. Panichkitkosolkul, "Bootstrap Confidence Intervals for the Parameter of the Poisson-Prakaamy Distribution with Their Applications," *WSEAS Transactions on Mathematics*, vol. 22, pp. 378–387, 2023. DOI: 10.37394/23206.2023.22.45.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

All authors made equal contributions to the current study.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research was funded by Suansunandha Rajabhat University.

Conflicts of Interest

The authors have no conflicts of interest to declare.