MATEMATIKA: MJIAM, 2020, Volume 36, Number 1, 51–67

Bayesian Mixture Poisson Regression for Modeling Spatial Point Pattern of Primary Health Centers in Surabaya

Tri Murniati*, Nur Iriawan and Dedy Dwi Prastyo

Department of Statistics, Faculty of Science and Data Analitics
Institut Teknologi Sepuluh Nopember
60111 Sukolilo, Surabaya, Indonesia

*Corresponding author: trimurniati1992@gmail.com

Article history
Received: 14 February 2019
Received in revised form: 14 February 2020
Accepted: 12 March 2020
Published online: 1 April 2020

Abstract Spatial Point Pattern describes the spatial location of Primary Health Centers (PHC) in Surabaya. The varying distribution of PHC locations in Surabaya causes its process to follow the Nonhomogeneous Poisson Process (NHPP). The NHPP intensity needs to be modeled to figure out the factors affecting the spread of PHC. The parameter of the NHPP intensity model estimated by building an algorithm based on Bayesian Markov Chain Monte Carlo to model the mixture Poisson regression. The result shows that two mixture components are significantly involved in the model along with four variables i.e., the total population, the number of clean households, the Accessibility Index, and the length of the road. It produces smaller Deviance Information Criterion (DIC) than Poisson regression.

Keywords Bayesian Markov Chain Monte Carlo; Mixture Poisson Regression; NHPP Intensity; Primary Health Centers

Mathematics Subject Classification 62F15, 74E30, 62J02, 62M30

1 Introduction

Spatial point pattern (SPPt) data is a collection of observed datasets based on random points of spatial locations. Many studies of SPPt data have been considered, such as forestry [1], population [2], criminology [3], and property sales [4]. The SPPt modeling can be done by involving covariate variables to determine whether the intensity distribution of the point location depends on the covariate variables [5]. The SPPt analysis is an example of a Poisson process in stochastic modeling. Based on the intensity, the Poisson process is categorized into Homogeneous Poisson Processes (HPP) and Nonhomogeneous Poisson Processes (NHPP). The SPPt classified as HPP has a constant intensity (expected value of a countable random variable) in each area, while NHPP has a spatially varied intensity in each area [6]. The intensity of NHPP that varies spatially can be influenced by the covariate as measured by the regression

parameter coefficient. Thus, the estimated NHPP intensity could be represented as a function of location diversity that is approximated by Poisson mixture regression. Mixture models appear when the measurement of random variables is performed based on two or more different conditions, or the sampling unit consists of sub-populations [7]. Finite mixture models can be either estimated within a frequentist framework or a Bayesian framework [8]. There are many studies of finite mixture models that work on different distribution, such as Mixture of GLMs [8], Bernoulli Mixture Regression Model [9], Mixtures of Normal Distributions [10], and Poisson Mixture Regression Model [11], [12], and [13]. Most of them use Bayesian framework to estimate parameters.

The Bayesian approach coupled with Markov Chain Monte Carlo (MCMC) algorithm is one of the popular methods to estimate parameters in mixture regression modeling. Bayesian models are widely applied to SPPt because they aim to build an intensity function for predicting the posterior distribution of the point pattern. Gibbs sampling is the most commonly used approach and it is done by augmenting the data with the unobservable variable of class membership. By using one of the MCMC methods that is Gibbs Sampling algorithm, the estimation of parameter regression is done based on the sample generated from the posterior distribution. The advantage of using the Bayesian approach in mixture regression is to obtain parameter interval estimation so that the significance of each parameter can be tested. Kusumaningrum, et al. [11] conducted SPPt analysis on the distribution pattern of community health centers in Surabaya. The city of Surabaya is divided based on the 8x4 grid to make up 32 regions with assigned covariate variables attributed to the characteristics of each community health center.

In this study, an analysis of the intensity of NHPP is carried out on points of dispersion location from Primary Health Centers (PHC) in Surabaya. The PHC includes Puskesmas (Community Health Centers) and Klinik Pratama (Privately-run Health Clinic), that is the closest health facility to the community. The existence of PHC becomes very important to provide accessible health services. The different characteristics of each sub-district, however, cause the spatial variation of the intensity of health facilities in Surabaya. Two dimensions of access need to be considered in the provision of health facilities namely availability or number of health facilities in a region and accessibility in relation to the distance between health facilities locations [14]. Surabaya has 31 sub-districts with a total population of 2.8 million people [15]. Ideally, the number of Puskesmas in the city of Surabaya should be at least 95; unfortunately, the available number at this time is only 63 [16]. The government has attempted to meet the needs of the health services profession by inviting private sectors to participate in developing health services by establishing Klinik Pratama. With the participation of private parties, the city of Surabaya has overcome the shortage of health facilities. However, the development of these private health facilities tends to be located near the city center. Hence, they are inaccessible to those who live far from the city center.

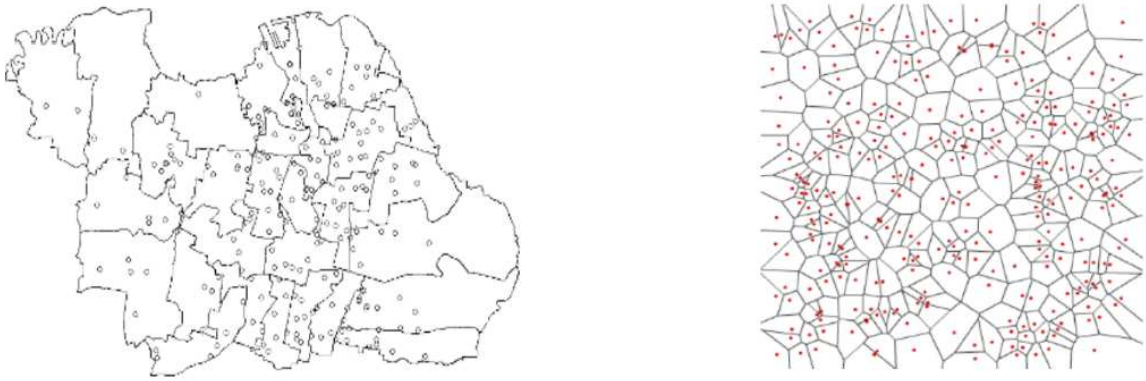
The balanced distribution of health facilities in urban areas is at the core of health services improvement in major cities. The exploration of spatial relationships between the location of the health center with the characteristics of the surrounding environment becomes an important factor for decision-makers, city planners, and healthcare stakeholders. Decisions about the location of health facilities are essential to the provision of health services as a basic need for the population. However, the variation in health facility coverage is frequently related to geographic characteristics of a population, their economic activities, and the different accessibility of each

region. The Geographic Information System (GIS), on the other hand, has a potential role in assessing the distribution of health services, particularly the effectiveness of coverage of health facilities relevant to population density. Various GIS analysis tools have been used extensively to analyze the distribution patterns of the existing health facilities and to find the new optimal locations [17]. The development of modeling for the allocation of health facilities through GIS allows researchers to model access dimensions using spatial statistical models. The GIS can help to explore the availability of PHC based on environmental characteristics such as population geography condition, population economic status, and accessibility [18]. GIS data can be used as the basic information in determining the type of health services needed by the surrounding community and as a consideration to determine the location of the new PHC.

2 Methodology

2.1 Spatial Point Process on R^d

Spatial point processes Y is a countable random variable from a space S where it is assumed that $S \subseteq R^d$. Spatial point processes are very useful as statistical models in analyzing point patterns, where points indicate the location of the study object (tree in the forest, birds nest, the case of disease, or crime). In practice, it can be observed as a point that is confined to a rectangular window space or irregular shape. The window object can be a tessellation of a polygon list describing the division of the administrative area of a state into states or provinces or can also be formed based on point data. It is called Dirichlet or Voronoi tessellation [5]. Figure 1(a) describes the city of Surabaya that is observed as a combination of objects of tessellations with sub-district boundaries. While Figure 1(b) describes the division of windows formed based on the Dirichlet or Voronoi tessellation function.



(a) (b)

Figure 1: (a) Tessellation Object City of Surabaya and (b) Voronoi Tessellation

2.2 Mixture Poisson Regression

The Poisson regression model is an example of a Generalized Linear Models (GLMs) with response Y, given the covariate $\mathbf{x}$, and has Poisson density function as follows:

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} I_A(y). \tag{1}$$

In (1), it would fulfill that $E(Y) = \lambda$ and has the link function $g(\lambda) = \log(\lambda) = \boldsymbol{x}\boldsymbol{\beta}$. Suppose $\mathbf{A} = \{0, 1, 2, \dots\}$ is the set of non-negative integers, and $I_{\mathbf{A}}(y)$ is the indicator function, that is $I_{\mathbf{A}}(y) = 1$ if y belongs to set $\mathbf{A}$ and $I_{\mathbf{A}}(y) = 0$ otherwise. The variance of Y is $\text{var}(Y) = E(Y) = \lambda$ and there exists an overdispersion when $\text{var}(Y) > E(Y)$ [19]. The Poisson Mixture Regression Model belongs to a GLMs, where the i-th response can be expressed as follows:

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\Phi}) = \sum_k w_k p_k(y_{i_k}|\boldsymbol{x}_{i_k}, \boldsymbol{\beta}_k), \tag{2}$$

$$p_k(y_{i_k}|\boldsymbol{x}_{i_k}, \boldsymbol{\beta}_k) = \frac{e^{-e^{\boldsymbol{x}_{i_k}\boldsymbol{\beta}_k}}(e^{\boldsymbol{x}_{i_k}\boldsymbol{\beta}_k})^{y_{i_k}}}{y_{i_k}!} T_{i_k}(y_i), \tag{3}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{w}, \boldsymbol{\beta}\}$, $\boldsymbol{w} = \{w_1, w_2, \dots, w_K\}$ as proportions of mixture components, K is the number of mixture component and for each k, $p_k(y_{i_k}|x_{i_k}, \boldsymbol{\beta}_k)$ is the density, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K\}$, $\boldsymbol{\beta}_k = \{\beta_{0k}, \beta_{1k}, \dots, \beta_{Qk}\}^T$ as the regression parameters, $\boldsymbol{x}$ is a $n \times Q$ matrix of independent variables, y_{i_k} and $\boldsymbol{x}_{i_k}$ are the i-th observations which come from the k-th component of mixture, Q is the number of covariate variables, n is number of observation, $i = 1, 2, \dots, n$, $k = 1, 2, \dots, K$, $i_k = 1, 2, \dots, n_K$, $n_1 + n_2 + \cdots + n_K = n$ and n_k is number observations in the k-th component of mixture. For the K-mixture component, it would have:

$$E(Y_i) = \sum_{k=1}^{K} w_k \lambda_{i_k}, \tag{4}$$

$$0 \leq w_k \leq 1 \text{ and } \sum_{k=1}^{K} w_k = 1,$$

$$\text{var}(Y_i) = E[\text{var}(Y_i|\boldsymbol{H}_k)] + \text{var}(Y_i|\boldsymbol{H}_k) = E(Y_i) + v_{i_k}, \tag{5}$$

where $\lambda_{i_k} = \exp(\boldsymbol{x}_{i_k}\boldsymbol{\beta}_k)$ is the average of the i-th response in the k-th component of mixture or $\lambda_{i_k} = (\lambda_k)_i$, and $\boldsymbol{H}_k$ is the component indicator vector of zeros and ones with $H_{i_k} = (H_k)_i$. It worths one when y_i is belong to the k-th component and zeros for the other, $v_{i_k} = 0$ when $\lambda_{i_1} = \lambda_{i_2} = \cdots = \lambda_{i_K}$ [19]. When we use logit or log-linear links to model H_{i_k}, it would be set one with probability $\pi_k(x)$. Therefore, it could represent λ_{i_k} which is the regression vector coefficients of the i-th response in the k-th component typically given as equation (6).

$$\text{logit}\big(\pi_k(x)\big) = \log\left(\frac{\pi_k(x)}{1 - \pi_k(x)}\right) = \boldsymbol{x}_{i_k}\boldsymbol{\beta}_k,$$

$$\log(\lambda_{i_k}) = \boldsymbol{H}_k^T \boldsymbol{\gamma}_k, \tag{6}$$

for $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ are unknown parameters. If there is Q covariate variables recorded from all spatial location (u_i) with K-components, then the form of the spatial Poisson mixture regression becomes as equation (7).

$$\begin{aligned}\hat{y}_i =& w_1\big[\log\big(\exp[\beta_{01} + \beta_{11}x_{1i_1}(u_{i_1})]\big)\big] + \\ & w_2\big[\log\big(\exp[\beta_{02} + \beta_{12}x_{1i_2}(u_{i_2})]\big)\big] + \cdots + \\ & w_K\big[\log\big(\exp[\beta_{0K} + \beta_{1K}x_{1i_K}(u_{i_K})]\big)\big].\end{aligned} \tag{7}$$

2.3 Bayesian Method

Bayesian statistics differ from the classical statistical theories because all unknown parameters in Bayesian are considered as random variables. The Bayesian analysis required the initial prior distribution derived from the information available previously available. The analysis was performed to obtain a posterior distribution based on observational data after it multiplies with the priors [20]. The relationship between the posterior distribution with the prior and likelihood distributions can be written as follows:

$$\text{Posterior} \quad \text{Distribution} \quad = \quad \text{likelihood} \quad \times \quad \text{Prior} \quad \text{Distribution}$$

If there are $\boldsymbol{\lambda}$ parameters given by the data information y which follows a Poisson distribution, then the probability posterior distribution for $\boldsymbol{\lambda}$ given y will be proportional to the multiplication between prior $\boldsymbol{\lambda}$ and the likelihood function given by y data. Based on the Bayes theorem, the posterior distribution is obtained based on the following equation.

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}), \tag{8}$$

where $p(\boldsymbol{\lambda})$ is the prior distribution of $\boldsymbol{\lambda}$ and $p(\boldsymbol{y}|\boldsymbol{\lambda})$ is the likelihood distribution as in (9).

$$p(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\lambda}). \tag{9}$$

As in (9), the likelihood function of the Poisson mixture regression model in (2) with K-components can be written as equation (10).

$$l_{mix} = \prod_{i=1}^{n} w_k p_k(y_{i_k}|\boldsymbol{x}_{i_k}, \boldsymbol{\beta}_k), \tag{10}$$

provided that $p_k(y_{i_k}|x_{i_k}, \boldsymbol{\beta}_k)$ as stated in (3), $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K$, $i_k = 1, 2, \ldots, n_K$, $n_1 + n_2 + \cdots + n_K = n$ and n_k is the number of observation in k-th sub-population, and $k = 1, 2, \ldots, K$ is the number of components. For 2-component of the spatial Poisson Mixture regression, the posterior distribution can be written as (11).

$$\begin{aligned}
p(\Phi|Y_i) &\propto p(y_i|\boldsymbol{w}, \boldsymbol{\beta}) \times p(\boldsymbol{w}) \times p(\boldsymbol{\beta}), \\
&= \prod_{i=1}^{n} \sum_{k=1}^{2} w_k p_k(y_{i_k}|x_{i_k}(u_{i_k}), \beta_k) \times p(w) \times p(\beta), \\
&= \prod_{i_k}^{n} \Big\{ w_1 \big[\log \big(\exp[\beta_{01} + \beta_{11} x_{1i_1}(u_{i_1}) + \cdots + \beta_{Q1} x_{Qi_1}(u_{i_1})] \big) \big] \\
&\quad + w_2 \big[\log \big(\exp[\beta_{02} + \beta_{12} x_{1i_2}(u_{i_2}) + \cdots + \beta_{Q2} x_{Qi_2}(u_{i_2})] \big) \big] \Big\} \times p(w) \times p(\beta),
\end{aligned} \tag{11}$$

where $p(\beta_{qk})$, $q = 0, 1, 2, \ldots, Q$ and $k = 1, 2$, is an individually independent of the Normal probability density function as the prior distribution for each regression parameter and $p(w_k)$ is an individually independent of the Dirichlet probability density function as the prior distribution of the k-th component of mixture proportion.

2.4 Gibbs Sampling

The Gibbs Sampling algorithm was firstly developed by Geman and Geman [21, 22]. The modified procedures performed on Gibbs Sampling to determine the membership of the k-th components mixture of Poisson could be written in Algorithm 1 [23].

Algorithm 1. Gibbs sampling algorithm for estimating the k-th components of mixture Poisson

1. Set $t = 0$ and give the state value of parameters of a mixture of Poisson for Markov chain $(\lambda_1^0, \ldots, \lambda_K^0, \boldsymbol{w}^0, \boldsymbol{H}_1^0, \ldots, \boldsymbol{H}_n^0)$. Usually, for the starting point, this value is randomly selected from the joint prior distribution.

2. For $t = 1, 2, \ldots, M$, update each parameter of mixture Poisson in every t-th iteration as follows:

   (a) Update Gibbs for λ_k^t, $k = 1, 2, \ldots, K$: sample λ_k^t from Gamma prior distribution, $Gamma(\alpha_k + \sum_{i=1}^n y_i H_{i_k}^{t-1}, \sum_{i=1}^n H_{i_k}^{t-1} + \theta_k)$, by using the updated value of $H_{i_1}^{t-1}, H_{i_2}^{t-1}, \ldots, H_{i_K}^{t-1}$. The prior distribution of Gamma is used as a conjugate prior for the parameter of the Poisson distribution λ_k^t.

   (b) Update Gibbs for proportion w_k^t: sample w_k^t from a prior distribution $Dir[(1 + \sum_{i=1}^n H_{i_1}^{t-1}), \ldots, (1 + \sum_{i=1}^n H_{i_K}^{t-1})]$ with the updated value of $H_{i_1}^{t-1}, H_{i_2}^{t-1}, \ldots, H_{i_K}^{t-1}$. The Dirichlet distribution is a conjugate prior for the multinomial distribution. $\boldsymbol{w}^t$ is a vector of multinomial parameters.

   (c) Update Gibbs for indicator $H_{i_k}^t$: sample $H_{i_k}^t$ from the prior distribution $Mult(1, r_{i_1}^t, \ldots, r_{i_K}^t)$, $i = 1, \ldots, n$, $k = 1, \ldots, K$, and $\sum_{k=1}^K r_{i_k}^t = 1$, where

   $$r_{i_k}^t = \frac{p(y_i|\lambda_k)p(w_k)}{p(y_i)}, fork = 1, \ldots, K, \tag{12}$$

   by using the updated value of λ_k^t and w_k^t.

   (d) Update a new Markov Chain state $(\lambda_1^t, \ldots, \lambda_K^t, \boldsymbol{w}^t, \boldsymbol{H}_1^t, \ldots, \boldsymbol{H}_n^t)$.

3. Increasing t, by setting $t = t + 1$

4. Repeat Steps 2-3 until the chain converges.

Stephens [10] in Iriawan [24] describes the K-components mixture procedures performed on Gibbs sampling for estimating the regression model with $(Q + 1)$ parameters, containing Q covariates and an intercept, by implementing their full conditional posterior distributions in each Gibbs iteration. This modified procedure is performed to estimate the parameters of Mixture Poisson Regression on Equation (11). It could be written in Algorithm 2.

Algorithm 2. Gibbs sampling algorithm for estimating the parameter of Mixture Poisson Regression

1. Setting the initial value of parameter Mixture Poisson Regression: $\boldsymbol{\Phi}^{(t)} = (\boldsymbol{w}, \boldsymbol{\beta})^{(t)}$, for $\boldsymbol{w} = (w_1, w_2, \ldots, w_K)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \ldots, \beta_{Q1})$, $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22}, \ldots, \beta_{Q2})$, $\ldots$, $\boldsymbol{\beta}_Q = (\beta_{0K}, \beta_{1K}, \ldots, \beta_{QK})$ on the iteration $t = 0$.

2. Generating component parameters for each mixture

   - Generating $\boldsymbol{w}^{(t+1)}$ from the full conditional posterior distributions of
   $$p\left(\boldsymbol{w}|y, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}\right)$$
   - Generating $\boldsymbol{\beta}_1^{(t+1)}$ from the full conditional posterior distributions of
   $$p\left(\boldsymbol{\beta}_1|y, \boldsymbol{w}^{(t+1)}, \boldsymbol{\beta}_2^{(t)}, \boldsymbol{\beta}_3^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}\right)$$
   - Generating $\boldsymbol{\beta}_2^{(t+1)}$ from the full conditional posterior distributions of
   $$p\left(\boldsymbol{\beta}_2|y, \boldsymbol{w}^{(t+1)}, \boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_3^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}\right)$$
   $$\vdots$$
   - Generating $\boldsymbol{\beta}_k^{(t+1)}$ from the full conditional posterior distributions of
   $$p\left(\boldsymbol{\beta}_k|y, \boldsymbol{w}^{(t+1)}, \boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)}, \ldots, \boldsymbol{\beta}_{k-1}^{(t+1)}, \boldsymbol{\beta}_{k+1}^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}\right)$$
   $$\vdots$$
   - Generating $\boldsymbol{\beta}_K^{(t+1)}$ from the full conditional posterior distributions of
   $$p\left(\boldsymbol{\beta}_K|y, \boldsymbol{w}^{(t+1)}, \boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)}, \ldots, \boldsymbol{\beta}_{K-1}^{(t+1)}\right)$$

3. Increasing t, by setting $t = t + 1$

4. Repeat step 2 and step 3 up to M times, where $M \to \infty$ or all parameters have reached their convergences.

For Poisson Mixture regression, we use the Normal distribution as an individually independent of conjugate prior for each component of parameter $\boldsymbol{\beta}$ and the Dirichlet distribution as an individually independent of prior for each component of parameter proportion $\boldsymbol{w}$. The hyper-prior parameter of $\boldsymbol{\beta}$ and $\boldsymbol{w}$ are set from their pseudo-prior after the data are modeled using the frequentist GLMs.

3 Research Variable

Research variables used in the analysis are the pattern of PHC location in the city of Surabaya along with covariate variables that potentially effect on the distribution of PHC. The covariate variables used are characteristics of each sub-district collected from Statistics Indonesia [15] and the Ministry of Health of the Republic of Indonesia [25]. The Response variable used for modeling is the number of Puskesmas and Klinik Pratama (PHC) observed at the tessellation object formed based on each sub-district boundary. There are 63 Puskesmas and 148 Klinik Pratama in Surabaya. The number of location points, Puskesmas and Klinik Pratama falling

on the object tessellations is observed as a random variable of a Poisson process. Each location point of Puskesmas and Klinik Pratama in two-dimensional space is expressed in latitude and longitude lines.

Table 1: Location of Puskesmas

Number	Name	Latitude	Longitude
1	Gayungan	-7.338074	112.718704
2	Kedurus	-7.319671	112.709634
3	Gununganyar	-7.340858	112.783992
⋮	⋮	⋮	⋮
62	Tambak Wedi	-7.217434	112.771583
63	Sememi	-7.248419	112.635390

Table 2: Location of Klinik Pratama

Number	Name	Address
1	Klinik Nurani Jaya 83 (JST)	K.H. Abdul Karim No 17 SBY
2	BP Widya Mandiri II (108)	Gubeng Kertajaya V C No 24
3	BP Klinik Kebangkitan (JST)	Manukan Madya 141 Tandes SBY
⋮	⋮	⋮
142	Klinik Rahap Bersalin Al-Azhar	Jl. Dupak Bandarejo No.23
143	Putri Rahayu	Jl. Mastrip IX No.9 Karang Pilang

Table 1 shows some of the data on the location point of Puskesmas obtained from the Publication of Surabaya Health Office, while Table 2 shows some of the data on the location of Klinik Pratama [16].

The covariate variables used in this research are some characteristics attributed to each sub-district. Description of each research variable is explained in Table 3.

4 Result and Discussion

4.1 PHC Data and Covariate Variables Exploration

PHC are the first-degree healthcare facilities including Puskesmas and Klinik Pratama. Health services provided by PHC include a general practitioner, maternal and child health services, dental services, etc. Due to their function as first-degree healthcare, PHC should ideally be built in every subdistrict. If a patient needs follow-up treatment, the doctors at PHC should refer the patient to the bigger hospitals.

The distribution of PHC location in Surabaya can be assumed as a point pattern. The analysis of this point pattern can be useful to know how the PHC locations spread in Surabaya. Since the PHC distribution is not evenly distributed in all sub-districts, the spatial point pattern

Table 3: Research Variables

Variable	Descriptions	Data Type
Number of PHC (Y)	Number of PHC located on each tessellation object	Count
X_1 (Population in each sub-district)	Total population in each sub-district in the city of Surabaya	Ratio
X_2 (Number of clean households)	Clean Households means the households that can maintain, improve, and protect the health of every household member from diseases and the environmental condition that does not support healthy living.	Ratio
X_3 (Accessibility Index in each subdistrict)	Accessibility Index is measured by using the travel time parameter as an indirect measurement system. The data is recorded based on the Google Maps software, and the method of measuring accessibility and depiction of the road network is based on the sub-district area, consisting of 31 sub-districts. The sub-district office is set as the center of sub-district and the road linking it is set as the route.	Ratio
X_4 (Length of road in good condition in km)	National/provincial/city roads that pass through the subdistrict are in good condition	Ratio

analysis focusing on the intensity pattern of the point in the area has been determined. The intensity shows the ratio of the number of points in an area to the population size. In an analysis involving administrative areas with a certain regional boundary, object tessellation is often used. The main reason for using object tessellation is the mechanism of a computational view that can change an area into a grid that has the same size as the region. The observed variables, therefore, can describe the conditions of each region. The main benefit of the analysis is that the results provide information on the district that need more PHC so that the government as a policymaker have sufficient information to take action in determining locations of new PHC in the future.

The SPPt analysis on PHC in Surabaya begins with transforming PHC data into the form of point patterns based on its location. The study area is divided into 31 tessellation objects corresponding to the number of sub-districts in Surabaya. Figure 2 shows the distribution of PHC in Surabaya. The red point displays the number of PHC in each location. The bigger the red circle point, the higher the number of PHC, see Figure 2(a). The location of PHC in Surabaya is still widely spread in Central Surabaya. It means that sub-districts in central Surabaya tend to have more PHC than do the other areas. Although the number of PHCs has met the minimum number of health facilities in Surabaya, the PHC locations tend to develop toward the proximity of the Central Surabaya area. This causes greater distance for people

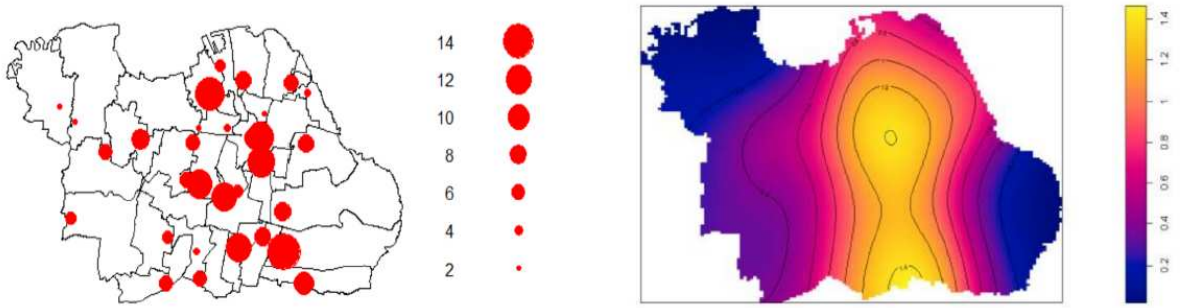


Figure 2: PHC Location Distribution in Surabaya: (a) Redpoint Displays the Number of PHC; (b) Coloring Shows the Smoothly Increasing the Intensity of PHC

living on the outskirts of Surabaya to reach the PHC. PHC intensity that shows PHC ratio per area in Figure 2(b) makes it clear that PHC development tends to move toward the center of Surabaya City.

Variables that represent each sub-district illustrate the social demographic, population, and accessibility conditions of the sub-district that describes in Figure 3. The total population variable indicates the number of people who have to get health services. Hence, the more densely populated an area, the greater the availability of health facilities. Figure 3(a) illustrates the condition of population distribution in the city of Surabaya. Most of Surabaya dwellers live in the area close to the Central Surabaya because the color is red and brick red which shows a higher population. Figure 3(c) shows the distribution of clean households. Clean households reflect the level of public awareness of health. The higher the level of public health awareness, the more the people carry out regular health checks so that the number of health facilities is increased. The distribution of clean households is relatively more evenly distributed in Surabaya.

The level of access in the sub-district is indicated by the accessibility index which is described by the linkage rate. The linkage rate means the distance that must be taken from one sub-district to another sub-district measured by the travel time [26]. High the linkage rate means the lower the level of access. High sub-district accessibility will encourage the growth of public facilities because the sub-districts are easier to reach. Figure 3(b) shows that the Accessibility Index in the centers of Surabaya area is lower than the other areas. This means that the centers of Surabaya have a higher level of access to health facilities. Figure 3(d) shows the infrastructure of roads in good conditions which reflects the results of economic development of a region, i.e. the higher availability of roads in good conditions will tend to increase the construction or improvement of other public facilities.

4.2 PHC Intensity Modelling Using Poisson Regression and Mixture Poisson Regression

Poisson regression modeling is employed to predict the count response data. Modeling was carried out on the intensity of PHC in Surabaya with four covariate variables. To do so, the assumptions that must be met in Poisson regression modeling are that the mean and variance data must be the same. Poisson regression modeled by the link function in equation (1).

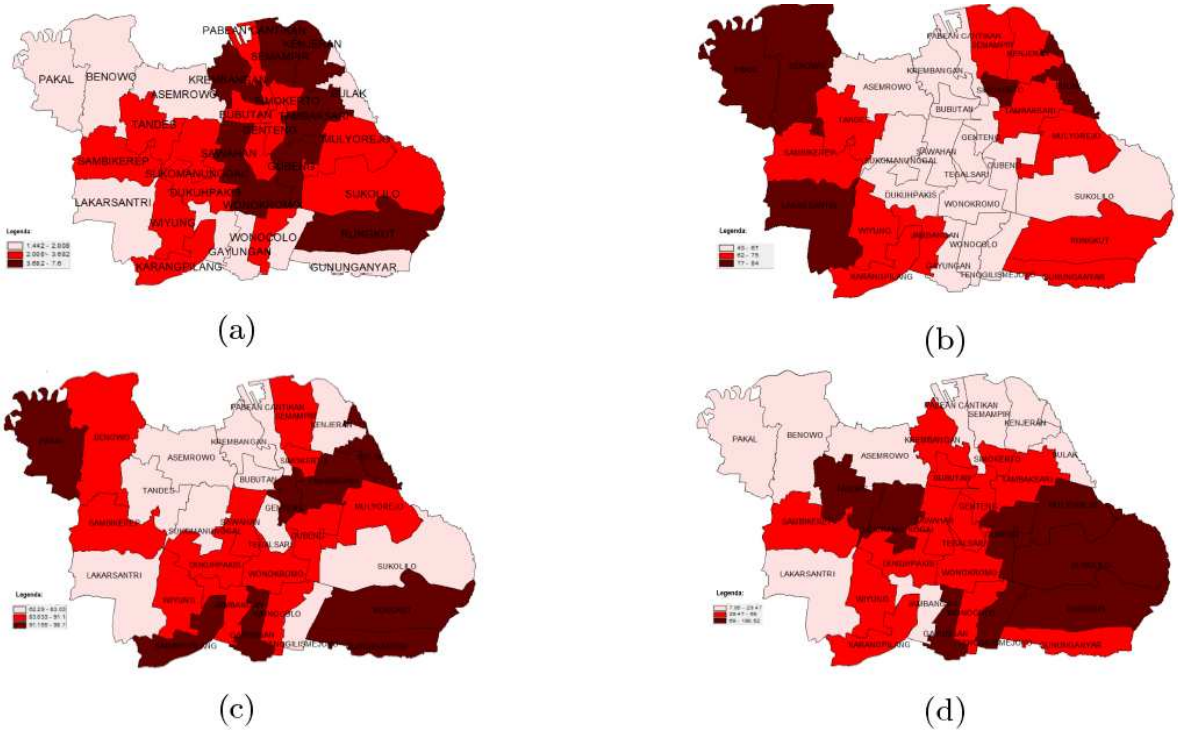


Figure 3: Covariate Variables: (a) Population in Each Sub-district; (b) Accessibility Index in Each Sub-district; (c) Number of Clean Households, and; (d) Length of Road in Good Condition in km

By using the Bayesian approach with MCMC as stated in Algorithm 2 is used to obtain the convergent results. The regression equation for PHC intensity is as follows:

$$\lambda_i = \exp\left[1.823 + 0.1926X_1(u_i) + 0.1787X_2(u_i) - 0.2361X_3(u_i) + 0.1478X_4(u_i)\right]. \tag{13}$$

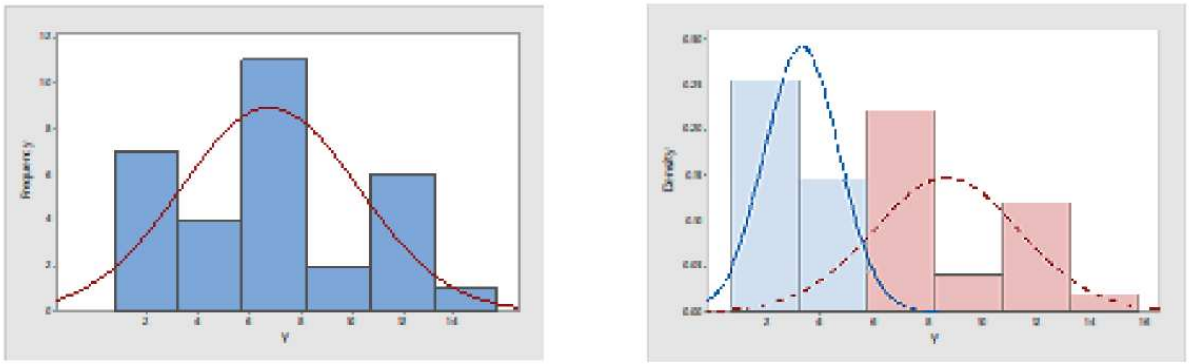
This model shows that PHC are established in areas with a dense population, a higher number of clean households, and higher accessibility. The assumption of no overdispersion is not fulfilled in this model.

A histogram of PHC number in Surabaya in Figure 4 shows the presence of a mixture distribution. The goodness of fit distribution by Chi-square test concludes that the response variable is not an unimodal Poisson distribution with the p-value of 0.006242096. Homogeneity testing with the quadrate count test resulted in Chi-square statistics of 114.67. It seems that the PHC intensity in each sub-district in Surabaya City follows the NHPP process. Overdispersion testing produces D-test statistics:

$$\text{D-test} = \frac{\text{Observed} \quad \text{variance}}{\text{Theoretical} \quad \text{variance}} \times (\text{number} \quad \text{observations-1}) = 53.01422, \tag{14}$$

which implies that the data of PHC are the overdispersed count data. Regression Poisson mixture must be employed to model the relationship between count-response variables and several covariate variables in which the response variables are from NHPP and or overdispersed count data.

The number of mixture components could be obtained visually based on the histogram and are as many as two components. These two components can be interpreted to imply



(a) (b)

Figure 4: PHC Histogram: (a) Histogram of the Number of PHC and; (b) Histogram of 2-component of the Number of PHC

that the sub-districts in Surabaya can be categorized in high and low intensities in terms of PHC availability. Membership of each component is determined by the Gibbs sampling procedure [21]. The result shows that the first component consists of seven sub-districts with many PHCs less than and equal to three in each sub-district, while the second component consists of 24 sub-districts with a total PHC of more than three in each sub-district. Figure 4(a) shows the histogram of PHC number, while Figure 4(b) shows the histogram of the number of PHC if there are two mixture components.

Mathematically, the Poisson mixture model for PHC in Surabaya can be written as follows:

$$Y_i \sim PM(w_k, \lambda_i),$$

$$\lambda_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}),$$

$$\begin{aligned} y_i &= w\big(\log\left[\lambda_1(u_{i_1})\right]\big) + (1-w)\big(\log\left[\lambda_2(u_{i_2})\right]\big), \\ &= w\big(\log\left[exp(\beta_{01} + \beta_{11}x_{1i_1}(u_{i_1}) + \beta_{21}x_{2i_1}(u_{i_1}) + \beta_{31}x_{3i_1}(u_{i_1}))\right]\big) \\ &\quad + (1-w)\big(\log\left[\beta_{02} + \beta_{12}x_{1i_2}(u_{i_2}) + \beta_{22}x_{2i_2}(u_{i_2}) + \beta_{32}x_{3i_2}(u_{i_2})\right]\big), \end{aligned}$$

$$\beta_{qk} \sim N\big(\mu_{[\beta_{qk}]}, \sigma^2_{[\beta_{qk}]}\big),$$

$\mu_{[\beta_{qk}]}$ and $\sigma^2_{[\beta_{qk}]}$ are hyper-parameters that are set as the pseudo-prior derived from regression parameter estimates using the frequentist GLMs approach, $q = 0, 1, 2, 3, 4$ and $k = 1, 2$. $w \sim Dir(1, 1)$ as an uninformative prior for the proportion of mixture component. Estimation of regression parameters for each component is carried out by using the fully computational Bayesian approach. The estimation process is done through repeated sampling through the form of a full conditional posterior distribution in equation (11). The structure of the regression Poisson mixture with four covariate variables model represent via DAG (directed acyclic graph) model in Figure 5 and then automatically generate the corresponding WinBUGS code from this graph. Determination of the posterior distribution and model parameter estimation are carried out using MCMC simulation until the irreducible, aperiodic and recurrent of chain conditions are obtained. These conditions are reached on the iteration of as much as 20000, with the sample thin one and the burn-in of 1000 iterations by using WinBUGS. The generated parameter β_{12}, as an example, are demonstrated as plots of the density plot, autocorrelation

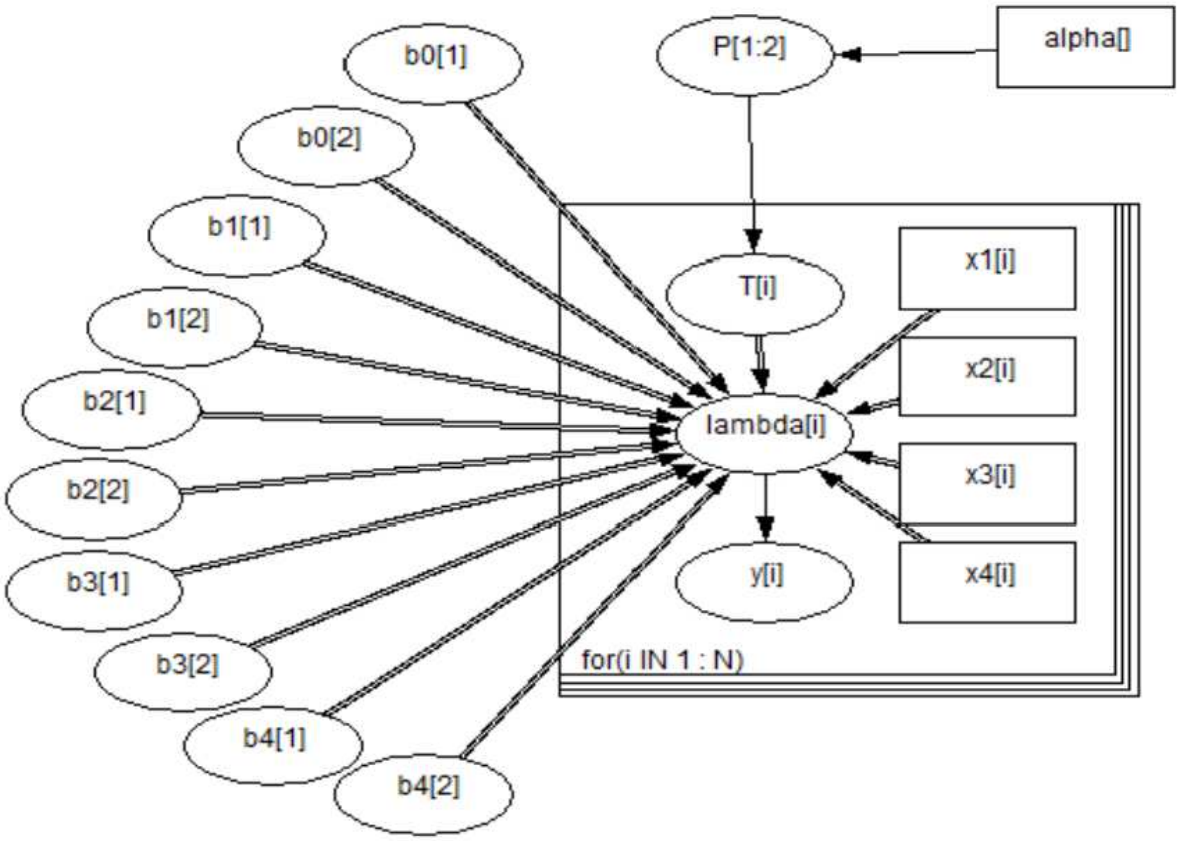


Figure 5: DAG Mixture Poisson Regression

plot, and history plot in Figure 6(a), Figure 6(b), and Figure 6(c) respectively. Figure 6(a) shows that diagnostic posterior density plots for Markov chain Monte Carlo (MCMC) process follow density plot of Normal distribution. The autocorrelation plot in Figure 6(b) indicates that there is no autocorrelation in samples generated by Gibb Sampling. Finally, the history plot in Figure 6(c) demonstrates that the chain is already in the state of a rapidly mixing processes which represents that the irreducible, aperiodic and recurrent conditions of the chain have already been reached perfectly. Typically, a parameter will appear to converge if the sample estimates form a tight horizontal band across this history plot.

Parameter significance testing is performed using a credible interval. If the credible interval does not hold zero then the parameter is declared significant. The estimated model of the Poisson mixture model in equation (7) for PHC (Y_i) in Surabaya is shown in equation (15). The number of PHC (Y_i) can be estimated by with the model in equation (15). As an example, the $x_{1i_1}(u_{i_1})$ represents that the total population in sub-district one has a contribution of 0.1874 in the first component of the mixture.

$$\begin{aligned}\hat{y}_i =& 0.2431 \times \left[1.3820 + 0.1874x_{1i_1}(u_{i_1}) + 0.2205x_{2i_1}(u_{i_1}) - 0.2351x_{3i_1}(u_{i_1})\right.\\ &\left. + 0.1716x_{4i_1}(u_{i_1})\right] + 0.7569 \times \left[1.9540 + 0.1732x_{1i_2}(u_{i_2}) + 0.1418x_{2i_2}(u_{i_2})\right.\\ &\left. - 0.1644x_{3i_2}(u_{i_2}) + 0.09193x_{4i_2}(u_{i_2})\right].\end{aligned} \tag{15}$$

The first component consisting of seven sub-districts with less than four PHCssub-district

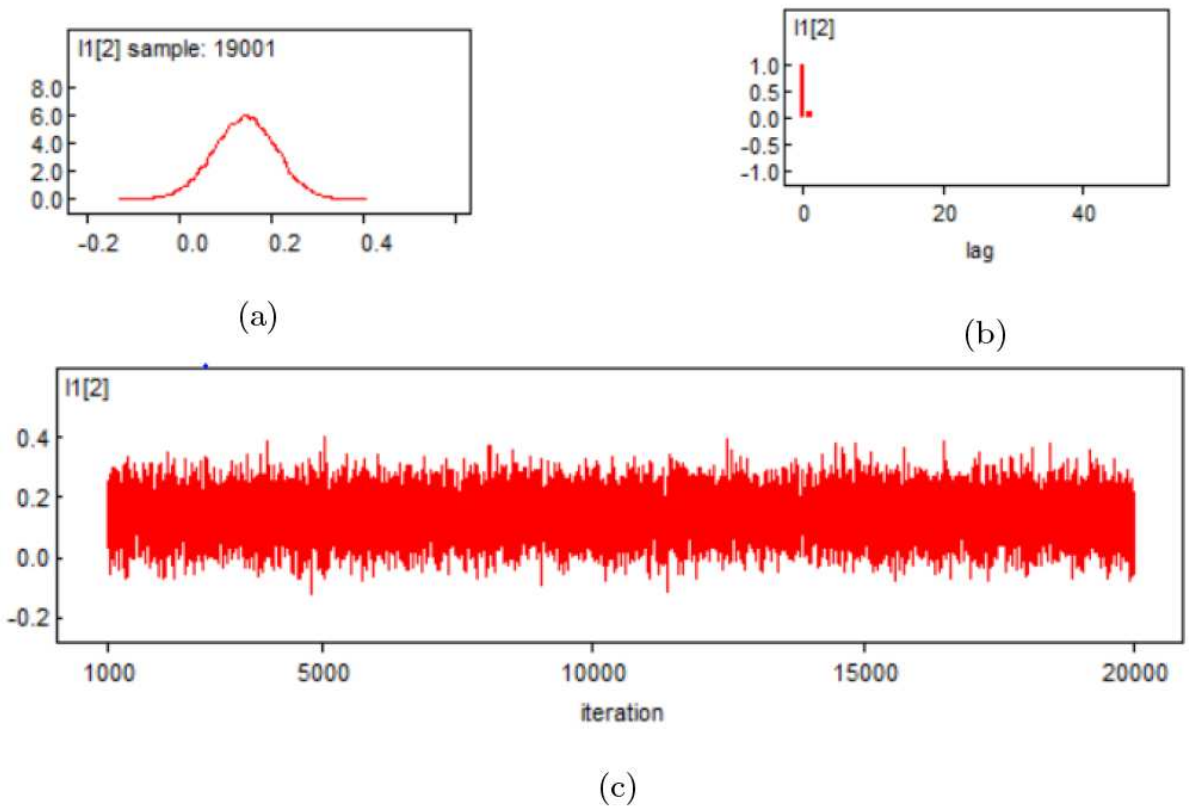


Figure 6: The Generated Posterior Sample of β_{12}: (a) The Density Plot; (b) The Autocorrelation Plot, and (c) The History Plot

of Simokerto, Asemrowo, Pakal, Benowo, Jambangan, Bubutan, and Bulakgives a smaller contribution to the model which is 24.3%, while the second component has a 75.7% contribution to the model. The covariate variables involved in modeling have a significant effect on the PHC distribution in the city of Surabaya. Variable of total population (X_1), number of clean households (X_2), and road length variable (X_4) have a significant positive effect, while Accessibility Index (X_3) has a negative influence on the PHC distribution. PHC development in Surabaya tends to be established in areas with a dense population, most clean households, and high accessibility.

Table 4 shows the estimated parameters of Poisson mixture regression in equation (7) which is shown in equation (15). It also shows the credible interval of each estimated parameters. The credible interval of all variables does not include zero value, which means that all variables are significant in the model.

4.3 Model Comparison

The model comparison between the Poisson Mixture regression model and Poisson regression is performed using the DIC goodness of fit test. The DIC of these two models is shown in Table 5 that concludes the second model is better than the first one due to its smallest DIC value.

Table 4: Estimation of Regression Parameter

Node	Mean	Sd	MC Error	2.5%	Median	97.5%
$P[1]$	0.243	0.0733	5.513×10^{-4}	0.116	0.2378	0.399
$P[2]$	0.757	0.0733	5.513×10^{-4}	0.601	0.7622	0.884
$b_0[1]$	1.382	0.2107	1.523×10^{-3}	0.951	1.3880	1.782
$b_0[2]$	1.954	0.0754	5.360×10^{-4}	1.805	1.9560	2.101
$b_1[1]$	0.187	0.0444	3.158×10^{-4}	0.101	0.1877	0.275
$b_1[2]$	0.173	0.0371	2.903×10^{-4}	0.099	0.1732	0.246
$b_2[1]$	0.221	0.0436	2.900×10^{-4}	0.134	0.2209	0.306
$b_2[2]$	0.142	0.0399	2.937×10^{-4}	0.063	0.1420	0.219
$b_3[1]$	-0.235	0.0435	2.934×10^{-4}	-0.321	-0.2351	-0.151
$b_3[2]$	-0.164	0.0401	3.009×10^{-4}	-0.243	-0.1644	-0.085
$b_4[1]$	0.172	0.0448	3.171×10^{-4}	0.083	0.1718	0.261
$b_4[2]$	0.092	0.0319	2.713×10^{-4}	0.019	0.0921	0.164

Table 5: DIC of the Model

Model	DIC Value
Poisson Regression	141.2410
Mixture Poisson Regression	137.3130

5 Conclusion

The analysis and discussion above have clearly demonstrated that the PHC distribution in the city of Surabaya is classified as the NHPP. This means that the PHC intensity can be modeled using Poisson Mixture regression. Demographic conditions represented by the population and the number of clean households, as well as the accessibility of the sub-districts illustrated by the number of linkages and good road conditions, affect the distribution of PHC in both of mixture components. The Poisson Mixture regression can be applied to model the overdispersed response variables.

6 Acknowledgment

The first author thanks Lembaga Pengelola Dana Pendidikan (LPDP) Ministry of Finance, Indonesia, for funding her Master study. Also, our thanks to the referees for their helpful comments.

References

[1] Waagepetersen, R. and Guan, Y. Two-step estimation for inhomogeneous spatial point processes. *Denmark: Journal of the Royal Statistical Society. Series B (Statistical*

Methodology). 2009. 71(3): 685–702.

[2] Chen, Y. and Ge, Y. Spatial point pattern analysis on the villages in China's poverty-stricken areas. *Procedia Environmental Sciences 27*. 2015. 98–105.

[3] Shirota, S., Mateu, J., and Gelfand, A.E. *Statistical Analysis of Origin-Destination Point Patterns: Modeling Car Thefts and Recoveries*. USA: arXiv preprint arXiv:1701.05863. 2017.

[4] Paci, L., Beamonte, M.A., Gelfand, A.E., Gargallo, P., and Salvador, M. Analysis of residential property sales using space-time point patterns. *Spatial Statistics*. 2017. 21: 149–165.

[5] Baddeley, A., Rubak, E., and Turner, R. *Spatial Point Patterns Methodology and Applications with R*. New York: CRC Press Taylor dan Francis Group. 2015.

[6] Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Pattern*. UK: John Wiley & Sons, Ltd. 2008.

[7] Carlin, B. P., and Louis, T. A. *Bayesian Methods for Data Analysis 3th Edition*. USA: CRC Press Taylor dan Francis Group. 2008.

[8] Grun, B. and Leisch, F. Finite mixtures of generalized linear regression models. In *Recent Advances in Linear Models and Related Areas, Shalabh and C. Heumann, Eds.*. Heidelberg. Springer. 2008. 205–230.

[9] Iriawan, N., Fithriasari, K., Ulama, B. S. S., Suryaningtias, W., Susanto, I. and Pravitasari, A. P. Bayesian Bernoulli Mixture Regression Model for Bidikmisi Scholarship Classification. In *Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information)*. 11/2 (2018), 67-76 DOI: http://dx.doi.org/10.21609/jiki.v11i2.536.

[10] Stephens, M. *Bayesian Methods for Mixtures of Normal Distributions*. Ph.D. Thesis. University of Oxford. 1997.

[11] Kusumaningrum, C.M., Iriawan, N., and Winahju, W.S. Pattern Analysis of Community Health Center Location in Surabaya using Spatial Poisson Point Process. In *AIP Conference Proceedings*. 2017. 1905-1.

[12] Wang, K., Yau, K. K. W., Lee, A. H. and Mc Lachlan, G. J. Two-component Poisson mixture regression modelling of count data with bivariate random effects. *Mathematical and Computer Modelling*. December 2007. 46(1112): 1468-1476.

[13] Wang P., Puterman M., Cokburn I. and Le N. Mixed Poisson regression models with covariate dependent rates. *Biometrics*. 52: 381-400. 1996.

[14] Nobles, M., Serban, N., dan Swann, J. Spatial Accessibility of Pediatric Primary Healthcare: Measurement and Inference. *The Annals of Applied Statistics*. 2014. 8(4): 1922–1946.

[15] Statistics Indonesia. *Kota Surabaya Dalam Angka 2016*. Surabaya. 2016.

[16] Surabaya Health Office. Puskesmas. Retrieved 3 17. 2018. Available: http:// dinkes.surabaya.go.id/portal/upt-dinas/puskesmas/. 2018.

[17] Mansour, S. Spatial analysis of public health facilities in Riyadh Governorate, Saudia Arabia: a GIS-based study to assess geographic variations of service provision and accessibility. *Geospatial Information Science*. 2016. 19(1): 26–38.

[18] Kwang-soo, L. and Kyeong-Jun, M. Hospital distribution in a metropolitan city: assessment by a geographical information system grid modeling approach. *Geospatial Health* 8(2). 2014. 537–544.

[19] Mufudza, C. and Erol, H. *Poisson Mixture Regression Models for Heart Disease Prediction*. Turkey: Computational and Mathematical Methods in Medicine. 2016.

[20] Ntzoufras, I. *Bayesian Modeling Using WinBUGS*. New Jersey: Wiley. 2009.

[21] Geman, S. and Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1984. 6(6): 721–741. doi:10.1109/TPAMI.1984.4767596.

[22] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. *Bayesian Data Analysis 3rd Edition*. Chapman & Hall Book, CRC Press, Taylor & Francis Group, Florida. 2014.

[23] Hamdah, D.F.A. *Bayesian Inference on Finite Mixtures of Poisson Distributions*. The Islamic University of Gaza: Ph.D. Thesis. 2015.

[24] Iriawan, N. Penaksiran Model Mixture Normal Univariabel: Suatu Pendekatan Mixture Bayesian dengan MCMC. In *Proceedings National Seminar and Conference VII Mathematics DIY & Jawa Tengah*. Yogyakarta. 2001.

[25] Ministry of Health of the Republic of Indonesia. *Profil Kesehatan Indonesia*. Jakarta. 2015.

[26] Mursalim. *Pengukuran Aksesibilitas Kecamatan di Wilayah Pemerintah Kota Surabaya*. Surabaya: Institut Teknologi Sepuluh Nopember. 2018.