mathematics

MDPI

Article

On Study of the Occurrence of Earth-Size Planets in Kepler Mission Using Spatial Poisson Model

Hong-Ding Yang [1,*], Yun-Huan Lee [2] and Che-Yang Lin [3]

[1] Institute of Statistics, National University of Kaohsiung, Kaohsiung 81148, Taiwan
[2] Department of Finance, Ming Chuan University, Taipei 11103, Taiwan
[3] Department of Business Administration, Yuanpei University of Medical Technology, Hsinchu 30015, Taiwan
* Correspondence: hongdingyang0111@gmail.com

Abstract: The problem of determining the occurrence rate for Earth-size planets orbiting Sun-like stars is emerging in the universe. We propose a methodology based on a spatial Poisson regression model with model parameters being inferred by the Bayesian framework to investigate this occurrence rate. We analyzed an exoplanet sample and its corresponding survey completeness data. Our results suggest that 46% of Sun-like stars have an Earth-size (i.e., 1–2 times Earth radii) planet with an orbital period of 5–100 days. Furthermore, we are also interested in the occurrence rate of Earth analogs hosted by GK dwarf stars (i.e., orbital period of 200–400 days and size 1–2 times Earth radii). After completeness correction, we obtained an occurrence rate of 0.18% based on the proposed methodology.

Keywords: conditional autoregressive model; Markov chain Monte Carlo; occurrence rate; spatial Poisson model

MSC: 62J05; 62J12; 62J20; 85-10; 85A35

Citation: Yang, H.-D.; Lee, Y.-H.; Lin, C.-Y. On Study of the Occurrence of Earth-Size Planets in Kepler Mission Using Spatial Poisson Model. *Mathematics* **2023**, *11*, 2508. https://doi.org/10.3390/math11112508

Academic Editors: Wen Zhang, Xiaofeng Xu, Jun Wu and Kaijian He

Received: 9 April 2023
Revised: 26 May 2023
Accepted: 29 May 2023
Published: 30 May 2023

1. Introduction

In spatial epidemiology, the spatial distribution of diseases is used to construct disease maps for finding the complex spatial patterns of interesting diseases. When Bayesian hierarchical models are used to investigate the disease mapping, various spatially structured random effects can be considered in models. Recently, we have been witnessing a resurgence of interest in disease mapping, and many efficient methods have been proposed in the literature (see Moraga and Lawson 2012 [1]; Duncan et al., 2017 [2]; Lawson 2018 [3]; Baer and Lawson 2019 [4]). To the best of our knowledge, the application of disease mapping concepts to explore related issues in astronomy within the context of spatial regression remains unaddressed. This knowledge gap is the driving force behind our investigation into whether spatial disease mapping techniques can be utilized to examine the occurrence of Earth-size planets in the Kepler survey. Disease mapping leverages neighboring region information for parameter estimation in epidemiology, leading to more accurate spatial predictions. In this study, we extended this approach by incorporating spatial random effects to capture the spatial correlation in the data. Interestingly, the incorporation of neighboring region information is still relatively unexplored in astronomy (e.g., Petigura et al., 2018 [5]).

To the best of our knowledge, however, how to apply the concepts of disease mapping to discuss the related issues in astronomy has not been adequately addressed under the spatial regression settings. This motivates us to explore whether the techniques of spatial disease mapping can be applied to investigate the occurrence of Earth-size planets in the Kepler survey.

The Kepler mission aims to explore the diversity of planets and planetary systems. The discovery of thousands of transiting planets and planet candidates by the Kepler mission drastically broadens our knowledge of exoplanets, especially in the category of

close-in ($\lesssim$1 AU) and small ($\lesssim$4 earth radii) planets around main-sequence dwarf stars (see Batalha 2014 [6]; Burke et al., 2014 [7]; Mullally et al., 2015 [8]). The inference of the occurrence of Earth-size planets is an interesting problem that has attracted the attention of astronomers because of the important theories regarding planet formation and evolution models (see Benz et al., 2014 [9]). Owing to the low false positive rate of the survey (see Fressin et al., 2013 [10]; Lissauer et al., 2014 [11]) while seeing different results from Santerne et al. (2016) [12] for giant-planet candidates, numerous works offered a window into the statistical studies of planet occurrence rates in terms of orbital periods and planet radius (see Dong and Zhu 2013 [13]; Fressin et al., 2013 [10]; Petigura et al., 2013 [5]; Burke 2015 [14]; Dressing and Charbonneau 2015 [15]; Silburt et al., 2015 [16]; Morton et al., 2016 [17]).

In this paper, we took the exoplanet sample and its corresponding survey completeness from Petigura et al., 2013 [5]. In the proposed methodology, we defined the planet occurrence to be based on the detection of a planet within a specified range of orbital period and orbital radius. To consider the spatial dependences of the data, we applied a spatial Poisson regression model (e.g., Besag et al., 1991 [18]; Chen and Yang 2011 [19]; Cressie 2015 [20]) to model the detection probability of an exoplanet. Further, to infer the posterior probability of detecting an exoplanet, a stochastic algorithm based on Markov chain Monte Carlo (MCMC) under the Bayesian framework was designed. Finally, the posterior inferences can simultaneously describe the number of exoplanets and the corresponding occurrence rate in the study region.

The remainder of this paper is organized as follows. In Section 2, we introduce a joint modeling methodology and present how to estimate parameters in the proposed model. Section 3 applies the proposed model to determine the occurrence rate of the Kepler planet. We conclude the paper with a discussion in Section 4.

2. Methodology

Let D be a bounded continuous random field in the $\Re^2$, which is partitioned into $n = n_1 \times n_2$ regular grids $D_1, \ldots, D_n$ with $D = \bigcup_{i=1}^{n} D_i$ and $D_i \bigcap D_j = \emptyset$ for $i \neq j$. Let Y_i, $i = 1, \ldots, n$, be a random variable that counts the number of exoplanets in grid D_i. For grid D_i, the expected number, E, of exoplanets can be easily evaluated by:

$$E = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Motivated by the concept of a standardized mortality ratio in epidemiology (see Kelsall and Wakefield 2002 [21]; Lawson 2018 [3]), a standardized occurrence ratio of exoplanets for the grid D_i is defined by

$$r_i = \frac{Y_i}{E}, \; i = 1, \ldots, n.$$

In general, one can simply use r_i as the occurrence rate of exoplanets in grid D_i. Here, one potential influential factor is that a large amount of gravity generally exists among planets, and the correlation of the data set among grids should be considered in estimating such an occurrence rate. Obviously, the quantity r_i does not take into account the dependence among $\{Y_1, \ldots, Y_n\}$. Thus, using r_i to estimate the occurrence rate of exoplanets of the grid D_i may yield inaccurate results. Motivated from existing works (see Kelsall and Wakefield 2002 [21]; Chen and Yang 2011 [19]; Moraga and Lawson 2012 [1]; Lawson 2018 [3]; Baer and Lawson 2019 [4]), a spatial conditional autoregressive (CAR) model (see Moraga and Lawson 2012 [1]; Cressie 2015 [20]; Lawson 2018 [3]; Baer and Lawson 2019 [4]) was applied, which was used to describe possible spatial correlations among $\{Y_1, \ldots, Y_n\}$. The estimates of the occurrence rate of exoplanets in the grid D_i, $i = 1, \cdots, n$, were then proposed.

2.1. Spatial Poisson Regression Model

For $i = 1, \ldots, n$, let R_i be the occurrence rate of exoplanets in grid D_i. Then, an intuitive model for Y_i given R_i; $i = 1, \ldots, n$, is a Poisson distribution as follows:

$$Y_i \mid R_i \sim \text{Poi}(R_i E). \tag{1}$$

In Equation (1), $R_i E$ represents the intensity rate of the Poisson process and $R_i > 0$ is the main parameter of interest in this research. In this paper, our goal was to incorporate the spatial dependence of $Y_1, \ldots, Y_n$ to estimate the unobserved variables $R_1, \ldots, R_n$. Suppose that there are p grid-level covariates observed in grid D_i denoted together with 1 for the intercept by $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$. As suggested in Basag et al. (1991) [18], the occurrence rate R_i of interest can be modeled in the following manner:

$$\ln(R_i) = \boldsymbol{x}_i' \boldsymbol{\beta} + \delta_i;\ i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients and δ_i is a spatial random error process. In spatial statistics, the spatial random errors $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ capture the spatial variation and can offer a local adjustment to the mean trend due to unobserved covariates. In general, we assume that $\boldsymbol{\delta}$ follows a multivariate Gaussian process as follows:

$$\boldsymbol{\delta} \mid \sigma^2, \phi \sim N\Big(\mathbf{0}, \sigma^2 \boldsymbol{V}(\phi)\Big), \tag{3}$$

where the $n \times n$ matrix $\boldsymbol{V}(\phi)$ is a spatial correlation matrix, ϕ is an unknown parameter, and σ^2 is a variance component. According to the CAR model, $\sigma^2 \boldsymbol{V}(\phi)$ given in Equation (3) can be further decomposed as

$$\sigma^2 \boldsymbol{V}(\phi) = (\boldsymbol{I} - \phi \boldsymbol{C})^{-1} \boldsymbol{M},$$

where $\boldsymbol{C} = (c_{ij})$ is an $n \times n$ spatial association matrix, $\boldsymbol{I}$ is an identity matrix, and $\boldsymbol{M} = \sigma^2 \boldsymbol{I}$. Under these settings, we have the following facts: (i) $(\boldsymbol{I} - \phi \boldsymbol{C})$ is nonsingular; (ii) when $\phi \in (\phi_{\min}, \phi_{\max})$, $(\boldsymbol{I} - \phi \boldsymbol{C})^{-1} \boldsymbol{M}$ is symmetric and positive-definite, where the upper and lower limits of ϕ are evaluated by the inverses of the smallest and the largest eigenvalues of the spatial association matrix. For the sake of simplicity, in this paper, we constructed $\boldsymbol{C}$ according to the rook contiguity structure; that is, the (i, j)th element of $\boldsymbol{C}$ is of the following form:

$$c_{ij} = \begin{cases} 1, & i \sim j; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Note that $i \sim j$ in Equation (4) represents that D_i and D_j are neighbors with a common boundary.

We define $N_i \equiv \{j \mid j \sim i\}$ to be the neighborhood set of grid D_i and $\boldsymbol{\delta}_{-i} \equiv (\delta_1, \ldots, \delta_{i-1}, \delta_{i+1}, \ldots, \delta_n)'$; then, the conditional distribution of δ_i conditioned on $\boldsymbol{\delta}_{-i}$ is given by

$$\delta_i \mid \sigma^2, \phi, \boldsymbol{\delta}_{-i} \sim N\left(\phi \sum_{j \in N_i} c_{ij} \delta_j, \sigma^2\right), \tag{5}$$

for $i = 1, \ldots, n$. Note that the joint distribution of $\delta_i \mid \sigma^2, \phi, \boldsymbol{\delta}_{-i}$, $i = 1, \ldots, n$ can be shown to be a multivariate Gaussian distribution as in Equation (3) based on the factorization theorem of Besag (1974) [22] and the properties of multivariate Gaussian distributions. Readers can better understand the correctness of Equation (5) by referring to De Oliveira (2012) [23] for a comprehensive and systematic introduction to the CAR model. It is obvious from Equation (5) that the spatial dependence is considered through the information derived from neighbors. Notice that the spatial Poisson regression model offers the advantage of

incorporating information from neighboring regions to enhance parameter estimation and prediction. Additionally, it is worth noting that the consideration of data correlation in recent literature is still relatively uncommon, as observed in studies such as Petigura et al. (2018) [24].

2.2. Prior Specifications and Posterior Distribution

Using the Bayesian approach, we set mutually independent prior distributions on parameters $\boldsymbol{\beta}$, σ^2, and ϕ as shown in Table 1. For $\boldsymbol{\beta}$ and σ^2, the hyper-parameters are pre-specified constants such that the corresponding priors are nearly flat. Based on the CAR model, the spatial dependence parameter ϕ must fall within (ϕ_{min}, ϕ_{max}) to ensure that $(\boldsymbol{I} - \phi \boldsymbol{C})^{-1}$ is a positive-definite matrix. However, ϕ_{min} can be less than zero, leading to a negative spatial correlation, which is rare in practice. Hence, we further restricted the spatial correlation parameter ϕ domain to $(0, \phi_{max})$, ensuring positive spatial correlation. This modification ensures that the model captures the desired spatial dependence structure and aligns with common practices in the field. According to the priors in Table 1, the joint prior distribution of $\boldsymbol{\beta}$, σ^2, and ϕ, denoted as $\pi(\boldsymbol{\beta}, \sigma^2, \phi)$, is given by

$$\pi(\boldsymbol{\beta}, \sigma^2, \phi) = \pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\phi) \propto \sigma^{-2(a+1)} \exp\left(-\frac{1}{b\sigma^2}\right); \ \sigma > 0, \ \phi \in (0, \phi_{max}). \tag{6}$$

Combining Equations (1)–(3) and Equation (6), the joint posterior distribution of σ^2, ϕ, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ conditioned on observed data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ satisfies:

$$\begin{aligned} p(\sigma^2, \phi, \boldsymbol{\beta}, \boldsymbol{\delta} \mid \boldsymbol{Y}) &= \frac{p(\sigma^2, \phi, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{Y})}{p(\boldsymbol{Y})} \\ &\propto \prod_{i=1}^{n} p(Y_i \mid R_i) p(\boldsymbol{\delta} \mid \sigma^2, \phi) \pi(\boldsymbol{\beta}, \sigma^2, \phi) \\ &\propto \exp\left(\sum_{i=1}^{n} Y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i) - E\sum_{i=1}^{n} \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i)\right) \\ &\quad \times \left(\det\left(\sigma^2 \boldsymbol{V}(\phi)\right)\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\delta}' \boldsymbol{V}(\phi)^{-1}\boldsymbol{\delta}\right) \\ &\quad \times \sigma^{-2(a+1)} \exp\left(-\frac{1}{b\sigma^2}\right). \end{aligned} \tag{7}$$

Because the joint posterior distribution in Equation (7) cannot be applied directly to generate posterior samples of model parameters, an alternative method called a Markov chain Monte Carlo (MCMC) method will be introduced in the following to generate posterior samples of model parameters.

Table 1. Priors for model parameters $\boldsymbol{\beta}$, σ^2, and ϕ.

Parameter	Prior Distribution	Support of Hyper-Parameter
$\boldsymbol{\beta}$	Non-informative prior	
σ^2	Inverse gamma (a, b)	$a, b > 0$
ϕ	Uniform $(0, \phi_{max})$	$\phi_{max} > 0$

2.3. Posterior Inferences of Model Parameters

To generate posterior samples of σ^2, ϕ, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$, the conditional posterior distributions of each parameter given all of the others are needed. One can then successively sample these conditional posterior distributions and obtain Markov chains in the parameter spaces that will converge to the joint posterior distribution of Equation (7) under Tierney's conditions (1994) [25].

Next, we summarize all necessary conditional posterior distributions for σ^2, ϕ, $\boldsymbol{\beta}$, and δ_i, $i = 1, \ldots, n$, based on Equations (1)–(7) as follows:

$$\begin{aligned}
p(\sigma^2 \mid \phi, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{Y}) &\propto p(\boldsymbol{\delta} \mid \sigma^2, \phi)\pi(\sigma^2) \\
&\propto \sigma^{-2(n/2+a+1)} \exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{b} + \frac{1}{2}\boldsymbol{\delta}'(\boldsymbol{I} - \phi\boldsymbol{C})\boldsymbol{\delta}\right)\right) \\
p(\phi \mid \sigma^2, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{Y}) &\propto p(\boldsymbol{\delta} \mid \sigma^2, \phi)\pi(\phi) \\
&\propto (\det(\boldsymbol{V}(\phi)))^{-1/2} \exp\left(\frac{\phi}{2\sigma^2}\boldsymbol{\delta}'\boldsymbol{C}\boldsymbol{\delta}\right) \\
p(\boldsymbol{\beta} \mid \sigma^2, \phi, \boldsymbol{\delta}, \boldsymbol{Y}) &\propto \prod_{i=1}^{n} p(Y_i \mid R_i)\pi(\boldsymbol{\beta}) \\
&\propto \exp\left(\sum_{i=1}^{n} Y_i \boldsymbol{x}_i'\boldsymbol{\beta} - E\sum_{i=1}^{n} \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i)\right) \\
p(\delta_i \mid \sigma^2, \phi, \boldsymbol{\beta}, \boldsymbol{\delta}_{-i}, \boldsymbol{Y}) &\propto p(Y_i \mid R_i)p(\delta_i \mid \sigma^2, \phi, \boldsymbol{\delta}_{-i}) \\
&\propto \exp\left(Y_i\delta_i - E\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i) - \frac{1}{2\sigma^2}\left(\delta_i^2 - 2\delta_i\phi\sum_{j\in N_i} c_{ij}\delta_j\right)\right)
\end{aligned}$$

We notice that $p(\sigma^2 \mid \phi, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{Y})$ is an inverse gamma distribution; that is, $\sigma^2 \mid \phi, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{Y} \sim IG(n/2 + a, (1/b + \boldsymbol{\delta}'(\boldsymbol{I} - \phi\boldsymbol{C})\boldsymbol{\delta}/2)^{-1})$. Therefore, a Gibbs sampling algorithm (see Geman and Geman 1984 [26]) can be used to generate the posterior samples of σ^2. However, $p(\phi \mid \sigma^2, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{Y})$, $p(\boldsymbol{\beta} \mid \sigma^2, \phi, \boldsymbol{\delta}, \boldsymbol{Y})$, and $p(\delta_i \mid \sigma^2, \phi, \boldsymbol{\beta}, \boldsymbol{\delta}_{-i}, \boldsymbol{Y})$, $i = 1, \ldots, n$, are not all standard distributions; hence, a Metropolis–Hastings algorithm (see Chib and Greenberg 1995 [27]) can be applied to ϕ, $\boldsymbol{\beta}$, and δ_i, respectively, to iteratively generate an ergodic Markov chain that yields the corresponding posterior samples. In particular, generating the posterior samples of ϕ is relatively difficult because ϕ appears in the covariance matrix $\boldsymbol{V}(\phi)$. In this paper, we treated ϕ as a discrete random variable that is defined on finite grid points from 0 to ϕ_{max}; hence, the values of matrix $\boldsymbol{V}(\phi)$ on these finite grid points can be computed in advance. For each step, the posterior sample of ϕ is generated from a probability mass function, which is based on the values of $(\det(\boldsymbol{V}(\phi)))^{-1/2} \exp\left(\frac{\phi}{2\sigma^2}\boldsymbol{\delta}'\boldsymbol{C}\boldsymbol{\delta}\right)$ evaluated on the finite grid points of $\phi \in (0, \phi_{max})$.

Based on the posterior samples of σ^2, ϕ, $\boldsymbol{\beta}$, and δ_i, $i = 1, \ldots, n$, the inferences of model parameters and the occurrence rate of exoplanets in grid D_i, $i = 1, \ldots, n$, can be obtained.

3. Application of the Proposed Methodology

To model the occurrence distribution of planets as a function of the planet period and radius, Petigura et al. (2013) [5] considered transiting planets that are all hosted by GK-type stars. They defined GK-type stars as those with surface temperatures of 4100 K $\leq T_{\text{eff}} \leq$ 6100 K and gravities of 4.0 cm/s^2 $\leq \log g \leq$ 4.9 cm/s^2. Furthermore, these planets are restricted to the brightest GK-type stars observed by *Kepler* (Kp = 10–15 mag). These 42,557 stars have the lowest photometric noise in the Kepler survey, thereby maximizing the detectability of Earth-size planets. In the present work, we mainly studied the occurrence rate of planets based on the catalog by Petigura et al. (2013) [5], which can compare our findings with their seminal work by adopting the same study region. Figure 1 shows the scatter plot of the data. Let x_1 be the orbital period (days), x_2 be the planet size (Earth radii), and $D = [6.25, 400] \times [1, 16]$ be the region of interest for this work; it is divided into the 6×4 grids shown in Figure 2. Let Y_i record the number of events in grid D_i for $i = 1, \ldots, 24$. Please note that the region D is the same as in Petigura et al. (2013) [5].

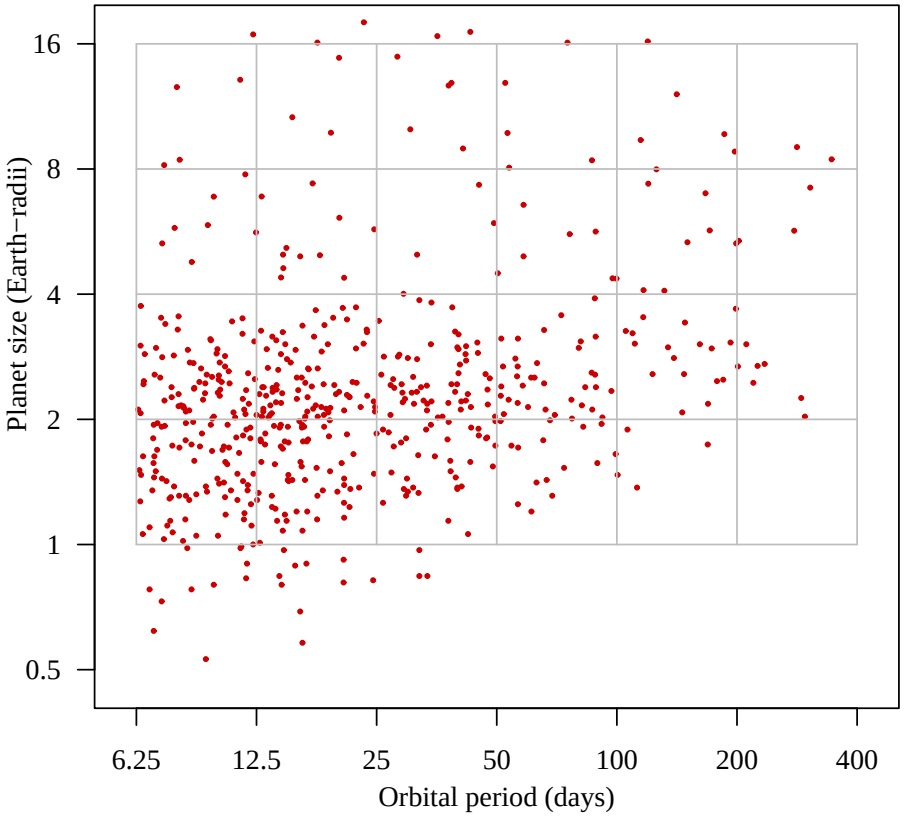

Figure 1. The scatterplot of exoplanets in the x_1–x_2 space and the 24 subregions D_i, $i = 1, \cdots, 24$.

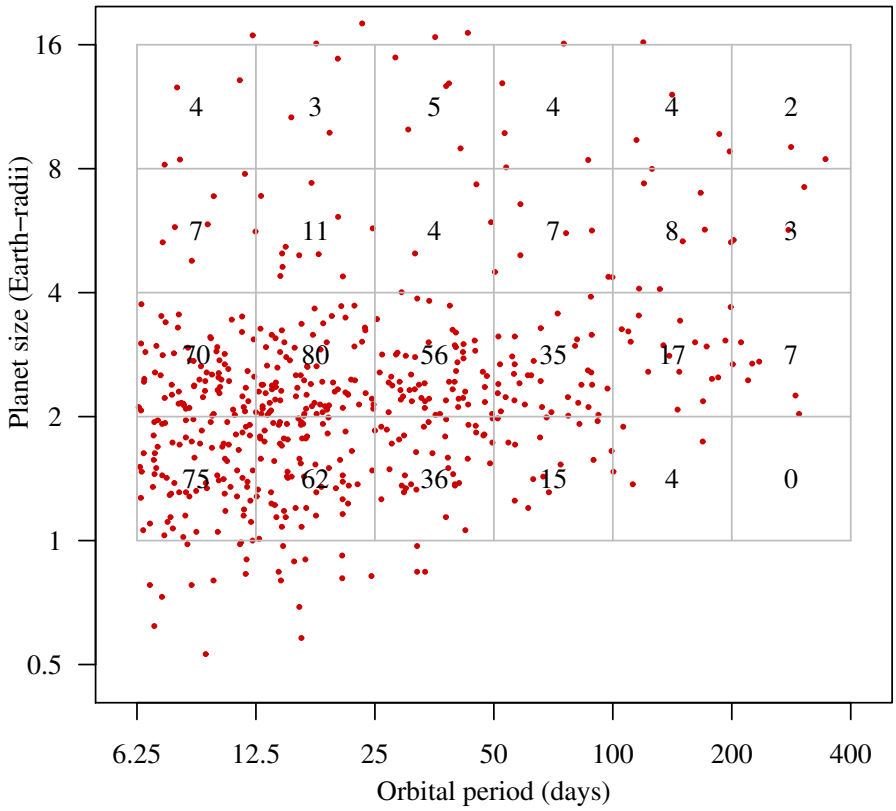

Figure 2. The values of Y_i for $i = 1, \cdots, 24$.

We applied the linear regression model illustrated in Equation (2) of Section 2.1 to model the occurrence rate R_i and considered two grid-level covariates, x_{i1} and x_{i2}, in the model, where x_{i1} and x_{i2} are, respectively, defined by the central points of the orbital period (days) and planet size (earth radii) of the grid D_i (i.e., the central coordinate of the grid D_i) for $i = 1, \ldots, 24$. As a result, the used model, called Model 1, is given by

$$\ln(R_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \delta_i;\ i = 1, \ldots, 24, \tag{8}$$

where β_0, β_1, and β_2 are unknown regression coefficients, and δ_i is a spatial random error process. Based on the Bayesian approach in Section 2.3, prior distributions of parameters in $\boldsymbol{\theta} = (\sigma^2, \phi, \beta_0, \beta_1, \beta_2)'$ are, respectively, set as follows:

$$\begin{aligned} \sigma^2 &\sim IG(3, 0.001) \\ \phi &\sim U(0, \phi_{\max}) \\ \beta_0 &\sim U(-20, 20) \\ \beta_1 &\sim U(-20, 20) \\ \beta_2 &\sim U(-20, 20) \end{aligned}$$

Note that $\phi_{\max}$ is 0.29 because the smallest eigenvalue of C is 3.42. Since we lacked additional information about the central tendencies of the parameters, we selected the hyper-parameter values for the prior distributions based on the preference for larger variances. Although larger variances may result in a slower convergence, the MCMC algorithm can still converge. Additionally, the larger variances allow for more flexibility and variation in the MCMC updates, enhancing the parameter space exploration.

Next, we first examined the hypothesized model (i.e., Equations (1)–(3)) that is suitable for analyzing the occurrence rate of Earth-size planets in the Kepler survey. In this paper, we conducted a simulation study based on the Pearson chi-squared test to illustrate the goodness of fit of the used model (i.e., Equation (8)); Model 1). In addition, as listed in the bottom of Table 2, a model (i.e., Model 2) with only the regressors and a model (i.e., Model 3) with only the spatial random error process were also used for comparison. Let $Y_i^{*(t)}; i = 1, \ldots, 24$, be independently generated from $\text{Poi}\left(R_i^{*(t)} E\right)$, with E being the expected number of exoplanets evaluated according to the observed data $\boldsymbol{Y}$, where $R_i^{*(t)}$ is an estimate of the occurrence rate R_i based on the posterior medians of $\boldsymbol{\theta}$ under the used model (i.e., Model 1, Model 2, or Model 3) and $t = 1, \ldots, 5$, represents the t-th simulation. For each simulation replicate, the goodness-of-fit test statistic is computed in the following manner:

$$\chi^{2(t)} \equiv \sum_{i=1}^{24} \frac{\left(Y_i^{*(t)} - R_i^{*(t)} E^{*(t)}\right)^2}{R_i^{*(t)} E^{*(t)}}$$

where $E^{*(t)}$ is the expected number of exoplanets evaluated based on the t-th simulated data $Y_i^{*(t)}; i = 1, \ldots, 24$. The simulation results are displayed in Table 2. First, we notice that Model 2 with only the regressors has a large $\chi^{2(t)}$ value for each simulation replicate. This indicates that Model 2 without considering the spatial correlation of the data is very inappropriate. Comparing the proposed model (i.e., Model 1) versus Model 3, they have relatively small $\chi^{2(t)}$ values and hence Model 1 and Model 3 are both appropriate for the analysis of the occurrence rate of Earth-size planets. Overall, the $\chi^{2(t)}$ values of Model 1 are slightly smaller than those of Model 3, which further suggests to us to use Model 1 (i.e., Equation (8)) to analyze the data set. Even if all the estimated regression coefficients are not significant (see Table 3), in general, the regressors should slightly contribute to evaluating the occurrence rate. Moreover, Figure 3 shows 95% credible intervals of Y_i; $i = 1, \ldots, 24$, for Model 1, Model 2, and Model 3. The results are in accord with Table 2; that is, Figure 3 reveals that Model 2 performs poorly and that Model 1 and Model 3

are fairly comparable. On the other hand, we notice that the data may contain potential biases that may arise from observational precision that results in inaccurate estimates of the underlying occurrence rates. In our proposed methodology, the random effects describe the spatial correlation in the data and are a suitable remedy for missing explanatory variables, addressing the limitations caused by uncollected vital variables. The simulation results indicate the effectiveness of our approach in mitigating potential biases and enhancing the model's explanatory power. Based on the results in Table 2 and Figure 3, Model 1 in Equation (8) is acceptable and hence we used it to analyze the occurrence rate of Earth-size planets in the next content.

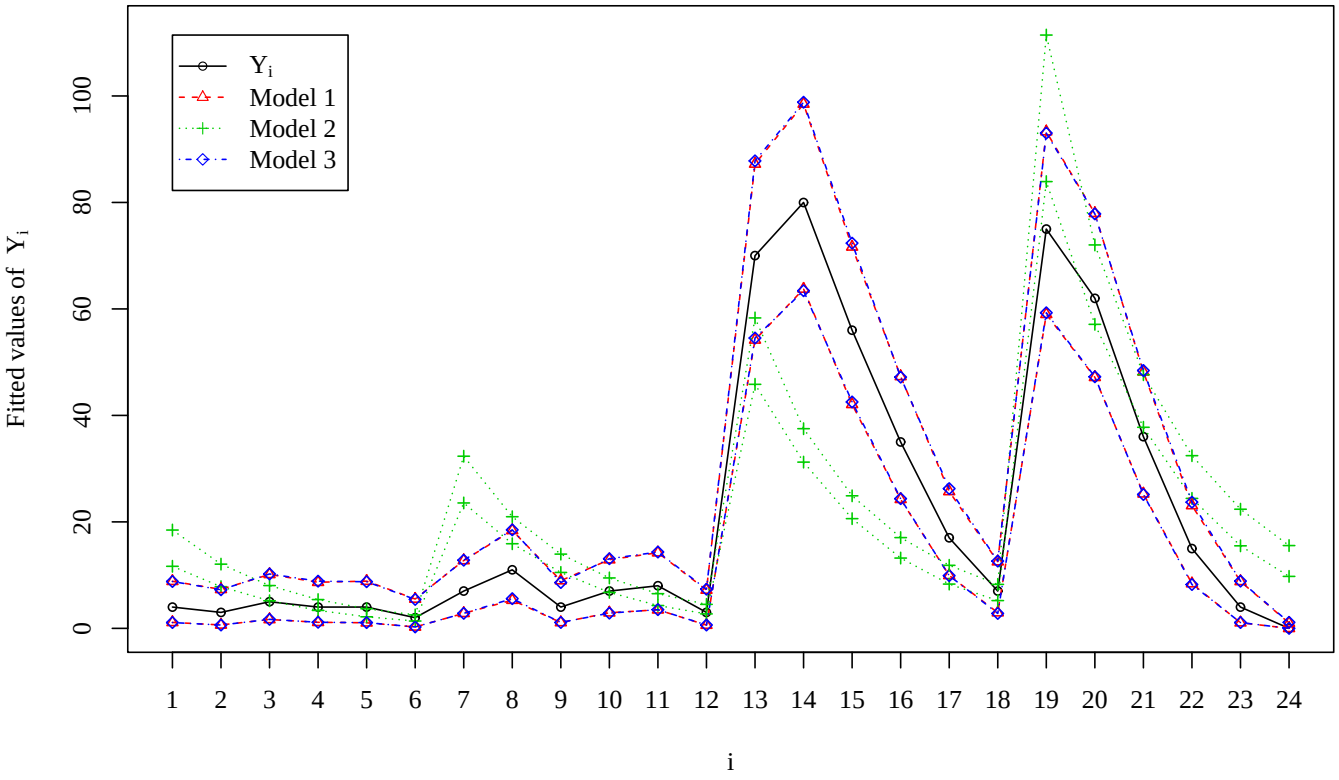

Figure 3. The 95% credible intervals for Models 1–3. Model 1: a model with the regressors and the spatial component; Model 2: a model with only the regressors; Model 3: a model with only the spatial component.

Table 2. The expected numbers and the values of chi-squared test statistics for Model 1, Model 2, and Model 3 based on the observed data $\boldsymbol{Y}$ and the simulated data $\boldsymbol{Y}^{*(t)} = \left(Y_1^{*(t)}, \ldots, Y_{24}^{*(t)}\right)'$, where t represents the t-th simulation with $t = 1, \ldots, 5$.

		$\boldsymbol{Y}^{*(t)}$					$\boldsymbol{Y}$
		1	**2**	**3**	**4**	**5**	
Model 1	$E^{*(t)}$	23.167	24.958	22.917	22.917	23.500	23.000
	$\chi^{2(t)}$	0.665	0.389	0.560	0.680	0.580	0.129
Model 2	$E^{*(t)}$	23.083	23.000	22.125	22.917	23.125	23.000
	$\chi^{2(t)}$	29.606	19.295	13.640	26.167	10.892	222.619
Model 3	$E^{*(t)}$	22.000	21.042	21.958	21.125	24.333	23.000
	$\chi^{2(t)}$	0.823	0.769	0.803	0.809	0.774	0.140

Note: Model 1: $\ln(R_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \delta_i$; Model 2: $\ln(R_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$; Model 3: $\ln(R_i) = \beta_0 + \delta_i$; $i = 1, \ldots, 24$.

Table 3. Summary of posterior inferences for model parameters.

Parameter	2.5%	5%	Median	95%	97.5%	Mean	S.D.
β_0	−6.052	−5.673	−0.283	2.852	3.128	−0.85	2.739
β_1	−0.957	−0.893	0.066	0.593	0.664	−0.021	0.453
β_2	−1.987	−1.696	0.444	1.935	2.222	0.332	1.097
σ^2	45.641	49.314	74.668	119.905	132.288	78.394	22.407
ϕ	0.029	0.029	0.116	0.232	0.261	0.122	0.068

Note: S.D. represents the standard deviations for each model parameter.

We implemented 200,000 iterations for the posterior calculations to obtain a convergent sequence and approximately independent posterior samples. The first 100,000 iterations were discarded as burn-in. Then, one has an approximately independent joint posterior sample size of 100,000 by subsampling every 10th scan. The execution time for 200,000 MCMC iterations was 56.26471 s on an i7-12700 2.10 GHz PC. The system environment was R language version 4.2.3 lined to Intel's Math Kernal Library (MKL) on Windows 11. The core codes of the MCMC process were implemented using custom-written code without relying on external packages.The trace plot in Figure 4 displays the logarithm values of Equation (7) for the 200,000 MCMC iterations. Given that the proposed model incorporates multiple parameters and random effect terms, we assessed the overall convergence of the MCMC process using these logarithm values. Notably, the trace plot reveals that it belongs to an interval within the 200,000 iterations, implying that the MCMC process has reached convergence. Table 3 presents posterior inferences based on 10,000 posterior samples for model parameters. Furthermore, the posterior means of $R_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i)$ for $i = 1, \ldots, 24$, are shown in Figure 5. Figure 6 displays the results with estimated occurrence rates P_i, $i = 1, \ldots, 24$ in each grid.

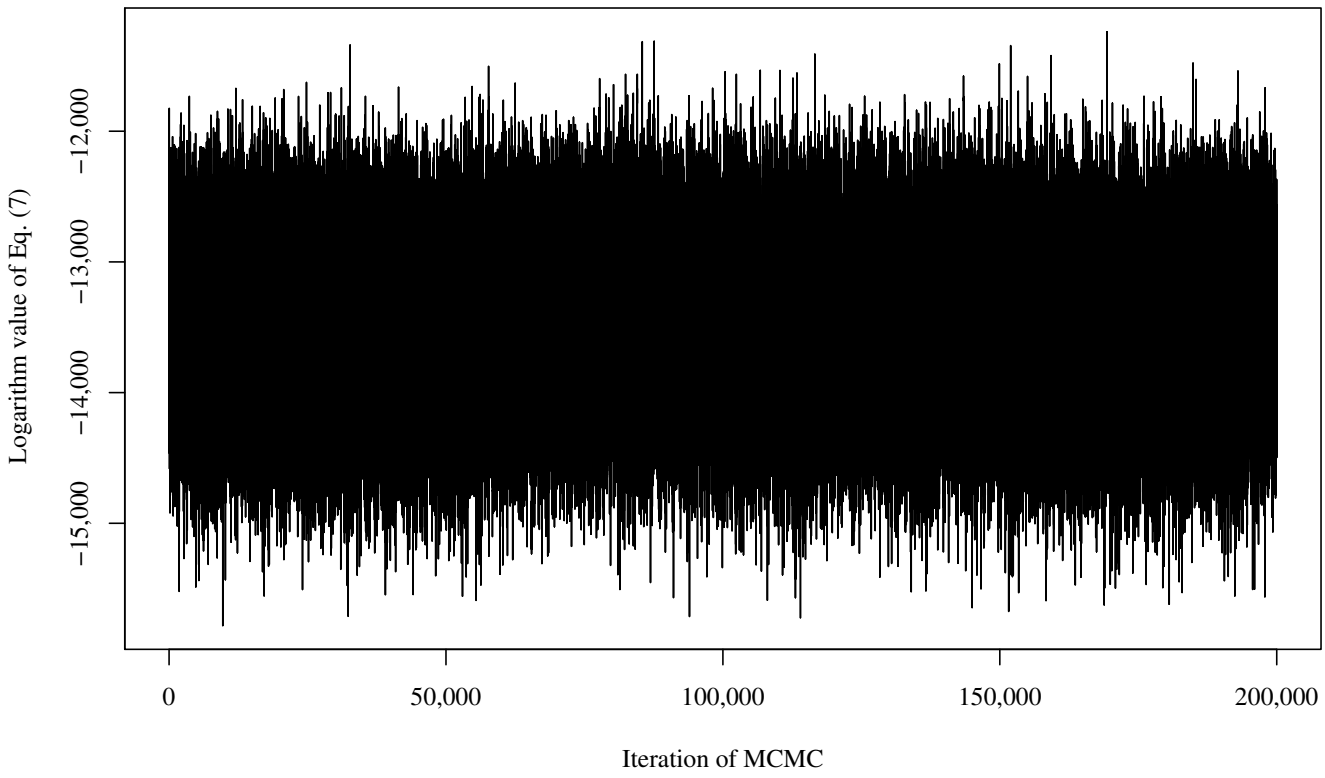

Figure 4. The logarithm trace plot of Equation (7) for the 200,000 MCMC iterations.

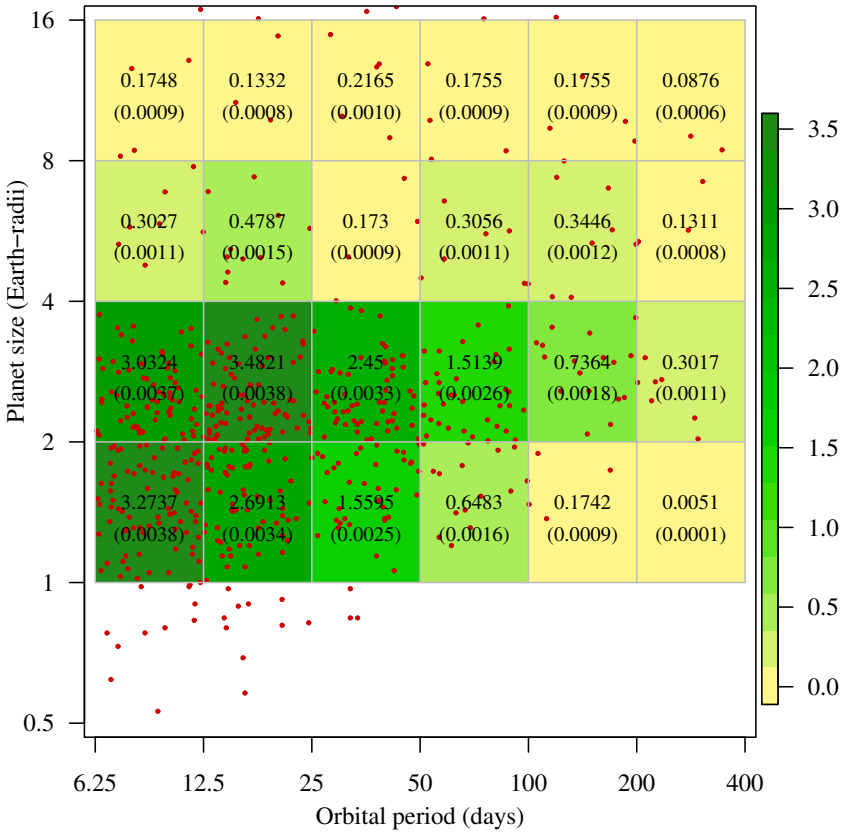

Figure 5. The posterior means of $R_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta_i)$ for $i = 1, \cdots, 24$.

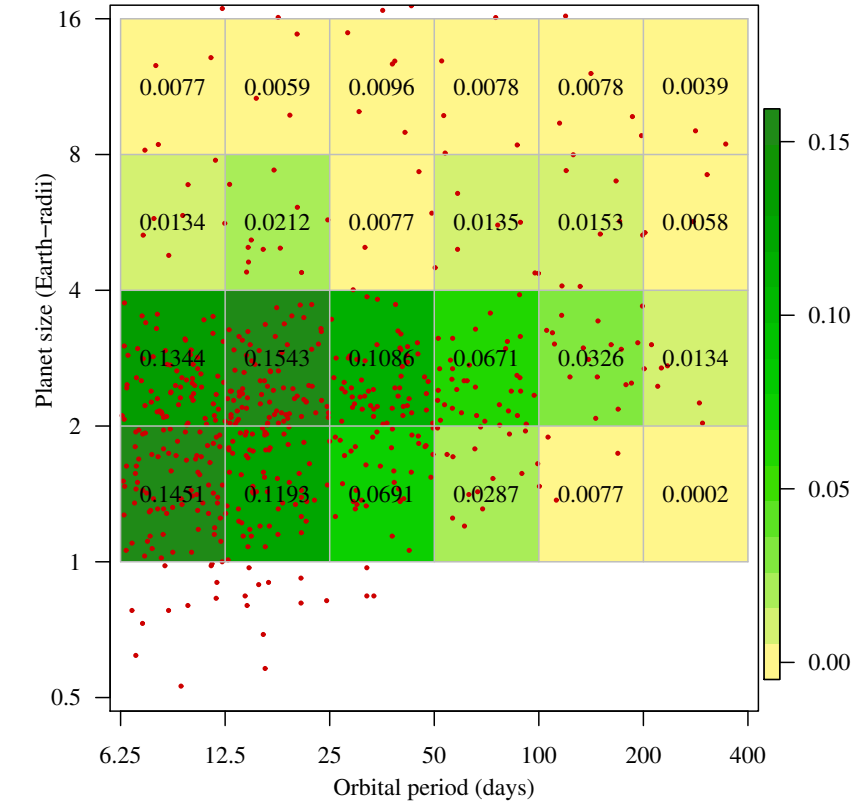

Figure 6. The estimated occurrence rates $P_i = R_i / \sum_{j=1}^{24} R_j$ for $i = 1, \cdots, 24$ without using completeness.

Next, we considered the variable detection efficiency (or completeness) in order to identify realistic occurrence rates. After obtaining the estimated occurrence rates in each cell shown in Figure 6, we further considered the survey completeness in order to identify realistic occurrence rates. The values of completeness function used here were constructed by Foreman-Mackey et al. (2014) [28]. We can thus obtain the true occurrence rates P_i^{tr}, $i = 1, \ldots, 24$ in each cell, as shown in Figure 7. Because the method proposed in this paper is presented as a totally different approach to that of Petigura et al. (2013) [5], we need to make a comparison with Petigura et al.'s method (2013) [5]. We computed realistic occurrence rates with different values of orbital period (P) and planet radius (R) and the corresponding realistic occurrence rates, as shown in Table 4. Note that case (i) in Table 4 corresponds to Jupiter-size planets.

From Table 4, we find that (1) for cases (ii), (iii), (iv), (vii), and (ix), the occurrence rates obtained from the proposed method are larger than those of Petigura et al. (2013) [5] by approximately a factor of two; (2) for cases (i), (v), and (vi), the occurrence rates obtained the proposed method are almost the same as Petigura et al.'s (2013) [5]; and (3) for cases (viii) and (x), the occurrence rates obtained by Petigura et al. (2013) [5] are larger than the proposed method herein. Because the proposed model considers the information of neighbors, the grid density is high, which will produce higher occurrence rates. On the contrary, if the grid density is low, lower rates will occur. Furthermore, both methods confirm the occurrence rates of planets with (i) P = 5–100 d and size 8–16 $R_\oplus$; (ii) P = 25–50 d and size 1–16 $R_\oplus$; and (iii) P = 50–100 d and size 1–16 $R_\oplus$.

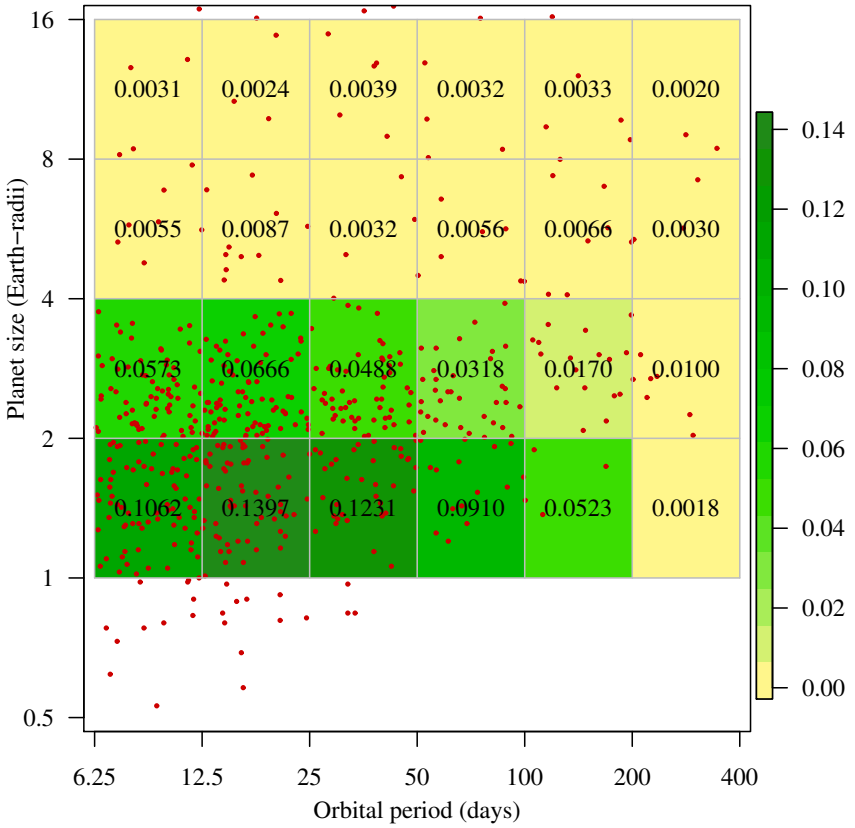

Figure 7. The true occurrence rates P_i^{tr} for $i = 1, \cdots, 24$.

Furthermore, we are interested in the occurrence rate of Earth analogs hosted by GK dwarf stars, i.e., P = 200–400 d and size 1–2 $R_\oplus$. From the scatter plot shown in Figure 1, there are no planets in this grid, and there are few planets in the neighborhood of this grid. Thus, it is reasonable that the occurrence rate of this grid is very small. After completeness correction, we find the occurrence rate to be 0.18% (please see case (viii) in

Table 4), whereas the values obtained by Petigura et al. (2013) [5], Foreman-Mackey et al. (2014) [28], and Chen and Hung (2019) [29] are 5.7%, 1.9%, and 2.5%, respectively. The proposed method indicates that 46% of Sun-like stars have an Earth-size (1–2 $R_\oplus$) planet with P = 5–100 d. This value is higher than Petigura et al.'s (2013) [5] due to the spatial model considering the information of neighbors. We further conducted an additional extrapolation of the hot Jupiter occurrence rate (i.e., the occurrence rate of 1–10 days and 8–24 $R_\oplus$) and compared it to the findings of Petigura et al. (2018) [24]. Their study reported a hot Jupiter occurrence rate of 0.57%, whereas our extrapolated estimate stands at 4.17%. According to the scalability of our proposed model, it provides an extrapolation with new data. To the best of our knowledge, utilizing neighboring data information for occurrence rate estimation in astronomy is a novel approach that has not been previously observed. According to the inference of Petigura et al. (2013) [5], we may imply that the nearest Earth-size planets in habitable zones of Sun-like stars are expected to orbit a star further than 12 light-years from Earth because we adopted the 46% occurrence rate.

Table 4. Comparison of realistic occurrence rates with different values of orbital period (P) and planet radius (R).

Case	Period (P)	Radius (R)	Petigura et al. (2013) [5]	The Proposed
(i)	5–100 d	8–16 $R_\oplus$	1.6%	1.26%
(ii)	5–100 d	1–2 $R_\oplus$	26%	46% *
(iii)	6.25–12.5 d	1–16 $R_\oplus$	8.9%	17.21% *
(iv)	12.5–25 d	1–16 $R_\oplus$	13.7%	21.74% *
(v)	25–50 d	1–16 $R_\oplus$	15.7%	17.9%
(vi)	50–100 d	1–16 $R_\oplus$	15.2%	13.16%
(vii)	6.25–25 d	1–2 $R_\oplus$	11.5%	24.59% *
(viii)	200–400 d	1–2 $R_\oplus$	5.7%	0.18%
(ix)	<50 d	1–2 $R_\oplus$	19.2%	36.9% *
(x)	200–400 d	2–4 $R_\oplus$	5%	1%

* represents the occurrence rates obtained by the proposed method are larger than Petigura et al. (2013) by approximately a factor of 2.

4. Discussion

Motivated by the study of Petigura et al. (2013) [5] on the prevalence of Earth-size planets orbiting Sun-like stars, we adopted a joint modeling approach to investigate the occurrence rates of planets around GK dwarfs. The inferred occurrence rate of Earth analogs around GK dwarfs increases to 46%. Compared with that of Petigura et al. (2013) [5], our approach increases the occurrence rate of Earth analogs by approximately a factor of two. Nevertheless, our model suggests that the occurrence rate for Kepler planets with radii between 1 and 2 earth radii and orbital periods between 50 and 400 days is 0.1451. Similar to most of the results in the literature, our occurrence rate of 0.1451 is also larger than the results computed by Petigura et al. (2013) [5], Dong and Zhu (2013) [13], and Foreman-MacKey et al. (2014) [28]. We cautiously contend that our proposed model exhibits a higher occurrence rate compared to other methods, attributed to the incorporation of spatial random effects. These effects effectively capture the spatial correlation in the data and moderately compensate for any missing explanatory variables. Applying our analysis to the entire Kepler planet sample ($Q1 - Q16$) will be left to future work. On the other hand, the current approach does not consider the influence of time on occurrence rates. All the data are treated as being from the same time point. Given the flexible nature of the proposed model, we plan to incorporate the effect of time using a Poisson process in future expansions. The expanded model will allow for dynamic predictions of variables over time.

Taking into account the survey incompleteness, we confirm the study of Petigura et al. (2013) [5]: the occurrence rates of planets with (i) P = 5–100 d and size 8–16 $R_\oplus$; (ii) P = 25–50 d and size 1–16 $R_\oplus$; and (iii) P = 50–100 d and size 1–16 $R_\oplus$. The inferred occurrence rates of Kepler planets suffer severely from systematic uncertainties (see Burke 2015 [14]). Follow-up spectroscopic observations of host stars will refine some of these

uncertainties, providing a planet sample with better stellar parameters and pipeline completeness for our model and others to revise the proposed model, and thus present better constraining theories of planet formation and evolution.

Author Contributions: Methodology, H.-D.Y.; software, H.-D.Y.; formal analysis, Y.-H.L. and C.-Y.L.; resources, H.-D.Y.; writing—original draft, H.-D.Y., Y.-H.L. and C.-Y.L.; writing—review and editing, Y.-H.L. and C.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the National Science and Technology Council, R.O.C., grant number MOST 109-2118-M-390-001, and MOST 111-2118-M-390-003.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moraga, P.; Lawson, A.B. Gaussian component mixtures and CAR models in Bayesian disease mapping. *Comput. Stat. Data Anal.* **2012**, *56*, 1417–1433. [CrossRef]
2. Duncan, E.W.; White, N.M.; Mengersen, K. Spatial smoothing in Bayesian models: A comparison of weights matrix specifications and their impact on inference. *Int. J. Health Geogr.* **2017**, *16*, 47. [CrossRef]
3. Lawson, A.B. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*; Chapman and Hall: London, UK; CRC: Boca Raton, FL, USA, 2018.
4. Baer, D.R.; Lawson, A.B. Evaluation of Bayesian multiple stage estimation under spatial CAR model variants. *J. Stat. Comput. Simul.* **2019**, *89*, 98–144. [CrossRef]
5. Petigura, E.A.; Howard, A.W.; Marcy, G.W. Prevalence of Earth-size planets orbiting Sun-like stars. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19273–19278. [CrossRef] [PubMed]
6. Batalha, N.M. Exploring exoplanet populations with NASA's Kepler Mission. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 12647–12654. [CrossRef]
7. Burke, J.E.; Inglis, A.J.; Perisic, O.; Masson, G.R.; McLaughlin, S.H.; Rutaganira, F.; Shokat, K.M.; Williams, R.L. Structures of PI4KIIIβ complexes show simultaneous recruitment of Rab11 and its effectors. *Science* **2014**, *344*, 1035–1038. [CrossRef] [PubMed]
8. Mullally, F.; Coughlin, J.L.; Thompson, S.E.; Rowe, J.; Burke, C.; Latham, D.W.; Batalha, N.M.; Bryson, S.T.; Christiansen, J.; Henze, C.E.; et al. Planetary candidates observed by Kepler. VI. Planet sample from Q1–Q16 (47 months). *Astrophys. J. Suppl. Ser.* **2015**, *217*, 31. [CrossRef]
9. Benz, W.; Ida, S.; Alibert, Y.; Lin, D.; Mordasini, C. Planet population synthesis. *arXiv* **2014**, arXiv:1402.7086.
10. Fressin, F.; Torres, G.; Charbonneau, D.; Bryson, S.T.; Christiansen, J.; Dressing, C.D.; Jenkins, J.M.; Walkowicz, L.M.; Batalha, N.M. The false positive rate of Kepler and the occurrence of planets. *Astrophys. J.* **2013**, *766*, 81. [CrossRef]
11. Lissauer, J.J.; Marcy, G.W.; Bryson, S.T.; Rowe, J.F.; Jontof-Hutter, D.; Agol, E.; Borucki, W.J.; Carter, J.A.; Ford, E.B.; Gilliland, R.L.; et al. Validation of Kepler's multiple planet candidates. II. Refined statistical framework and descriptions of systems of special interest. *Astrophys. J.* **2014**, *784*, 44. [CrossRef]
12. Santerne, A.; Moutou, C.; Tsantaki, M.; Bouchy, F.; Hébrard, G.; Adibekyan, V.; Almenara, J.M.; Amard, L.; Barros, S.; Boisse, I.; et al. SOPHIE velocimetry of Kepler transit candidates-XVII. The physical properties of giant exoplanets within 400 days of period. *Astron. Astrophys.* **2016**, *587*, A64. [CrossRef]
13. Dong, S.; Zhu, Z. Fast Rise of "Neptune-Size" Planets (4–8 $R_{\oplus}$) from P 10 to 250 days—Statistics of Kepler Planet Candidates up to 0.75 AU. *Astrophys. J.* **2013**, *778*, 53. [CrossRef]
14. Burke, C.J.; Christiansen, J.L.; Mullally, F.; Seader, S.; Huber, D.; Rowe, J.F.; Coughlin, J.L.; Thompson, S.E.; Catanzarite, J.; Clarke, B.D.; et al. Terrestrial planet occurrence rates for the Kepler GK dwarf sample. *Astrophys. J.* **2015**, *809*, 8. [CrossRef]
15. Dressing, C.D.; Charbonneau, D. The occurrence of potentially habitable planets orbiting M dwarfs estimated from the full Kepler dataset and an empirical measurement of the detection sensitivity. *Astrophys. J.* **2015**, *807*, 45. [CrossRef]
16. Silburt, A.; Gaidos, E.; Wu, Y. A statistical reconstruction of the planet population around Kepler solar-type stars. *Astrophys. J.* **2015**, *799*, 180. [CrossRef]
17. Morton, R.; Verth, G.; Fedun, V.; Shelyag, S.; Erdélyi, R. Evidence for the photospheric excitation of incompressible chromospheric waves. *Astrophys. J.* **2013**, *768*, 17. [CrossRef]
18. Besag, J.; York, J.; Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **1991**, *43*, 1–20. [CrossRef]
19. Chen, C.S.; Yang, H.D. A joint modeling approach for spatial earthquake risk variations. *J. Appl. Stat.* **2011**, *38*, 1733–1741. [CrossRef]
20. Cressie, N. *Statistics for Spatial Data*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
21. Kelsall, J.; Wakefield, J. Modeling spatial variation in disease risk: A geostatistical approach. *J. Am. Stat. Assoc.* **2002**, *97*, 692–701. [CrossRef]

22. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Methodol.* **1974**, *36*, 192–225. [CrossRef]
23. De Oliveira, V. Bayesian analysis of conditional autoregressive models. *Ann. Inst. Stat. Math.* **2012**, *64*, 107–133. [CrossRef]
24. Petigura, E.A.; Marcy, G.W.; Winn, J.N.; Weiss, L.M.; Fulton, B.J.; Howard, A.W.; Sinukoff, E.; Isaacson, H.; Morton, T.D.; Johnson, J.A. The California-Kepler survey. IV. Metal-rich stars host a greater diversity of planets. *Astron. J.* **2018**, *155*, 89. [CrossRef]
25. Tierney, L. Markov chains for exploring posterior distributions. *Ann. Stat.* **1994**, *22*, 1701–1728. [CrossRef]
26. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741. [CrossRef] [PubMed]
27. Chib, S.; Greenberg, E. Understanding the metropolis-hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
28. Foreman-Mackey, D.; Hogg, D.W.; Morton, T.D. Exoplanet population inference and the abundance of Earth analogs from noisy, incomplete catalogs. *Astrophys. J.* **2014**, *795*, 64. [CrossRef]
29. Chen, J.H.; Hung, W.L. Parametrizing the Kepler exoplanet period-radius distribution with the bivariate normal inverse Gaussian distribution. *J. Appl. Stat.* **2019**, *46*, 725–736. [CrossRef]