stats

MDPI

Article

Theoretical Advancements in Small Area Modeling: A Case Study with the CHILD Cohort

Charanpal Singh [†,‡] **and Mahmoud Torabi** [*,‡]

Department of Community Health Sciences, University of Manitoba, Winnipeg, MB R3E 0W3, Canada; charanjeet.singh@umanitoba.ca

* Correspondence: mahmoud.torabi@umanitoba.ca

† Current address: S113 Medical Services Building, 750 Bannatyne Ave., Winnipeg, MB R3E 0W3, Canada.

‡ These authors contributed equally to this work.

Abstract: Developing accurate predictive models in statistical analysis presents significant challenges, especially in domains with limited routine assessments. This study aims to advance the theoretical underpinnings of longitudinal logistic and zero-inflated Poisson (ZIP) models in the context of small area estimation (SAE). Utilizing data from the Canadian Healthy Infant Longitudinal Development (CHILD) study as a case study, we explore the use of individual- and area-level random effects to enhance model precision and reliability. The study evaluates various covariates' impact (such as mother's asthma, mother wheezed, mother smoked) on model performance to predict child's wheezing, emphasizing the role of location within Manitoba. Our main findings contribute to the literature by providing insights into the development and refinement of small area models, emphasizing the significance of advancing theoretical frameworks in statistical modeling.

Keywords: Bayesian; longitudinal data; regression; small area estimation; zero-inflated Poisson

check for updates

Academic Editor: Wei Zhu

Received: 26 March 2025
Revised: 2 May 2025
Accepted: 12 May 2025
Published: 16 May 2025

Citation: Singh, C.; Torabi, M. Theoretical Advancements in Small Area Modeling: A Case Study with the CHILD Cohort. *Stats* **2025**, *8*, 39. https://doi.org/10.3390/stats8020039

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Predictive modeling in statistics often involves estimating parameters within complex data structures. This paper focuses on the theoretical development of longitudinal logistic and zero-inflated Poisson (ZIP) models, addressing challenges posed by small sample sizes and sparse data. The context of asthma prevalence in Manitoba serves as a case study to illustrate these models' application. However, the primary objective is to advance the theoretical framework and methodologies underlying these models, beyond asthma-specific applications.

Asthma is a chronic disease characterized by airway inflammation, leading to variable expiratory airflow limitation and symptoms such as shortness of breath, chest tightness, coughing, and wheezing [1]. In Canada, asthma prevalence is highest among individuals aged 1–34 years [2], with incidence rates gradually increasing each year, particularly among younger demographics [3,4]. Manitoba, a sparsely populated province with the majority of its population concentrated in Winnipeg, presents unique challenges in studying asthma due to small sample sizes in many areas.

This study aims to develop robust statistical models for small area estimation (SAE) to assess childhood wheezing—a symptom of asthma—and its associated risk factors, leveraging longitudinal data from the Canadian Healthy Infant Longitudinal Development (CHILD) study. By focusing on methodological and statistical techniques, this study seeks

to enhance the precision and accuracy of estimating wheezing severity across different regions of Manitoba.

To address the issue of small sample sizes, we employed small area models to analyze wheezing rates across different regions. A "small area" is defined as any domain for which direct estimates of adequate precision cannot be produced [5]. Small area models leverage auxiliary variables and random area effects to improve predictions of wheezing rates and severity in geographical areas [5]. Our goal is to refine the theoretical framework and methodologies of these models, advancing the field of small area estimation and providing more accurate and reliable estimates for public health research.

Longitudinal logistic models are designed to handle binary or ordinal response variables, particularly when data are collected over multiple time points. These models account for within-subject correlations and allow for the inclusion of random effects. In this study, we utilized hierarchical Bayesian logistic models to estimate the probability of wheezing severity at different follow-ups, incorporating individual-, area-, and interaction-level random effects.

ZIP models are suitable for count data with an excess number of zeroes, which is common in longitudinal health studies. These models assume that the population consists of two groups: those who never experience the event (always zero) and those who may experience the event (Poisson-distributed counts). The ZIP models in this study incorporate area and follow-up level effects to address overdispersion and excess zeros in wheezing episode counts.

The Canadian Healthy Infant Longitudinal Development (CHILD) study provided a unique opportunity to explore these theoretical advancements. The CHILD study includes detailed longitudinal data on childhood wheezing and associated risk factors, collected from a cohort of children born between 2008 and 2012 and followed up until 2.5 years of age. The data were first accessed for this study on 6 June 2020. The time points for this study were prenatal, 3, 6, 12, 18 months and 2 and 2.5 years. The Manitoba sub-cohort was used for this study and members were recruited within a radius of Winnipeg and Morden-Winkler; families who moved out of this region were still followed. For this particular study, participants were assumed to have the same location throughout the study. This was done using postal codes from the 2-year follow-up, since this follow- up contained the most unique postal codes. Furthermore, individuals with incomplete follow-ups for postal codes or missing postal codes were removed, which resulted in a sub-cohort of 690 families. The responses from the home environment questionnaire indicated that between 80 and 100 participants moved between each follow-up in Manitoba. The primary outcome of the CHILD study was expert physician diagnosed asthma, but wheezing was one of the secondary outcomes. However, it can be challenging to make a diagnosis of asthma in children aged five years since it is not possible to have routine airflow limitation measurement within younger age groups [1]. Wheezing symptoms, however, can still be present in children under five years old. Thus, the difficulty of diagnosing asthma in younger children means that wheezing can be used as an indicator of "possible" asthma diagnosis in the future [6].

For the purpose of this study, Manitoba was divided into small geographical regions (areas) and thus more information was provided regarding which regions had higher rates of wheezing and which factors contributed. The province has five regional health authorities (RHAs) created from the initial eleven RHAs in 2012. These five RHAs are further divided into 96 regional health authority districts (RHADs); these RHADs were intended to be used as small areas for this study. However, for this study, only 60 of the 96 RHADs were recruited by the CHILD cohort. Thus, the 60 recruited RHADs were used as small areas in this study and the RHADs which were not recruited were greyed out in

the figures throughout this study. The Manitoba sub-cohort, comprising 690 families, was used to analyze wheezing severity and its predictors.

By employing hierarchical Bayesian logistic models and ZIP models, we aimed to identify significant predictors of wheezing severity and rates, such as maternal asthma, smoking, and environmental factors. The integration of individual- and area-level covariates allowed us to capture the variability within small areas and enhance the precision of our estimates.

2. Small Area Models

2.1. Linear Mixed Model

Basic types of small area models are area-level and unit-level models. Area-level models are also called aggregate models, which relate small area means to area-specific auxiliary covariates [5,7,8]. Unit-level models relate the unit values within a small area to unit-specific auxiliary variables. Unit-level auxiliary variables provide individual-level information [9]. Assuming the response of the model is y_k, the parameter of interest is a function of the small area means $\bar{Y}_k$ for area k such that $\mu_k = g(\bar{Y}_k)$. In this case, $g(\cdot)$ is the link function that connects the response y_k with the covariates and random components of the model. In the case of the linear model, the link function $g(\cdot)$ is the identity, for the case of the logistic model, the link function $g(\cdot)$ is logit, and for the Poisson model, the link function $g(\cdot)$ is log. It is crucial to understand that the end goal of small area analysis is to estimate the small area parameter μ_k. Assume the indirect estimator of μ_k is

$$\widehat{\mu_k} = g\left(\widehat{\bar{Y}}_k\right) = \mu_k + e_k, \quad k = 1, \ldots, K, \tag{1}$$

where K is the number of small areas and the error terms e_k $(k = 1, \ldots, K)$ are independently and normally distributed with mean 0 and variance ψ_k. Note that it is assumed that ψ_k is known to avoid identifiability issues. Further, if area-level auxiliary data $\boldsymbol{x}_k = \left(x_{1k}, \ldots, x_{pk}\right)^T$ are available then a linear model can be formed as

$$\mu_k = \mathbf{x}_k^T \boldsymbol{\beta} + b_k v_k, \quad k = 1, \ldots, K, \tag{2}$$

where b_k's are the known positive constants; these are to set optional different random effects for each small area. The v_k term represents the area-specific random effects with the assumption that all v_k's are independent and identically distributed (iid) with mean 0 and variance σ_v^2, as denoted by $v_k \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_v^2\right)$. Note that v_k's are assumed to be normally distributed but are not limited to this distribution. Lastly, $\boldsymbol{\beta}$ is a vector of regression coefficients. Combining (1) with (2) results in

$$\widehat{\mu_k} = \mathbf{x}_k^T \boldsymbol{\beta} + b_k v_k + e_k, \quad k = 1, \ldots, K, \tag{3}$$

which is the final form of the basic area-level model and is also referred to as the Fay–Herriot (FH) model [10]. Now, if there are unit-specific auxiliary data available within each small area, then the basic unit-level linear model can be used to model small areas. Assuming that i is any element, say an individual within any small area k, and n_k represents the sample size in each area, then the model is

$$y_{ik} = \mathbf{x}_{ik}^T \boldsymbol{\beta} + v_k + e_{ik}, \quad i = 1, \ldots, n_k, \quad k = 1, \ldots K. \tag{4}$$

In Equation (4), $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})^T$ are the auxiliary data and $\boldsymbol{\beta}$ are the corresponding regression coefficients. Like the FH model, v_k represents the area-level random effects while the sampling error is $e_{ik} \sim N\left(0, t_{ik}\sigma_e^2\right)$ with known constants t_{ik} but unknown vari-

ance σ_e^2. It should be kept in mind that the parameters of interest like the FH model are the small area means, $\overline{Y}_k$. Note that one can also incorporate area-level and unit-level covariates into the unit-level model (4) as is the case for the CHILD cohort.

In general, to estimate small area means, the best predictor (BP) is used. The BP of a parameter of interest, such as the small area mean is the conditional expectation of the parameter given the data and the model parameters. Thus, the BP assumes that the model parameters are known. The BP does not have a closed form for most non-linear models. Thus, numerical methods were needed to estimate such models. In particular, Bayesian methods were used for this study; hence, diffuse priors were implemented on the model parameters β, σ_v^2, and σ_e^2.

2.2. Logistic Mixed Model

Suppose now that y_{ik} is binary, and we are interested in the small area proportions P_k, $k = 1, \ldots K$, then the hierarchical Bayesian (HB) area-level logistic model [7,8] is

$$y_k \mid p_k \sim \text{Binomial}(p_k, n_k),$$

$$\text{logit}(p_k) = \mathbf{x}_k^T \boldsymbol{\beta} + v_k, \text{with } v_k \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$$

and priors,

$$f(\boldsymbol{\beta}) \propto 1 \text{ and } \sigma_v^{-2} \sim \text{Gamma}(a, b) \tag{5}$$

Note the diffuse prior chosen for the inverse of variance parameters is a gamma distribution with shape and location parameters 0.001 ($a = b = 0.001$). If unit-specific auxiliary data are available within each small area, and we are interested in the small area proportions $\overline{Y}_k = P_k = \sum_{i=1}^{N_k} y_{ik}/N_k$, $k = 1, \ldots K$, where N_k is the number of individuals in area k, then the unit-level logistic model is

$$y_{ik} \mid p_{ik} \sim \text{Bernoulli}(p_{ik}),$$

$$\text{logit}(p_{ik}) = \mathbf{x}_{ik}^T \boldsymbol{\beta} + v_k, \text{with } v_k \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$$

and priors,

$$f(\boldsymbol{\beta}) \propto 1 \text{ and } \sigma_v^{-2} \sim \text{Gamma}(a, b) \tag{6}$$

2.3. Poisson Mixed Model

Now, we consider the count data in the response by assuming that y_k are area-level counts and are distributed Poisson variables [7,8]. The parameter of interest in this model would be λ_k, the true average rate of wheezing in the *kth* area. Thus, the model becomes

$$log(\lambda_k) = \mathbf{x}_k^T \boldsymbol{\beta} + v_k, \tag{7}$$

where $k = 1, \ldots, K$ and $v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$. The same priors as in Section 2.2 were used for β and σ_v^2.

2.4. Inference

A Monte Carlo Markov Chain (MCMC) was used to estimate the parameters of interest for the above models (5), (6), and (7). We used the R package Rjags to implement the MCMC. All models in thus study were run on Rjags; the initial models had 200,000 iterations with a thinning degree of 18. Diagnostic tools such as the Gelman–Rubin test were used to determine if the model converged to the target posterior distribution. The Gelman–Rubin

diagnostic evaluates MCMC convergence by analyzing the differences between multiple Markov chains—a small within-chain variance indicates convergence [11]. Thus, multiple chains were sampled from each, with burn-in to ensure accurate results. If the initial model did not converge according to the Gelman diagnostics for all covariates, then the model was updated each time with 10,000 iterations. The model was updated until convergence was reached. Additionally, the 95% credible intervals for the posterior distributions were calculated using the equal-tail interval method, which is the default method in Rjags. Particular attention was made to the choice of priors and structure of the random effects. The Deviance Information Criterion (DIC) was used for model comparison and to determine model adequacy. DIC is often viewed as a trade-off between model adequacy and complexity [12]. DIC statistics were used throughout the model building stages of this study.

3. Proposed Statistics Models

For the CHILD study, our interest was to determine which factors affected the rate of wheezing and severity over time (follow-ups) along with geographical information. Wheezing severity was either a binary or an ordinal variable; thus, longitudinal logistic area models were required [13–15]. Furthermore, a model to analyze the number of wheezing episodes was needed for each follow-up. Thus, extensions of the logit (6) and Poisson (7) models were needed for this study. In general, the models have different link functions dependent on the type of response variable. In the context of the CHILD study, the response y_{ijk} was dependent on participant i, follow-up j, and area k. Assuming, without loss of generality, the model has three random effects which can depend on any combination of i, j and k and with response variable y_{ijk} and corresponding X_{ijk}, then, letting $\mu_{ijk} = E(y_{ijk})$, we obtain,

$$g(\mu_{ijk}) = \boldsymbol{X}_{ijk}^{T}\boldsymbol{\beta} + f_{ij} + v_k + I_{ijk}. \tag{8}$$

In Equation (8), f_{ij} was the random effect of the individual and follow-up. In this model, we assume the disturbance is an AR(1) (autoregressive-one) effect. I_{ijk} is the interaction random effect of all three levels interacting with one another and v_k is simply the random effect of each small area.

From model (8), $g(\cdot)$ is the link function that connects y_{ijk} with the covariates and random components of the model. The link functions of interest for this study were the logit and log, as shown below in Equations (9) and (10), for the logistic and Poisson models, respectively. The logit link function is used when the response variable is assumed to follow a binomial, multinomial or ordinal distribution (e.g., a child wheezed or not in our CHILD study). The log link function is primarily used when the response is assumed to follow a Poisson distribution (e.g., wheezing episodes in our CHILD study). So, accordingly, we have

$$g\left(\mu_{ijk}\right) = \text{logit}\left(\mu_{ijk}\right) = \log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) \tag{9}$$

$$g(\mu_{ijk}) = \log\left(\mu_{ijk}\right) \tag{10}$$

The three types of models of interest for this study were as follows.

3.1. Ordinal Logistic Model

A logistic model with the assumption that the response, y_{ijk}, is ordinal was used to quantify how much of an impact location and time had on wheezing severity. Here, y_{ijk} represented the childhood wheezing severity of child i in a particular area k at follow-up j.

Now, the ordinal logistic model is described. In our CHILD study, y_{ijk} have three levels $l = 1, 2, 3$; thus, there are two categories $c = 1, 2$ (with another level as a reference category). We then obtain

$$\log\left[\frac{\text{Prob}\left(y_{ijk} \leq c\right)}{\text{Prob}\left(y_{ijk} > c\right)}\right] = a_c + \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{ijk} + f_{ij} + v_k + I_{ijk}, \tag{11}$$

where i represents the individual, j is follow-up, and k is the small area, and

$$\begin{gathered} v_k \sim \mathcal{N}(0, \sigma_v^2) \\ f_{ij} = \rho f_{ij-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_f^2\right) \\ f_{ij} \mid f_{ij-1} \sim \mathcal{N}\left(\rho f_{ij-1}, \sigma_f^2\right) \\ I_{ijk} \sim \mathcal{N}(0, \sigma_I^2) \\ -1 < \rho < 1. \end{gathered}$$

Thus, the two logit equations with $Q(m) = \text{Prob}\left(y_{ijk} \leq c\right)$ are

$$\begin{aligned} \text{logit}[\text{Q}(1)] &= a_1 + \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{ijk} + f_{ij} + v_k + I_{ijk} \\ \text{logit}[\text{Q}(2)] &= a_2 + \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{ijk} + f_{ij} + v_k + I_{ijk}. \end{aligned}$$

Note, in this case, the vector of covariates $\boldsymbol{X}_{ijk}$ does not include the intercept term since this is represented by a_1 and a_2. In general, a_c was a level specific intercept with the restriction that $a_1 < \cdots < a_c$ such that $c = 1, \cdots, l-1$.

3.2. Binary Logistic Model

Due to the lack of viable data, the severity of wheezing response variable was transformed to two levels. This was accomplished by merging responses 2 and 3 into one level, as the numbers in these two levels were small compared to level 1. The resulting model is described as follows:

$$y_{ijk} \mid p_{ijk} \sim \text{Bernoulli}\left(p_{ijk}\right).$$

Then, $p_{ijk} = \text{Prob}(y_{ijk} = 1)$ is the probability that the individual i had wheezing severity at follow-up j and small area k. Then, the logistic model is

$$\log\left[\frac{p_{ijk}}{1 - p_{ijk}}\right] = \boldsymbol{X}_{ijk}^T\boldsymbol{\beta}_{ijk} + f_{ij} + v_k, \tag{12}$$

where

$$\begin{gathered} v_k \sim \mathcal{N}(0, \sigma_v^2) \\ f_{ij} = \rho f_{ij-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_f^2\right) \\ f_{ij} \mid f_{ij-1} \sim \mathcal{N}\left(\rho f_{ij-1}, \sigma_f^2\right) \\ -1 < \rho < 1, \end{gathered}$$

note that we ignored the interaction term I_{ijk} in model (12) to reduce the number of parameters and variability in the model.

3.3. Zero-Inflated Poisson Model

A different model was required to model the count data from the CHILD study. While a Poisson regression model could have been used, there were issues with this approach, including excess zeros in the data and overdispersion. As a result, a Zero-Inflated Poisson (ZIP) model was used to allow for excess zeros in the dependent variable

and overdispersion [16]. It was assumed that there were two types of participants in the study: those with always-zero counts and those with counts predicted by the standard Poisson. The ZIP model is a mixture of probabilities from the two groups and allows for both overdispersion and excess zeros. Area and follow-up level effects were introduced to the ZIP model, similar to previous logistic models, but to reduce the number of parameters, I_{ijk} was not included in the model.

$$\Theta_{ijk} = \boldsymbol{X}_{ijk}^T \boldsymbol{\beta}_{ijk} + f_{ij} + v_k, \tag{13}$$

where

$$\begin{gathered} v_k \sim \mathcal{N}\left(0, \sigma_v^2\right) \\ f_{ij} = \rho f_{ij-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_f^2\right) \\ f_{ij} \mid f_{ij-1} \sim \mathcal{N}\left(\rho f_{ij-1}, \sigma_f^2\right) \\ -1 < \rho < 1. \end{gathered}$$

Then, the ZIP model becomes

$$\text{Prob}\left(y_{ijk} = m\right) = \begin{cases} \pi_{ijk} + \left(1 - \pi_{ijk}\right) \exp\left(-\lambda_{ijk}\right) & \text{if } m = 0 \\ \left(1 - \pi_{ijk}\right) \frac{\lambda_{ijk}^{y_{ijk}} \exp\left(-\lambda_{ijk}\right)}{y_{ijk}!} & \text{if } m > 0 \end{cases} \tag{14}$$

such that $\log(\lambda_{ijk}) = \Theta_{ijk}$, $logit(\pi_{ijk}) = \Theta_{ijk}$, and m is the number of wheezing episodes. Here, λ_{ijk} represents the expected rate of the Poisson component, while π_{ijk} represents the probability of excess zeroes. Then, the expected value of the ZIP model is $E\left(y_{ijk}\right) = (1 - \pi_{ijk})\lambda_{ijk}$. Similarly, the variance of the ZIP model is $\text{Var}\left(y_{ijk}\right) = \lambda_{ijk}(1 - \pi_{ijk})(1 + \pi_{ijk}\lambda_{ijk})$.

In ZIP models, it is possible to choose different predictors for the true and excess zeros if it is expected that different variables are driving the presence or absence of the outcome in the data. For simplicity, the same covariates for both components were fitted in the ZIP model, as shown by the expression Θ_{ijk}. Typically, we are less interested in the modeling of false zeros; thus, a diffuse Bayesian prior for π_{ijk} was also implemented as an alternative model such that $logit(\pi_{ijk}) = logit(\pi) = \gamma_0$, $\gamma_0 \sim \mathcal{N}$(0,10,000).

3.4. Predicted Proportions and Rates

In addition, to study the effects of the covariates over time, we were also interested to investigate which areas (regions) are more at risk in terms of wheezing episodes and also their severity. Thus, we needed area-level models to predict the proportions and rates of wheezing. We considered the area-level binomial logit model, where y_k was the number of children who wheezed at area k for the first follow-up. The parameter of interest was $p_k = h(\mathbf{x}_k^T \boldsymbol{\beta} + v_k)$, where $h(t) = 1/\{1 + \exp(-t)\}$. Since the true population parameters were not known, they were estimated using the logit-model by $\hat{p}_k^{HB} = h(\mathbf{x}_k^T \hat{\boldsymbol{\beta}} + \hat{v}_k)$ directly from the MCMC samples. Here, $\hat{\boldsymbol{\beta}}$ was the vector of the estimated fixed effects and $\hat{v}_k$ was the predicted area-specific random effect. Note that $\hat{p}_k^{HB}$ represents the probability or proportion of wheezing in area k.

Similarly, for the ZIP model, our interest was the average count per area. We again restricted our ZIP model to the first follow-up. The model was described as

$$\Theta_{ik} = \mathbf{x}_{ik}^T \boldsymbol{\beta} + v_k \tag{15}$$

where $\mathbf{x}_{ik}$ are the unit-level covariates and $v_k \sim \mathcal{N}\left(0, \sigma_v^2\right)$.

Then, the ZIP model became

$$\text{Prob}(y_{ik} = m) = \begin{cases} \pi_{ik} + (1 - \pi_{ik}) \exp(-\lambda_{ik}) & \text{if } m = 0 \\ (1 - \pi_{ik}) \frac{\lambda_{ik}^{y_{ik}} \exp(-\lambda_{ik})}{y_{ik}!} & \text{if } m > 0 \end{cases} \tag{16}$$

such that $\log(\lambda_{ik}) = \Theta_{ik}$ and $\text{logit}(\pi_{ik}) = \Theta_{ik}$. Then, we assumed the simple case with covariates $\mathbf{x}_{ik}$ at area-level such that $\mathbf{x}_{ik} = \mathbf{x}_k$. Then, the area-level ZIP model became

$$\Theta_k = \mathbf{x}_k^T \boldsymbol{\beta} + v_k, \tag{17}$$

where $\mathbf{x}_k$ were the area-level covariates.

Then,

$$\text{Prob}(y_k = m) = \begin{cases} \pi_k + (1 - \pi_k) \exp(-\lambda_k) & \text{if } m = 0 \\ (1 - \pi_k) \frac{\lambda_k^{y_k} \exp(-\lambda_k)}{y_k!} & \text{if } m > 0 \end{cases} \tag{18}$$

Now, our parameter of interest was $\lambda_k = \exp(\mathbf{x}_k^T \boldsymbol{\beta} + v_k)$. Using the parameter estimates of the ZIP model, this became $\hat{\lambda}_k^{HB} = \exp(\mathbf{x}_k^T \hat{\boldsymbol{\beta}} + \hat{v}_k)$. Here, $\hat{\lambda}_k^{HB}$ represents the average number of wheezing episodes in area k.

4. Results

Simulation studies were conducted on the ordinal logistic and ZIP models to evaluate the performance of the proposed models. Furthermore, a simulation was conducted on a simpler binary logistic model. The results of the simulations are displayed in Appendix B. Due to the lack of viable data, the severity of the wheezing response variable was transformed to run a binary logistic model by merging responses 2 and 3 into one level. However, the binary logistic model yielded inflated odds ratios. After model building using a step-up approach, a simpler Hierarchical Bayes unit-level model was fitted [5]. The response variable was the severity of wheezing at 3-month follow-up, along with covariates at the prenatal and 3-month time points. The model findings are presented in Table 1. The unit-level model results demonstrated that maternal asthma and smoking during pregnancy increased the probability of developing wheezing, while living in close proximity to a farm and having furry pets reduced the odds of developing wheezing. Maternal asthma during pregnancy emerged as the sole significant predictor in this model.

Table 1. Unit -level binary logistic model (odds ratios were calculated by taking the exponent of the covariate parameter estimates).

Parameter	Estimate	95% Credible Interval (Lower, Upper)	Odds Ratio
Intercept	−2.74	(−3.25, −2.27)	-
Mother asthma (prenatal)	1.20	(0.63, 1.76)	3.31
Mother smoked (prenatal)	1.05	(−0.02, 2.01)	2.85
Farm (3 months)	−0.36	(−1.47, 0.60)	0.70
Furry pets (3 months)	−0.28	(−0.86, 0.28)	0.76
σ_v^2	0.37	(0.04, 0.95)	-

To analyze the wheezing rates of children in the cohort, ZIP models were fitted. Model building was performed using a step-up approach by adding covariates that were statistically and clinically significant one at a time. The results of the most viable final ZIP model are shown in Table 2. The unit-level ZIP was considered the most viable model since the preliminary models indicated no autocorrelation between follow-ups, and thus the

AR(1) was not needed in the model. The unit-level ZIP model used a diffuse prior for the logit component of the model, which accounted for the excess zeros in the data.

Table 2. Simple unit-level ZIP model (rate ratios were calculated by taking the exponent of the covariate parameter estimates).

Parameter	Estimate	95% Credible Interval (Lower, Upper)	Rate Ratio
Intercept	−1.22	(−2.17, −0.35)	-
Mother asthma (prenatal)	0.84	(0.18, 1.52)	2.32
Mother wheezed (prenatal)	0.02	(−0.63, 0.65)	1.02
Mother smoked (prenatal)	1.01	(−0.38, 2.44)	2.75
Farm (3 months)	0.90	(0.10, 1.78)	2.45
γ_0	1.45	(1.01, 1.88)	-
$1-\pi$	0.19	(0.13, 0.27)	-
σ_v^2	1.37	(0.86, 2.07)	-

The results of unit-level ZIP model in Table 2 indicated that there were a high number of false zeros in the data, since the estimate of $1-\pi$ meant that only 19% of individuals have true zero counts. This was to be expected since many of the individuals did not wheeze; thus, their number of wheezing episodes was 0. The results also implied that maternal wheezing seemed to marginally increase the expected rate of wheezing episodes in Manitoba. However, maternal smoking and asthma greatly increased the expected rate of wheezing episodes, while only maternal asthma was statistically significant. Interestingly, living near a farm also increased the rate of wheezing episodes among those who already had wheezing symptoms. The key distinction between the ZIP and logistic model was that the ZIP model was focused on individuals in the cohort who already had wheezing, as opposed to the odds of developing wheezing, which was the focus of the logistic model.

Proportions and Rates of Wheezing

The area-level models were fitted after individual-level auxiliary variables were converted to area-level covariates. It should be mentioned that aggregating the sample data resulted in the loss of some information, accounting for why some covariates had only a marginal contribution to the odds of wheezing. However, the main purpose of the area-level models was the prediction of wheezing proportions and rates for each RHAD. The results of the final area-level (after model building) are reported in Table 3.

The results of Table 3 show that the proportion of prenatal maternal asthma and wheezing in an area were marginally associated with increased odds of developing wheezing, while living near a farm slightly decreased the odds of developing wheezing. Interestingly, maternal smoking had no effect on the odds of developing wheezing. The proportion of predicted wheezing in each area k (=1, $\cdots$, K) was obtained by $\hat{p}_k^{HB} = h(\mathbf{x}_k^T\hat{\boldsymbol{\beta}} + \hat{v}_k)$ by using the results of Table 3. Then, the predicted proportions were mapped to their corresponding RHADs in Manitoba (Figure 1). Recall that for this study, we had information for 60 (out of 96) RHADs; thus, the rest appeared as missing in the figures. For comparison purposes, the area-level wheezing proportions were obtained from the data by dividing wheezing severity at 3 months by the sample size for each area, as shown in Figure 2. The proportions obtained from the raw data indicated some areas with almost no proportions of wheezing, while the predicted proportions obtained from the model showed slightly higher proportions. The reason for this was that the area-level logistic model predicted

the proportions of wheezing severity for each small area along with taking into account the auxiliary information available for each area. This resulted in smoothed predictions towards the mean, which is shown in Figure 1.

Table 3. Final area-level binary logistic model.

Parameter	Estimate	95% Credible Interval (Lower, Upper)	Odds Ratio
Intercept	−3.27	(−4.09, −2.52)	-
Mother wheezed (prenatal)	0.01	(−0.01, 0.04)	1.01
Mother asthma (prenatal)	0.02	(0.01, 0.05)	1.02
Mother smoked (prenatal)	0.00	(−0.03, 0.03)	1.00
Farm	−0.01	(−0.03, 0.00)	0.99
σ_v^2	0.25	(0.03, 0.84)	-

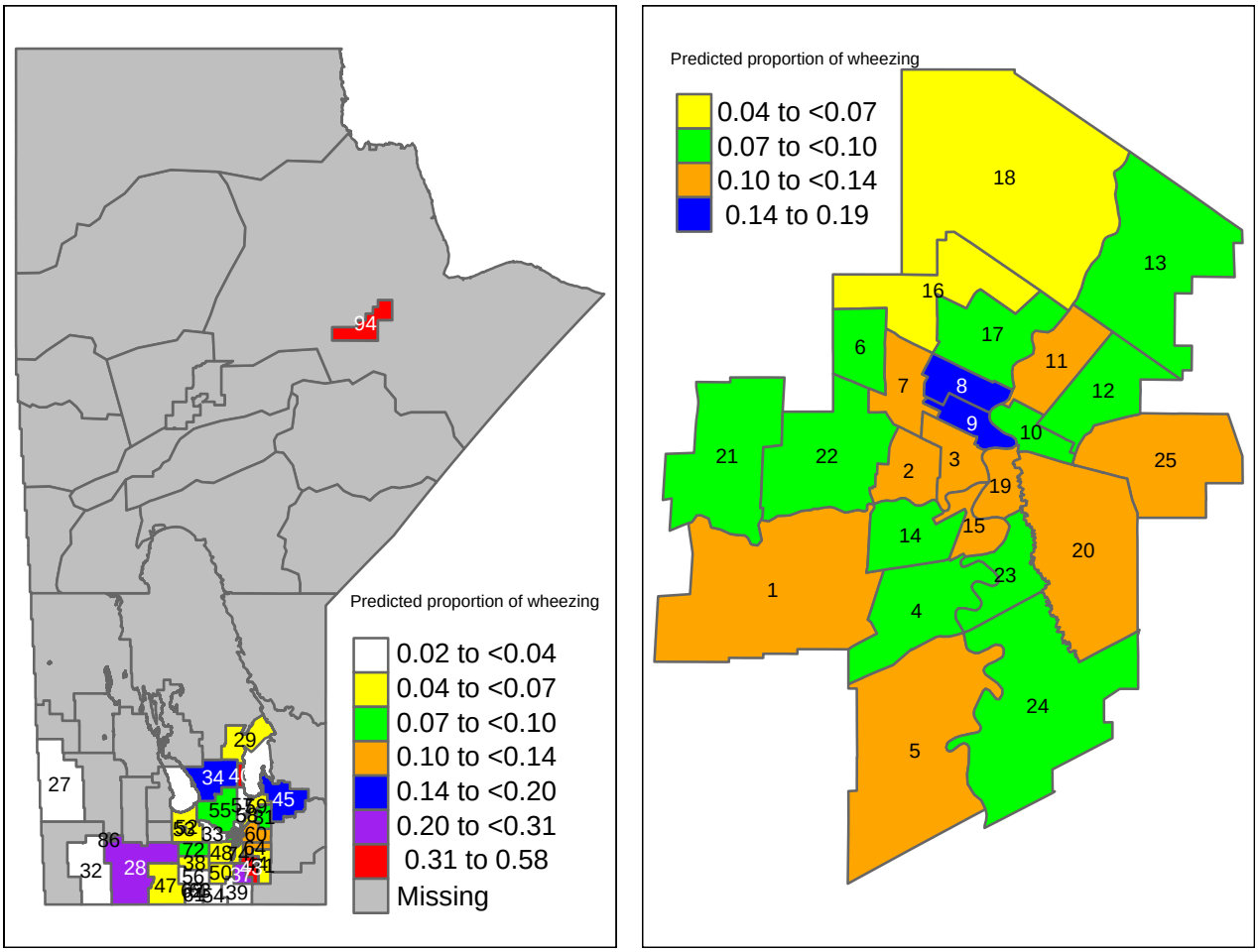


Figure 1. Predicted proportions of wheezing severity using the final area-level logistic model in Manitoba. Winnipeg-specific proportions are shown on the right panel (region codes are given in Table A2).

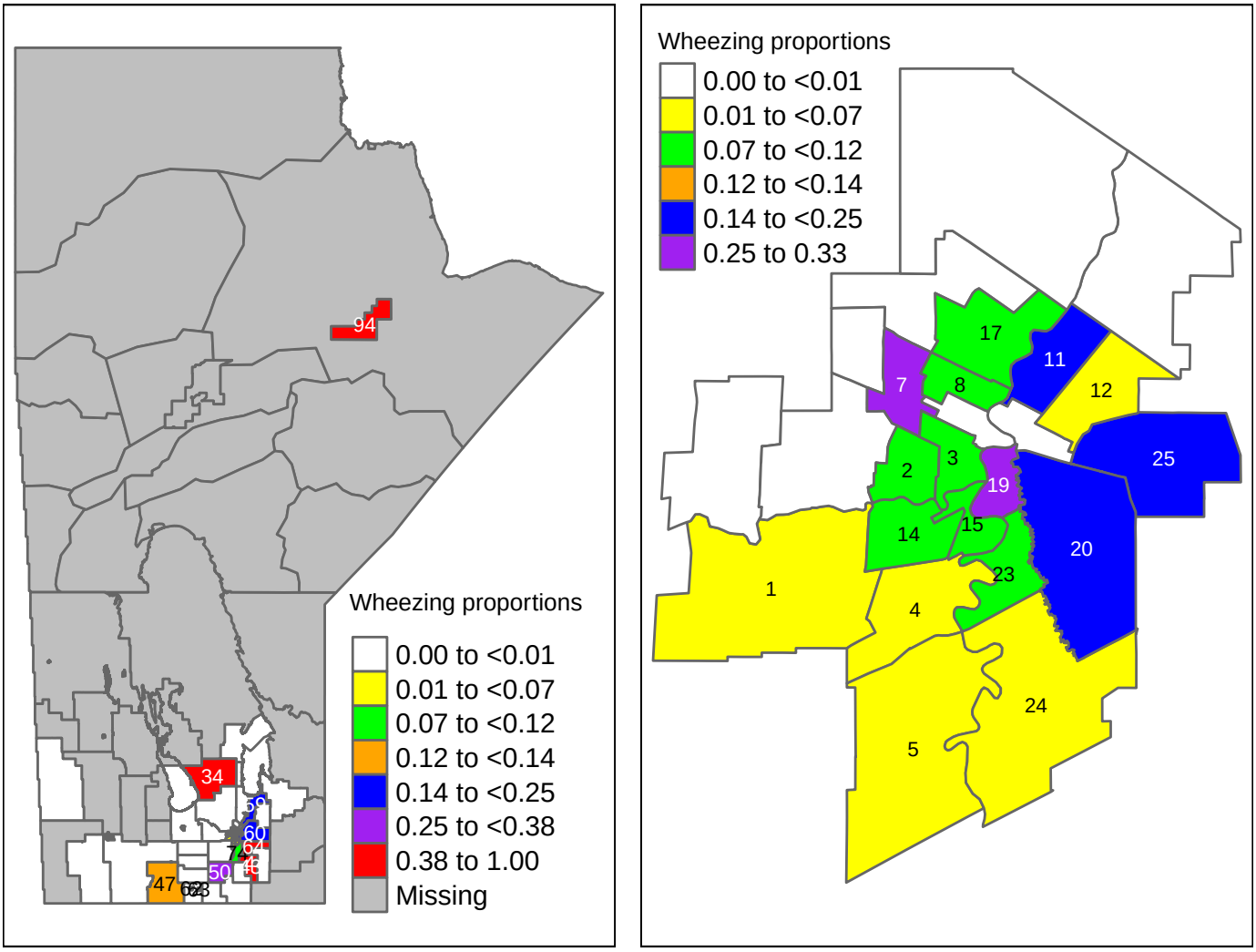


Figure 2. Raw proportion of wheezing severity obtained from the data in Manitoba (aggregated wheezing severity at 3 months divided by sample size in each area). Winnipeg-specific proportions are shown on the right panel (region codes are given in Table A2).

The results of the final area-level ZIP (after model building steps) are reported in Table 4. The area-level ZIP model used maternal asthma and farm as covariates for both the logit (excess zero) and Poisson component, and thus did not use a prior for the excess zero component of the model. The area-level ZIP model indicated that living within a 100 m from a farm decreased the expected rate of wheezing episodes by 3%, while prenatal maternal asthma increased the expected rate of wheezing episodes by 1%.

Table 4. Final area-level ZIP model.

Parameter	Estimate	95% Credible Interval (Lower, Upper)	Rate Ratio
Intercept	0.95	(−0.02, 1.90)	-
Farm	−0.03	(−0.60, 0.01)	0.97
Mother asthma (prenatal)	0.01	(−0.02, 0.03)	1.01
σ_v^2	1.17	(0.76, 1.75)	-
DIC	164.7	-	-

Using the results of Table 4, the predicted average number of wheezing episodes in each area k (=1, $\cdots$, K) was calculated using $\hat{\lambda}_k^{HB} = \exp\left(x_k^T \hat{\boldsymbol{\beta}} + \hat{v}_k\right)$. These predicted average wheezing episodes were then mapped to their corresponding RHADs in Manitoba (Figure 3). For comparison purposes, the area-level wheezing episodes at 3 months obtained from the raw data were also mapped (Figure 4). The average number of wheezing episodes obtained directly from the data indicated that many areas have no or a small number of wheezing episodes. In particular, regions 27, 28, and 45 had no wheezing episodes while the smoothed predictions using the proposed new ZIP model (Figure 3) had at least four episodes of wheezing in those areas.

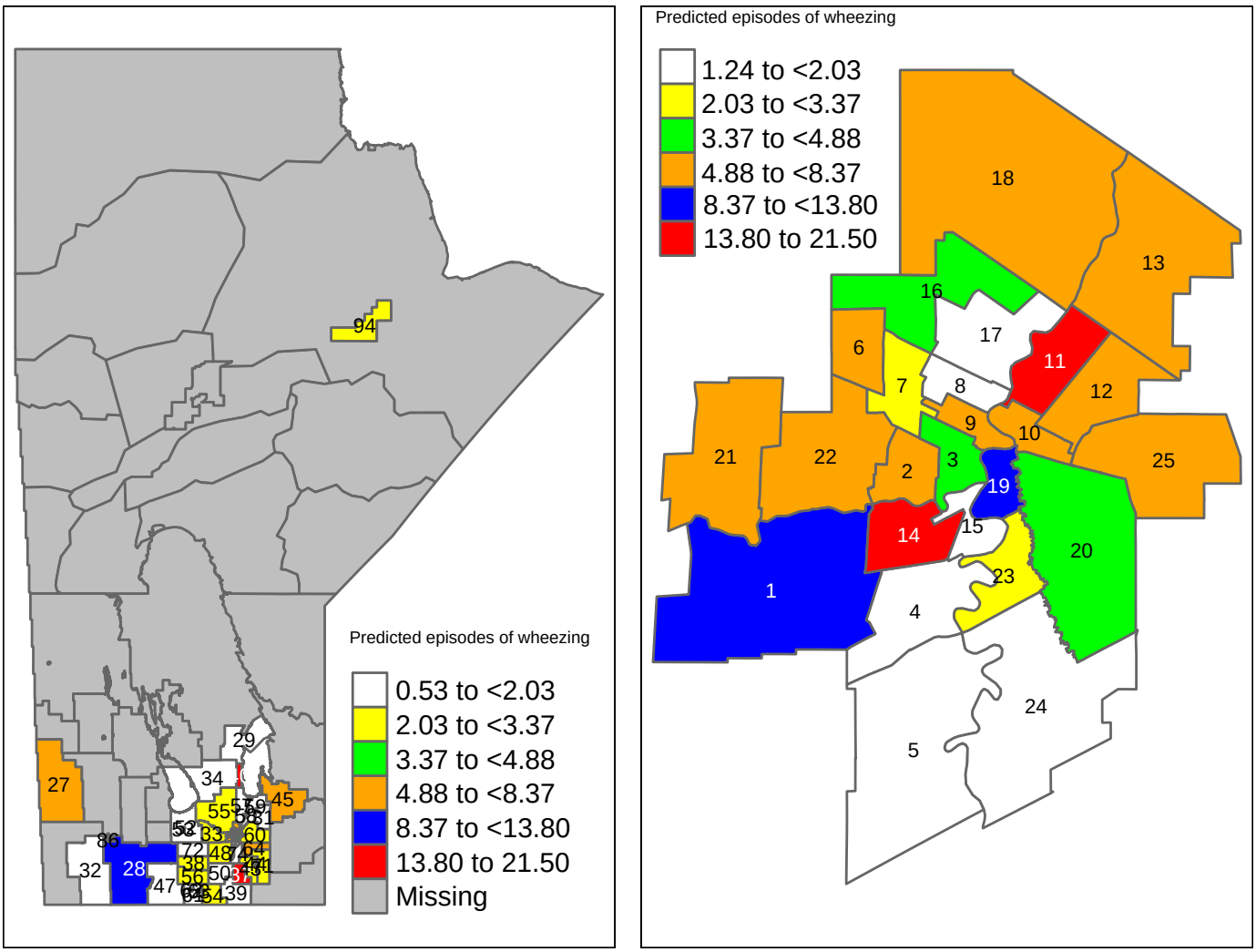


Figure 3. Predicted average number of wheezing episodes in Manitoba using the final area-level (new) ZIP model. Winnipeg-specific proportions are shown on the right panel (region codes are given in Table A2).

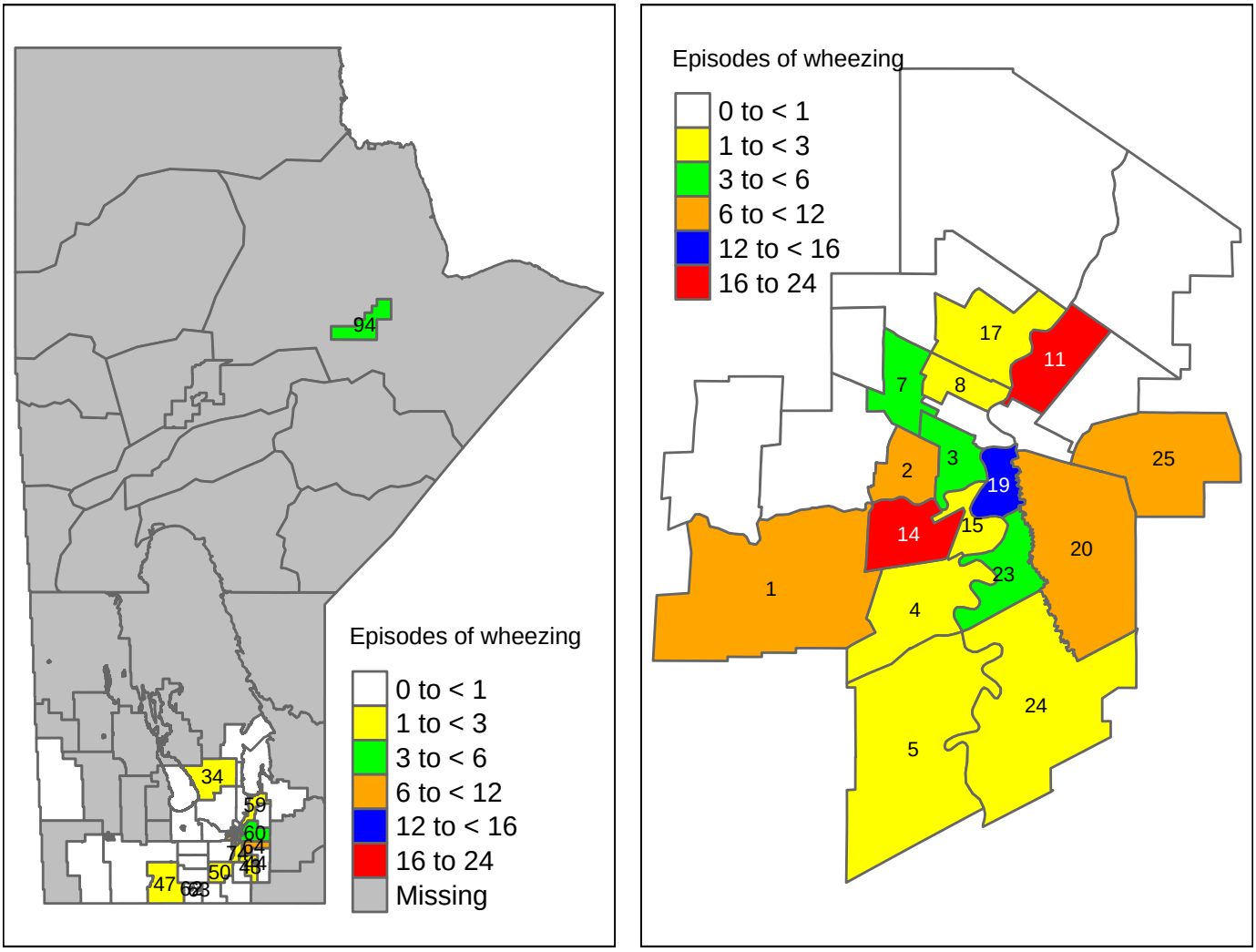


Figure 4. Raw number of wheezing episodes obtained from the data in Manitoba (aggregated wheezing episodes at 3 months). Winnipeg-specific proportions are shown on the right panel (region codes are given in Table A2).

5. Discussion

The use of advanced statistical modeling techniques, particularly small area estimation (SAE) models, has proven to be valuable in this study. The application of Bayesian methods to both area-level and unit-level data has allowed for a more nuanced understanding of childhood wheezing patterns across Manitoba.

The simulation studies revealed that the proposed models performed well in terms of bias and the mean squared error (MSE) of the model parameter estimates. One of the key strengths of this study is the rigorous model validation process, including the use of the Gelman–Rubin diagnostic for ensuring MCMC convergence and the DIC statistics for model comparison. These steps have ensured that the models used are both statistically sound and robust. The choice of the hierarchical Bayesian approach provided flexibility in modeling the complex structure of the data, accommodating both individual and area-level random effects. We have selected non-informative priors within the models to prioritize data-driven influence over subjective bias.

The inclusion of auxiliary variables, such as maternal health history and environmental factors, has enhanced the precision of the estimates. These variables have acted as significant predictors, explaining a substantial portion of the variation in wheezing

outcomes. The models demonstrated that maternal asthma and smoking are crucial risk factors, significantly increasing the odds and rates of wheezing in children.

Given the sparsity of the population in many regions of Manitoba, the use of SAE models was particularly appropriate. These models effectively drew strength from related areas, improving the estimates in regions with small sample sizes. This methodological approach is crucial for public health research in sparsely populated areas, ensuring that findings are reliable and actionable. As a practical guideline, we recommend using the proposed ordinal logistic model (Section 3.1) for longitudinal data with ordinal outcomes when the goal is to predict outcomes at the area level. Similarly, we suggest the proposed ZIP model (Section 3.3) for longitudinal count data with many zeros, with the same area-level prediction objective.

The implementation of the ZIP model addressed the issue of overdispersion and excess zeros in the count data of wheezing episodes. This model differentiated between children who never wheezed and those who did, providing a more accurate representation of the data. The ZIP model's findings regarding the influence of environmental factors, such as proximity to farms, have important implications for understanding the triggers of wheezing in susceptible populations.

Notably, we encountered complex computational challenges in analyzing longitudinal data, particularly with high-dimensional covariates/areas and relatively small sample sizes. The interaction between these factors posed significant obstacles to achieving robust and efficient analytical methods. To overcome these difficulties, we leveraged high-performance computing resources to address computational constraints and employed small area estimation techniques to mitigate issues related to sample size limitations.

6. Conclusions

The methodological rigor applied in this study highlights several key findings and their implications for public health policy:

The study underscores the critical role of maternal health, particularly asthma and smoking, in influencing wheezing outcomes in children. These findings highlight the need for targeted interventions aimed at improving maternal health as a strategy to reduce childhood wheezing and asthma.

The use of Bayesian SAE models has provided more accurate and reliable estimates of wheezing prevalence and severity across Manitoba. These models are particularly valuable for public health surveillance in regions with sparse populations, offering insights that would not be possible with traditional methods.

The study has revealed the dual role of environmental factors, such as farm proximity, in influencing wheezing outcomes. While farm living appears to reduce the odds of developing wheezing, it increases the frequency of episodes among those already affected. These nuanced findings can inform targeted public health interventions and policies.

The study highlights the potential for further research using more comprehensive and longitudinal data. Future studies could expand on this work by incorporating additional environmental and genetic factors, as well as exploring the long-term impacts of early childhood wheezing on asthma development.

These innovative techniques can also be applied across different provinces and regions to examine variations in asthma and wheezing over both geographic areas and time.

Author Contributions: Conceptualization, methodology, software, resources, validation, visualization, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, C.S.; Conceptualization, methodology, validation, supervision, writing—review and editing, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Health Research Ethics Board (HREB) of the University of Manitoba, with ethics number HS23359 (H2019:429). The approval was granted on 28 October 2024, and is valid until 15 November 2025. The project, titled "Modelling Severity of Wheezing in Small Areas in Manitoba Using Longitudinal Data", underwent annual review and was found to be ethically acceptable for research involving human participants.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The dataset used in this study cannot be shared publicly due to copyright and privacy concerns for participants involved in the CHILD Cohort Study. For more details on data access policies, please visit https://childstudy.ca/for-researchers/data-access/ (accessed on 25 March 2025). The study utilized individual-level participant data, which are classified as controlled access.

Acknowledgments: We are very grafeful to Elinor Simons and Zeinab Mashreghi for their help in guiding and focusing the clinical and statistical objectives of this paper throughout multiple revisions and rewrites. We also sincerely thank the three reviewers for their constructive and insightful comments, which have significantly enhanced the quality and clarity of this manuscript. We thank the CHILD Cohort Study (CHILD) participant families for their dedication and commitment to advancing health research. CHILD was initially funded by CIHR and AllerGen NCE. Visit CHILD at childstudy.ca.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Appendix A. Supplementary Tables

Table A1. Severity of wheezing frequency distribution for six follow-ups within the CHILD cohort. Scale of 0–3, 0 being no evidence of wheezing.

	Severity of Wheezing					
Severity	**3 Months**	**6 Months**	**1 Year**	**18 Months**	**2 Years**	**2.5 Years**
0	618	596	576	567	600	581
1	34	29	26	2	31	24
2	19	16	19	6	17	12
3	8	7	12	2	13	9
Missing	11	42	57	112	29	64

Table A2. Region code with associated location and small area number.

Code	Region	Small Area
W002	Assiniboine South	1
W11A	Downtown	2
W11B	Downtown	3
W03A	Ft. Garry	4
W03B	Ft. Garry	5
W09A	Inkster	6
W09B	Inkster	7
W10A	Point Douglas	8
W10B	Point Douglas	9
W07A	River East	10
W07B	River East	11
W07C	River East	12
W07D	River East	13

Table A2. *Cont.*

Code	Region	Small Area
W12A	River Heights	14
W12B	River Heights	15
W08A	Seven Oaks	16
W08B	Seven Oaks	17
W08C	Seven Oaks	18
W05A	St. Boniface	19
W05B	St. Boniface	20
W01A	St. James-Assiniboia	21
W01B	St. James-Assiniboia	22
W04A	St. Vital	23
W04B	St. Vital	24
W006	Transcona	25
WE34	Souris River	26
WE31	Asessippi	27
WE36	Spruce Woods	28
IE42	Arborg/Riverton	29
WE32	Little Saskatchewan	30
IE31	Beausejour	31
WE33	Turtle Mountain	32
SO14	Cartier/SFX	33
IE43	St. Laurent	34
WE13	Riding Mountain	35
WE15	Dauphin	36
SO25	St Pierre/DeSalaberry	37
SO22	Carman	38
SO26	Red River South	39
IE41	Gimli	40
WE35	Whitemud	41
WE11	Duck Mountain	42
SO45	Hanover	43
SO44	Steinbach	44
IE32	Pinawa/LDB	45
SO11	Seven Regions	46
SO31	Lorne/Louise/Pembina	47
SO23	MacDonald	48
WE12	Porcupine Mountain	49
SO24	Morris	50
SO12	MacGregor	51
SO13	Rural Portage	52
SO15	City of Portage la Prair	53
SO33	Altona	54
IE21	Stonewall/Teulon	55
SO36	Roland/Thompson	56
IE22	Wpg Beach/St. Andrews	57
IE11	Selkirk	58
IE23	St. Clements	59
IE24	Springfield	60
SO32	Stanley	61
SO34	Morden	62
SO35	Winkler	63
SO42	Tache	64
WP21	Winnipeg Churchill	65
IE52	Fisher/Peguis	66
NO11	Flin, Snow, Cran, Sher	67

Table A2. *Cont.*

Code	Region	Small Area
NO13	LL/MC, LR, O-P(SIL), PN(GVL)	68
IE51	Powerview/PF	69
NO21	GR/Mis, ML/Mos, Eas/Che	70
SO43	Ste Anne/LaBroquerie	71
SO21	Nortre Dame/St. Claude	72
NO14	Thomp, Myst Lake	73
SO41	Niverville/Richot	74
IE61	Northern Remote	75
SO46	Rural East	76
NO22	Puk/Mat Col CN	77
NO23	SayD(TL), Bro/BL, NoL(Lac)	78
NO26	Bu(OH), MS(GR), GLN/GLFN	79
NO28	Norway House/NH CN	80
NO27	Cross Lake/Pimi CN	81
NO31	Island Lake	82
IE33	Whiteshell	83
WE14	Agassiz Mountain	84
WE16	Swan River	85
WE23	Bdn Downtown	86
WE24	Bdn South End	87
WE22	Bdn North Hill	88
WE25	Bdn East End	89
WE21	Bdn West End	90
IE53	Eriksdale/Ashern	91
NO12	The Pas/OCN, Kels	92
NO15	Bay Line	93
NO16	Gillam, Fox Lake CN	94
NO25	Sham, YorkF, Tat(SPL)	95
NO24	Nelson House/NCN	96

Appendix B. Simulation Studies: Theory and Results

Appendix B.1. Ordinal Logistic Model

A simulation study was performed in which data were simulated with an ordinal response; the response represented the wheezing severity scale found on the CHILD study questionnaires. The dependent variable or response variable y_{ijk} was randomly generated with 1600 random values, all ranging from 1–3 using the integer uniform distribution or trinomial distribution. The probability of generating each integer $1, 2, 3$ was $\frac{1}{3}$; thus, equal probabilities. The aim of the simulation was to investigate the performance of the proposed ordinal logistic model with a similar set-up as the data structure from the CHILD cohort. The ordinal logistic model used for the simulation is described as follows. Let y_{ijk} have three levels $l = 1, 2, 3$; thus, there were two categories $c = 1, 2$ (with another level as a reference category). We then obtain

$$\log\left[\frac{\text{Prob}\left(y_{ijk} \leq c\right)}{\text{Prob}\left(y_{ijk} > c\right)}\right] = a_c + \beta_1 X_i + \beta_2 X_{ij} + \beta_3 X_{ijk} + f_{ij} + v_k + I_{ijk}, \tag{A1}$$

where i represents the individual, j is follow-up, and k is small area, and

$$
\begin{gathered}
v_k \sim \mathcal{N}(0, \sigma_v^2) \\
f_{ij} = \rho f_{ij-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_f^2\right) \\
f_{ij} \mid f_{ij-1} \sim \mathcal{N}\left(\rho f_{ij-1}, \sigma_f^2\right) \\
I_{ijk} \sim \mathcal{N}(0, \sigma_I^2) \\
-1 < \rho < 1.
\end{gathered}
$$

Thus, the two logit equations with $Q(m) = \text{Prob}\left(y_{ijk} \leq c\right)$ are

$$
\begin{aligned}
\text{logit}[\text{Q}(1)] &= a_1 + \beta_1 X_i + \beta_2 X_{ij} + \beta_3 X_{ijk} + f_{ij} + v_k + I_{ijk} \\
\text{logit}[\text{Q}(2)] &= a_2 + \beta_1 X_i + \beta_2 X_{ij} + \beta_3 X_{ijk} + f_{ij} + v_k + I_{ijk}.
\end{aligned}
$$

For the simulation, there were 20 individuals, 4 follow-ups, and 20 small areas. The covariates X_i, X_{ij}, and X_{ijk} were generated from a binary distribution with the probability of generating each integer $0, 1$ being equal to $\frac{1}{2}$. The simulation was conducted by fitting the simulated data with the ordinal logistic model using Rjags. Diffuse priors were used for all the model parameters. Coefficients from the fitted model were then to be used as true parameters, since a starting point was needed to begin the simulations.

These estimates were then used as the true parameters for the simulation study. The initial goal was to evaluate the performance of the model parameters. The simulation procedure was as follows:

Algorithm A

Step 1: Consider the estimates from Table A3 as the true parameters, represented by the vector δ

Step 2: Use the true parameters to generate the random effects, which, in turn, generate the new response values

Step 3: Refit the model using the new generated response values

Step 4: Extract the estimates from the newly fitted model

Step 5: Repeat the cycle by going to step 2

Table A3. Ordinal Simulation Study: Three chains are used with 60,000 samples from each chain. Furthermore, the Gelman diagnostics indicated convergence of model parameters. These estimates were used as the true parameters for the ordinal logistic model.

Parameter	Estimate
a_1	-0.685
a_2	0.772
β_1	-0.090
β_2	0.068
β_3	0.170
ρ	0.378
σ_v	0.145
σ_f	0.162
σ_I	0.169

In the above steps, the model was essentially reversed in the sense that the aim was to simulate values of the response y_{ijk} using the ordinal model described before and thus replacing the original values of y_{ijk} with the simulated values of y_{ijk} and then refitting the same ordinal model; this was performed 500 times. Each simulation run produced its own

estimates ($\hat{\delta}_r; r = 1, \cdots, 500$) of the model parameters. Then, the bias, relative bias (R. Bias), and MSE (Mean Square Error) were calculated using the three formulas below, assuming element-wise operation of the vectors. The bias and relative bias were used to quantify the average difference between the true parameters and the estimates within the simulations runs. The MSE was used to assess, on average, how far an estimator was from the true parameter within the simulation runs.

$$\text{Bias} = \frac{1}{500}\sum_{r=1}^{500}\left(\delta - \hat{\delta}_{\mathbf{r}}\right)$$

$$\text{Relative Bias}(\%) = \frac{1}{500}\sum_{r=1}^{500}\frac{\left(\delta - \hat{\delta}_{\mathbf{r}}\right)}{\delta} \times 100$$

$$\text{MSE} = \frac{1}{500}\sum_{r=1}^{500}\left(\delta - \hat{\delta}_{\mathbf{r}}\right)^2$$

The results from Table A4 indicated the model parameter estimates have very small bias and MSE. The relative bias for σ_v is higher than the other parameters, but overall it is acceptable. A possible explanation is that the number of small areas may not have been enough in this simulation set-up to reflect the performance of small area variation in the proposed model. Thus, the results indicate that the estimated parameters are working well in the proposed model.

Table A4. Bias, relative bias (R. Bias), and MSE of model parameter estimates based on 500 simulation runs from the proposed ordinal logistic model.

Parameter	True	Bias	R. Bias (%)	MSE
a_1	-0.685	-0.022	3.21	0.037
a_2	0.772	0.010	1.24	0.038
β_1	-0.090	0.003	-3.33	0.011
β_2	0.068	-0.002	-2.94	0.009
β_3	0.170	-0.010	-5.88	0.009
ρ	0.378	-0.019	-4.95	0.015
σ_v	0.145	0.028	19.24	0.015
σ_f	0.162	0.011	6.48	0.001
σ_I	0.169	-0.006	-3.55	0.001

Further simulation studies were performed, which assumed the responses are binary and count; thus, logit and log link functions were used.

Appendix B.2. Binary Logistic Model

The model is described as

$$y_{ijk} \mid p_{ijk} \sim \text{Bernoulli}\left(p_{ijk}\right),$$

where y_{ijk} is the response of wheezing for individual i, follow-up j, and small area k. Then, $p_{ijk} = \text{Prob}(y_{ijk} = 1)$ is the probability that the individual i had a wheezing event at follow-up j, and in small area k. Now the logistic model is described as

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \beta_0 + \beta_1 X_i + \beta_2 X_{ij} + \beta_3 X_{ijk} + f_{ij} + v_k, \tag{A2}$$

where

$$v_k \sim \mathcal{N}(0, \sigma_v^2)$$
$$f_{ij} = \rho f_{ij-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_f^2\right)$$
$$f_{ij} \mid f_{ij-1} \sim \mathcal{N}\left(\rho f_{ij-1}, \sigma_f^2\right)$$
$$-1 < \rho < 1.$$

Now, f_{ij} is the random effect of the individual and follow-up. Usually, this is the time-specific disturbance, but in the case of this model, the disturbance is an autoregressive-one or AR(1) model and v_k is simply the random effect of each small area. A simulation study was performed in which data were simulated with a binary response; the response represented whether each child within the cohort wheezed or not. The aim of the simulation was to verify the estimates produced by a binary logistic model with data representing the CHILD cohort. For the simulation, there were 20 individuals, 4 follow-ups, and 20 small areas. The dependent variable y_{ijk} was randomly generated with 1600 random binary values from a Bernoulli distribution. The probability of generating each integer $0, 1$ was $\frac{1}{2}$; thus, equal probabilities. The covariates X_i,X_{ij}, and X_{ijk} were also generated from a binary distribution with the probability of generating each integer $0, 1$ being equal to $\frac{1}{2}$. The estimated parameters are reported in Table A5.

Table A5. Binary Logistic Simulation Study: Three chains were used, with 60,000 samples from each chain. Furthermore, the Gelman diagnostics indicated convergence of model parameters. These estimates were used as the true parameters for the binary logistic model.

Parameter	Estimate
β_0	-0.298
β_1	-0.118
β_2	0.270
β_3	0.0168
ρ	0.0117
σ_v	0.111
σ_f	0.128

The goal was to evaluate the performance of the model parameters in the proposed logistic model. The simulation procedure followed from Algorithm A. In the simulation steps, the model was essentially reversed in the sense that the aim was to simulate the values of the response y_{ijk} using the binary logistic model described before, thus replacing the original values of y_{ijk} with the simulated values of y_{ijk} and then refitting the same binary logistic model; this was performed 500 times. Each simulation run produced its own estimates $(\hat{\delta}_r; r = 1, \cdots, 500)$ of the model parameters. Then, the bias, relative bias (R. Bias), and MSE were calculated (Table A6).

The results from Table A6 indicated the model parameter estimates had small bias and MSE. The relative bias for ρ was higher than the other parameters, but it was still acceptable. Note that the model is sensitive to the number of follow-ups in the dataset. Ultimately, the number of areas, follow-ups and individuals chosen show how well the simulation performed. Thus, the overall results indicated that the estimated parameters were working well in the proposed binary logistic model.

Table A6. Bias, relative bias (R. Bias), and MSE of model parameter estimates based on 500 simulation runs from the proposed binary logistic model.

Parameter	True	Bias	R. Bias (%)	MSE
β_0	-0.298	0.009	-3.00	<0.0001
β_1	-0.118	-0.002	-1.80	<0.0001
β_2	0.270	-0.001	-0.41	<0.0001
β_3	0.0168	-0.0002	-8.40	<0.0001
ρ	0.0117	0.0025	22.70	<0.0001
σ_v	0.111	0.0005	0.47	<0.0001
σ_f	0.128	0.0006	0.62	<0.0001

Appendix B.3. Zero-Inflated Poisson Model

We also simulated the Zero-Inflated Poisson (ZIP) model, in which the response was the number of wheezing episodes for each child in the cohort. We considered the same covariates as discussed in the previous logistic model (Appendix B.2). For this simulation, we were less interested in the modeling of false zeros; therefore, a diffuse prior for π_{ijk} was implemented such that $logit(\pi_{ijk}) = logit(\pi) = \gamma_0$, $\gamma_0 \sim \mathcal{N}(0, 10,000)$. Again, for the simulation, there were 20 individuals, 4 follow-ups, and 20 small areas. The dependent variable y_{ijk} was randomly generated with 1600 random values from the ZIP distribution with $\pi = 0.4$ and $\lambda = 4$, which give the range of possible values for y_{ijk} as between 0 and 13. The covariates X_i, X_{ij}, and X_{ijk} were also generated from a binary distribution with the probability of generating each integer $0, 1$ being equal to $\frac{1}{2}$. The statistical measures used to test the credibility of the models were the bias and MSE. Now, let

$$\Theta_{ijk} = \beta_0 + \beta_1 X_i + \beta_2 X_{ij} + \beta_3 X_{ijk} + f_{ij} + v_k \tag{A3}$$

where

$$\begin{gathered} v_k \sim \mathcal{N}\left(0, \sigma_{\mathrm{v}}^2\right) \\ f_{ij} = \rho f_{ij-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_{\mathrm{f}}^2\right) \\ f_{ij} \mid f_{ij-1} \sim \mathcal{N}\left(\rho f_{ij-1}, \sigma_{\mathrm{f}}^2\right) \\ -1 < \rho < 1. \end{gathered}$$

Then, the ZIP model becomes,

$$\operatorname{Prob}\left(y_{ijk} = m\right) = \begin{cases} \pi + (1-\pi)\exp\left(-\lambda_{ijk}\right) & \text{if } m = 0 \\ (1-\pi)\frac{\lambda_{ijk}^{y_{ijk}}\exp\left(-\lambda_{ijk}\right)}{y_{ijk}!} & \text{if } m > 0 \end{cases} \tag{A4}$$

such that $\log(\lambda_{ijk}) = \Theta_{ijk}$ and $\operatorname{logit}(\pi) = \gamma_0$ where $\gamma_0 \sim \mathcal{N}(0, 10,000)$.

The model parameter estimates are presented in Table A7. Then, the simulation procedure follows Algorithm A. Similar to previous simulations, the model was essentially reversed. Each simulation run produced its own estimates $(\hat{\delta}_r; r = 1, \cdots, 500)$ of the model parameters. Then, the bias, relative bias (R. Bias), and MSE were calculated (Table A8).

The results from Table A8 indicate that the model parameter estimates behaved reasonably well; however, some parameters had relative bias above 30 percent, which, after further investigation, was due to ρ not being statistically significant for this particular simulated dataset. When using such a model with real data, particular attention should be placed on the AR(1) and this should be removed if not significant.

Table A7. ZIP Simulation Study: Three chains are used with 60,000 samples from each chain. Furthermore, the Gelman diagnostics indicated convergence of model parameters. These estimates are to be used as the true parameters for the ZIP model.

Parameter	Estimate
β_0	1.085
β_1	−0.034
β_2	0.066
β_3	0.085
γ_0	−0.304
ρ	0.657
$1-\pi$	0.575
σ_v	0.639
σ_f	0.042

Table A8. Bias, relative bias (R. Bias), and MSE of model parameter estimates based on 500 simulation runs from the proposed ZIP model.

Parameter	True	Bias	R. Bias (%)	MSE
β_0	1.085	0.270	24.9	0.252
β_1	−0.034	−0.003	8.40	<0.0001
β_2	0.066	0.002	3.30	<0.0001
β_3	0.085	-0.003	−3.50	<0.0001
γ_0	−0.304	<0.0001	0.10	<0.0001
ρ	0.657	−0.224	−34.10	0.194
$1-\pi$	0.575	<0.0001	<0.0001	<0.0001
σ_v	0.639	−0.210	−34.00	0.355
σ_f	0.042	<0.0001	0.30	<0.0001

References

1. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention. *Eur. Respir. J.* **2008**, *31*, 143–178. [CrossRef] [PubMed]
2. Public Health Agency of Canada. *Report from the Canadian Chronic Disease Surveillance System: Asthma and Chronic Obstructive Pulmonary Disease (COPD) in Canada, 2018*; Public Health Agency of Canada: Ottawa, ON, Canada, 2018.
3. Statistics Canada. Table 13-10-0096-08 Asthma, by Age Group, 2022. [CrossRef]
4. Nahum, U.; Gorlanova, O.; Decrue, F.; Oller, H.; Delgado-Eckert, E.; Böck, A.; Schulzke, S.; Latzin, P.; Schaub, B.; Karvonen, A.M.; et al. Symptom trajectories in infancy for the prediction of subsequent wheeze and asthma in the BILD and PASTURE cohorts: A dynamic network analysis. *Lancet Digit. Health* **2024**, *6*, e718–e728. [CrossRef] [PubMed]
5. Rao, J.; Molina, I. *Small Area Estimation*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2015. [CrossRef]
6. Richelle, J. Wheezing in Early Life and Validating the Asthma Predictive Index. Bachelor Thesis, University of Manitoba, Winnipeg, MB, Canada, 2014.
7. Jiang, J.; Torabi, M. Sumca: Simple, unified, Monte-Carlo-assisted approach to second-order unbiased mean-squared prediction error estimation. *J. R. Stat. Soc. Ser. B* **2020**, *82*, 467–485. [CrossRef]
8. Torabi, M.; Lele, S.R.; Prasad, N.G. Likelihood inference for small area estimation using data cloning. *Comput. Stat. Data Anal.* **2015**, *89*, 158–171. [CrossRef]
9. Battese, G.E.; Harter, R.M.; Fuller, W.A. An error-components model for prediction of county crop areas using survey and satellite data. *J. Am. Stat. Assoc.* **1988**, *83*, 28–36. [CrossRef]
10. Fay, R.E.; Herriot, R.A. Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *J. Am. Stat. Assoc.* **1979**, *74*, 269–277. [CrossRef]

11. Alkan, N. Assessing convergence diagnostic tests for Bayesian Cox regression. *Commun. Stat. Simul. Comput.* **2017**, *46*, 3201–3212. [CrossRef]
12. Chan, J.C.C.; Grant, A.L. On the Observed-Data Deviance Information Criterion for Volatility Modeling. *J. Financ. Econom.* **2016**, *14*, 772–802. [CrossRef]
13. Torabi, M.; Shokoohi, F. Likelihood inference in small area estimation by combining time-series and cross-sectional data. *J. Multivar. Anal.* **2012**, *111*, 213–221. [CrossRef]
14. Torabi, M.; Shokoohi, F. Hierarchical Bayes estimation in small area estimation using cross-sectional and time-series data. *J. Stat. Comput. Simul.* **2014**, *84*, 605–613. [CrossRef]
15. Shokoohi, F.; Torabi, M. Semi-parametric small-area estimation by combining time-series and cross-sectional data methods. *Aust. N. Z. J. Stat.* **2018**, *60*, 323–342. [CrossRef]
16. Torabi, M. Zero-inflated spatio-temporal models for disease mapping. *Biom. J.* **2017**, *59*, 430–444. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.