UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 4 Nomor 3 Tahun 2026 Halaman 346-356.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Hierarchical Bayesian Modeling with IAR Hexagonal Grids for Reconstructing Incomplete OD Matrices

Eko Primadi Hendri[1*], Sarah Fadhlia[2], Edi Santosa[3], Rachmat Sadili[4], Sudirman Anggada[5]
[1,3,4,5]Department of Road Transportation Management, Politeknik Transportasi Darat Indonesia - STTD, Bekasi, Indonesia
[2]Department of Guidance and Counseling, Universitas Indraprasta PGRI, Jakarta, Indonesia
*Corresponding author: eko.primadi@ptdisttd.ac.id

Submitted : 11 Mei 2026
Revised : 13 Agustus 2026
Accepted : 31 Agustus 2026

ABSTRACT

Origin-Destination (OD) matrix derivation in open Bus Rapid Transit systems such as Transjakarta is fundamental for urban mobility studies. However, this process is often complicated by insufficient observations, particularly missing tap-out data, leading to sparsity, zero-inflation, and overdispersion in the resulting matrices. This paper considers the probabilistic reconstruction of sparse OD matrices with spatial dependencies. As previous imputation methods have limitations, such as ignoring network topology or being affected by the Modifiable Areal Unit Problem (MAUP), this research proposes a hierarchical Bayesian approach incorporating an Intrinsic Autoregressive (IAR) prior within an isotropic hexagonal (H3) tessellation framework. A Negative Binomial distribution describes overdispersed count data and latent spatial intensities, while missing destinations are jointly estimated using Markov Chain Monte Carlo (MCMC). The proposed Spatial IAR model achieves stable convergence with a maximum R-hat value of 1.015, while the independent non-spatial model also achieves satisfactory convergence with a maximum R-hat value of 1.012. The Spatial IAR model achieves lower WAIC and LOOIC values (12635.11 and 12635.77, respectively) than the independent model (12658.30 for both), supporting the inclusion of spatial dependence. Prior sensitivity analysis shows that the main results remain relatively stable across alternative exponential prior specifications. Posterior Predictive Checks confirm that the spatial model reproduces the overdispersion and zero-inflation characteristics of the observed data. Overall, spatial regularization provides a framework for reconstructing sparse transportation data. As the analysis uses a synthetic transportation transaction dataset, the findings should be interpreted as a methodological evaluation rather than empirical evidence of Transjakarta passenger mobility patterns.

Keywords: *Hierarchical Bayesian; Incomplete Data; Intrinsic Autoregressive; Origin - Destination Matrix.*

This is an open access article under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2022 by author and Universitas Negeri Padang.

I. INTRODUCTION

In spatial statistics and big data modeling, mobility traces from smart card-based Automated Fare Collection (AFC) systems form high-dimensional spatial interaction networks that are commonly represented as Origin-Destination (OD) matrices. These matrices offer a straightforward foundation to study urban mobility patterns and are critical for applications including travel demand modelling, fleet allocation, and transportation planning, especially in large-scale systems like Transjakarta Bus Rapid Transit (BRT) (Gordon et al., 2018; Shalit et al., 2023).

However, the extraction of OD matrices in open BRT systems is usually hampered by inadequate observations, notably the missing tap-out data. This leads to large missing values and excessive sparsity when the research region is discretised in fine spatial units. Consequently, the OD matrix is characterised by many zero values and overdispersion (variance higher than the mean), making statistical modelling challenging (Zhuang et al., 2022).

Previous research have tried to handle this problem using listwise deletion, independent regression models, machine learning techniques or traditional spatial econometrics. Nevertheless, these approaches either overlook spatial dependencies, do not provide measures of uncertainty, or use spatial units such as square grids, which are susceptible to directional bias and the Modifiable Areal Unit Problem (MAUP) effects (Comber et al., 2022; Ni & Zhang, 2024; Tessema et al., 2023). Models without appropriate spatial regularisation tend to overfit, become unstable and have inaccurate inference (Blume et al., 2022; Donegan, 2022).

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

To tackle these challenges, this paper provides a spatial hierarchical Bayesian framework for sparse and imperfect OD matrix reconstruction. Missing destinations are modelled as latent variables and are estimated simultaneously with model parameters. To deal with overdispersed count data, we use a Negative Binomial distribution and model spatial dependence with an Intrinsic Autoregressive (IAR) prior over an isotropic hexagonal (H3) tessellation that minimises directional bias and facilitates sharing of information across neighbouring zones (Banerjee et al., 2014; Brodsky, 2018).

This research develops a framework for probabilistically reconstructing a sparse OD matrix using a synthetic transportation transaction dataset while incorporating spatial dependencies. The paper's main contributions are: (1) a hierarchical Bayesian model that integrates IAR and H3 to jointly address sparsity and missing data, (2) demonstration of the importance of spatial regularisation in stable MCMC inference, and (3) a reconstructed OD matrix with quantified uncertainty that demonstrates the methodological potential of the proposed framework for transportation analysis

II. MATERIAL AND METHODS

A. Data and Problem Characteristics

The dataset in this research is a synthetic Transjakarta Transportation Transaction dataset from Kaggle (Dikisahkan, 2023) which contains the coordinates of boarding (tap-in) and alighting (tap-out) of Bus Rapid Transit (BRT) passengers (Blume et al., 2022). The dataset was generated using the Faker library to provide a synthetic representation of passenger transactions (Dikisahkan, 2023). The use of this dataset provides a controlled setting for evaluating the proposed Bayesian modelling framework and its inferential performance through a structured Bayesian workflow (Betancourt, 2020; Gabry et al., 2019; Vehtari et al., 2021). The analysis is challenged by two problems: (1) the very sparse OD matrix owing to geographic aggregation and (2) missing destination coordinates contained in the synthetic dataset. Out of 25971 trips, 933 (3.59%) do not have destination coordinates.

Because the missing destination records are part of the published synthetic dataset (Dikisahkan, 2023), this study does not explicitly model the underlying missing-data mechanism (e.g., MCAR, MAR, or MNAR). Instead, the incomplete destination records are treated as latent variables within the proposed Bayesian hierarchical framework, allowing the Origin–Destination (OD) matrix to be reconstructed without excluding incomplete trips (Gelman et al., 2013). The objective of this study is to evaluate the proposed framework using the available synthetic dataset rather than to investigate the effects of specific missing-data mechanisms. Accordingly, the proposed framework is evaluated under the missing-data structure inherent in the published dataset, while the assessment of alternative missing-data mechanisms is considered outside the scope of this study (Little & Rubin, 2019).

B. Spatial Representation using Hexagonal Grid Systems

Spatial interaction analysis using continuous geographic coordinates has to be discretised to minimise computing complexity and sparsity in the Origin-Destination (OD) matrix. In this work we use the H3 hierarchical hexagonal tessellation (Brodsky, 2018), Its isotropic qualities include equal centroid distances and 6 equidistant neighbours, which leads to a stable neighbourhood structure and removes directional bias in a square grid. We define a spatial graph based on cell adjacency that facilitates modelling spatial autocorrelation in Bayesian inference, where let H be the set of hexagonal cells, each trip with coordinates (x, y) is mapped into a cell via an aggregating function f.

The hexagon size is defined at H3 resolution 8 (≈0.73 km² per cell), which corresponds to a typical 400-800 metre pedestrian catchment area (El-Geneidy et al., 2014; Ibraeva et al., 2020). This resolution was chosen to maintain sufficient spatial detail while avoiding excessive sparsity in the OD matrix. Using a finer resolution, such as H3 resolution 9, would divide the study area into a larger number of smaller cells, increasing the number of zero-flow OD pairs and the computational burden. In contrast, a coarser resolution, such as H3 resolution 7, would reduce sparsity but would also merge neighbouring travel patterns into larger spatial units, resulting in a loss of local spatial detail (Comber et al., 2022; Tessema et al., 2023). Therefore, H3 resolution 8 was considered an appropriate balance between spatial detail, data sparsity, and computational efficiency for the characteristics of the study area. Although hexagonal tessellation helps reduce directional bias and mitigates some effects of the Modifiable Areal Unit Problem (MAUP), it does not eliminate MAUP entirely because the results remain influenced by the spatial resolution selected for aggregation (Comber et al., 2022; Tessema et al., 2023).

C. Origin-Destination Matrix Construction and Incomplete Data Representation

After geographical discretisation, transaction data are aggregated into a hexagon-based Origin-Destination (OD) matrix Y, where each element y_{ij} denotes the number of journeys from origin i to destination j, with observed values

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

indicated as y_{ij}^{obs} (Gordon et al., 2018). Trips with unknown destination owing to tap-out anomalies are not deleted but modelled as latent variables (Blume et al., 2022). Let m_i be the number of missed trips from origin i. These are modelled as a multinomial distribution:

$$y_i^{miss} \sim \text{Multinomial}(m_i, \theta_i) \tag{1}$$

where θ_{ij} represents the probability of traveling from i to j with $\sum_j \theta_{ij} = 1$. These probabilities are linked to the latent trip intensity λ_{ij} through

$$\theta_{ij} = \frac{\lambda_{ij}}{\sum_k \lambda_{ik}} \tag{2}$$

where λ_{ij}denotes the expected trip intensity. The total flow is then defined as

$$y_{ij}^{total} = y_{ij}^{obs} + y_{ij}^{miss} \tag{3}$$

This framework enables joint estimation of missing trips during posterior inference, guaranteeing adequate uncertainty propagation and leading to a more robust OD matrix, in line with recent Bayesian workflow principles (Betancourt, 2020; Gabry et al., 2019; Gelman et al., 2020).

D. Hierarchical Statistical Model Formulation

The model is formulated in a two-level hierarchical Bayesian framework. The first level models total trip flows y_{ij} using a Negative Binomial to accommodate overdispersion in the count data:

$$y_{ij} \sim \text{NegativeBinomial}(\lambda_{ij}, \phi) \tag{4}$$

where $\lambda_{ij} = \mathbb{E}[y_{ij}]$ and the variance is defined as

$$\text{Var}(y_{ij}) = \lambda_{ij} + \frac{\lambda_{ij}^2}{\phi} \tag{5}$$

At the second level, missing trips are allocated using a multinomial structure, while the latent intensity is modeled through a log-link function:

$$\log(\lambda_{ij}) = \beta_0 + u_i + v_j \tag{6}$$

where β_0 represents the global baseline, u_i and v_j denote origin and destination effects. respectively. Both spatial effects are assigned Intrinsic Autoregressive (IAR) priors to capture spatial dependence among neighbouring hexagonal cells (Krisztin & Piribauer, 2021; LeSage & Pace, 2008). The corresponding spatial standard deviation parameters, σ_u and σ_v, are assigned weakly informative exponential priors to regularise the magnitude of the spatial random effects while allowing sufficient flexibility for the observed data to influence the posterior estimates. To ensure identifiability and avoid confounding during MCMC estimation, sum-to-zero constraints are imposed:

$$\sum_i u_i = 0 \text{ and } \sum_j v_j = 0 \tag{7}$$

so that spatial effects are interpreted as deviations from the global average, ensuring stable and consistent parameter estimation. The implementation of the IAR priors, exponential prior specification for the spatial standard deviation parameters, and the corresponding posterior inference in Stan are described in Section E.

E. Hexagonal Grid-Based Bayesian Prior Specification

The hexagonal neighborhood structure is represented by an adjacency matrix A, where $a_{ij} = 1$ if cells i and j are adjacent and 0 otherwise, with node degree $d_i = \sum_j a_{ij}$ forming a diagonal matrix D. The spatial dependence is defined through the intrinsic precision matrix (spatial Laplacian) $Q = D - A$. To ensure numerical stability and a valid posterior distribution, sum-to-zero constraints are imposed on the spatial effects, maintaining Q as positive semi-definite (Donegan, 2022). Spatial heterogeneity at origin and destination levels is modeled using Intrinsic Autoregressive (IAR) priors:

$$u_i \mid u_{-i}, \tau_u \sim N\left(\frac{\sum_{j\sim i} u_j}{d_i}, \frac{1}{\tau_u d_i}\right) \text{ and } v_j \mid v_{-j}, \tau_v \sim N\left(\frac{\sum_{k\sim j} v_k}{d_j}, \frac{1}{\tau_v d_j}\right) \tag{8}$$

where $\tau_u$and τ_v control the strength of spatial smoothing. This specification is consistent with Tobler's First Law of Geography where neighbouring zones may communicate information, and the isotropic hexagonal structure guarantees that spatial smoothing is unbiased in different directions (Banerjee et al., 2014; Brodsky, 2018; Comber et al., 2022). The spatial standard deviation parameters, σ_u and σ_v, are assigned weakly informative exponential priors, $\sigma_u \sim$

©Department of Statistics, Universitas Negeri Padang
DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Exponential(λ) and $\sigma_v \sim$ Exponential(λ), where λ is the exponential rate parameter controlling the degree of shrinkage applied to the spatial random effects.

Since the intrinsic precision matrix $Q$is singular, the IAR prior is intrinsically improper and requires an identifiability constraint for posterior inference (Banerjee et al., 2014; Morris et al., 2019). To address this issue, the Stan model defines the origin and destination spatial effects through unconstrained latent variables, denoted by u^{raw} and v^{raw}. These latent variables are centred around their sample means and scaled by their respective standard deviations to obtain the spatial effects used in the model, $u_i = \sigma_u(u_i^{raw} - \bar{u}^{raw})$ and $v_j = \sigma_v\left(v_j^{raw} - \bar{v}^{raw}\right)$, where $\bar{u}^{raw} = \frac{1}{N}\sum_{i=1}^{N} u_i^{raw}$ and $\bar{v}^{raw} = \frac{1}{N}\sum_{j=1}^{N} v_j^{raw}$. This parameterization naturally satisfies the sum-to-zero constraint, $\sum_{i=1}^{N} u_i = 0$ and $\sum_{j=1}^{N} v_j = 0$, which resolves the identifiability issue while preserving the spatial dependence among neighbouring hexagonal cells. In addition, weak soft sum-to-zero priors, $\sum_{i=1}^{N} u_i^{raw} \sim \mathcal{N}(0,\ 0.001N)$ and $\sum_{j=1}^{N} v_j^{raw} \sim \mathcal{N}(0,\ 0.001N)$, were included to improve numerical stability during HMC–NUTS sampling (Donegan, 2022; Morris et al., 2019).

F. Posterior Inference and Data Augmentation

The joint posterior distribution is defined as

$$p(\Theta, Y^{miss} \mid Y^{obs}) \propto p(Y^{obs}, Y^{miss} \mid \lambda, \phi)\, p(u \mid \tau_u)\, p(v \mid \tau_v)\, p(\Omega) \tag{9}$$

where Θ indicates the collection of latent parameters and $p(\Omega)$ represents the prior distributions for the model hyperparameters, including the exponential priors assigned to the spatial standard deviation parameters (σ_u and σ_v) described in Section E, together with the remaining model prior. The zero-sum parameterization of the spatial effects described in the previous section ensures posterior identifiability despite the singular intrinsic precision matrix, while preserving the underlying IAR spatial dependence. Consequently, posterior inference can be performed efficiently without directly sampling from the singular precision matrix. Inference is performed using Hamiltonian Monte Carlo (HMC) with the No-U-Turn Sampler (NUTS) which efficiently explores high-dimensional parameter spaces and estimates model parameters and missing data jointly in a unified Bayesian framework (Betancourt, 2020; Gelman et al., 2020; Vehtari et al., 2021).

G. Identification of Spatial Hotspots and Coldspots

In addition to rebuilding the missing data, the estimated origin spatial effects (u_i) are also utilised to identify the mobility clusters within the Transjakarta network. We use the mean ± one standard deviation criterion based on the posterior mean of the estimated origin spatial effects, following threshold-based approaches for hotspot identification (Fiechter et al., 2020; Sussman et al., 2019). Let $\bar{u}_i$denote the posterior mean of the origin spatial effect for zone i, with overall mean $\mu_u$and standard deviation σ_u. The upper and lower thresholds are defined as $T_{\text{upper}} = \mu_u + \sigma_u$ and $T_{\text{lower}} = \mu_u - \sigma_u$, respectively. If $\bar{u}_i \geq T_{\text{upper}}$, then the zone is termed a hotspot; if $\bar{u}_i \leq T_{\text{lower}}$, it is considered a coldspot; otherwise, it is regarded as neither a hotspot nor a coldspot. This method makes it simple to identify and visualise areas with relatively high and low estimated spatial effects.

H. Model Evaluation and Benchmarking

Model validity is evaluated using the Rconvergence diagnostic with a threshold of $\hat{R} < 1.05$ to ensure MCMC convergence (Vehtari et al., 2021), followed by Posterior Predictive Checking (PPC) using

$$p(y^{rep} \mid y) = \int p(y^{rep} \mid \theta)\, p(\theta \mid y)\, d\theta \tag{10}$$

to assess goodness-of-fit (Gelman et al., 2020). Predictive performance is further evaluated using WAIC and LOO-CV, defined as

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}) \text{ and LOOIC} = -2(\text{lppd}_{loo} - p_{loo}) \tag{11}$$

where lppd evaluates predictive accuracy with penalties for complexity. Differences in expected log predictive density ($\Delta elpd$) and its standard error are used to compare models. The proposed spatial model with an IAR prior is benchmarked against an independent Negative Binomial model without spatial dependence, to isolate the contribution of the hexagonal structure, and evaluate its role in reducing overfitting and improving inference stability (Donegan, 2022; Krisztin & Piribauer, 2021).

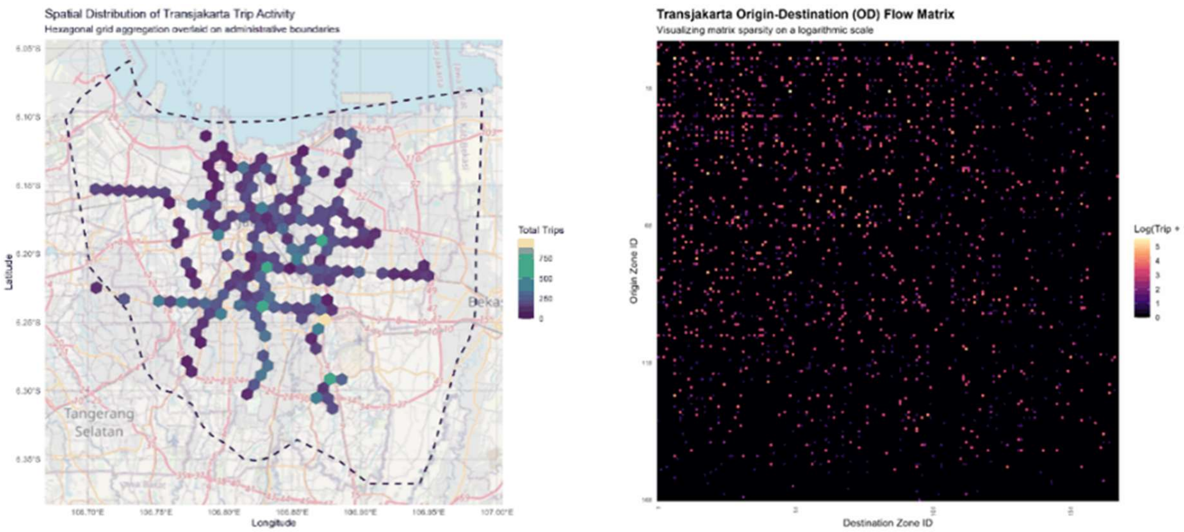
©Department of Statistics, Universitas Negeri Padang
DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

III. RESULT AND DISCUSSION

A. Spatial Representation and OD Matrix Characteristics

Before the evaluation of the Bayesian imputation model, the spatial structure and sparsity patterns of the data should be checked. To counteract the MAUP bias and provide a consistent spatial topology, the Transjakarta network is discretised using the H3 hexagonal grid scheme (Brodsky, 2018; Comber et al., 2022). **Figure 1(a)** shows that trip activity throughout 169 hexagonal zones is heavily concentrated in major metropolitan corridors, with intensity decreasing towards periphery locations.



(a) (b)

Figure 1. (a) Spatial distribution of Transjakarta trip activity, (b) Origin-Destination matrix heatmap

The associated Origin-Destination (OD) matrix is quite sparse, a problem that is often seen in open BRT systems owing to the absence of tap-out data (Blume et al., 2022; Gordon et al., 2018). **Figure 1(b)** we show this scenario as a logarithmic heatmap, where zeros are abundant and there are nodes of destination that are completely dormant. Such extreme sparsity hints the need to include an Intrinsic Autoregressive (IAR) spatial prior to allow robust reconstruction of missing mobility flows.

B. Parameter Estimation and Latent Network Reconstruction

Estimation in this part, we analyse the hierarchical Bayesian model with an Intrinsic Autoregressive (IAR) prior for reconstructing sparse Origin-Destination (OD) matrices. To evaluate the robustness of the prior specification, a prior sensitivity analysis was performed by varying the exponential hyperparameter assigned to the spatial standard deviation parameters (σ_u and σ_v) using $\lambda = 0.5$, 1.0, and 2.0. Such analyses are recommended in Bayesian modeling to evaluate whether posterior inference is unduly influenced by reasonable changes in prior assumptions (Betancourt, 2020; Gelman et al., 2020). The results are presented in **Table 1**.

Table 1. Prior sensitivity analysis under alternative exponential prior specifications.

Prior	α	σu	σv	κ	WAIC	LOOIC
Exp(0.5)	2.540	0.189	0.329	0.517	12636.23	12637.13
Exp(1.0)	2.541	0.183	0.327	0.517	12635.21	12635.91
Exp(2.0)	2.541	0.167	0.328	0.517	12635.11	12635.77

Based on **Table 1**, varying the exponential hyperparameter had only a limited effect on the posterior estimates. The estimates of α and κ remained highly consistent across all prior specifications. Likewise, the estimates of σu and σv changed only slightly under the different exponential priors. The WAIC and LOOIC values also showed only minor

©Department of Statistics, Universitas Negeri Padang
DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

differences among the three prior specifications, indicating similar predictive performance. Overall, the sensitivity analysis suggests that the proposed Bayesian IAR model is robust to reasonable changes in the prior specification, implying that the posterior inference is influenced more by the observed data than by the choice of prior (Betancourt, 2020; Gelman et al., 2020). For consistency, the Bayesian IAR model with the Exp(2) prior specification is adopted for all subsequent analyses presented in this study.

The trace plots in **Figure 2** further support convergence, with all MCMC chains being sufficiently well-mixed and no divergent behaviour, demonstrating the effective posterior exploration of the Hamiltonian Monte Carlo (HMC) approach (Gabry et al., 2019). Posterior sampling was conducted using four independent chains, with each chain run for 2,000 iterations, including 1,000 warm-up iterations and 1,000 post-warm-up sampling iterations. In addition, the four chains overlap closely throughout the post-warmup iterations, with no visible trends or chain separation, suggesting that the MCMC sampling was stable for the main model parameters (α, σu, and σv).

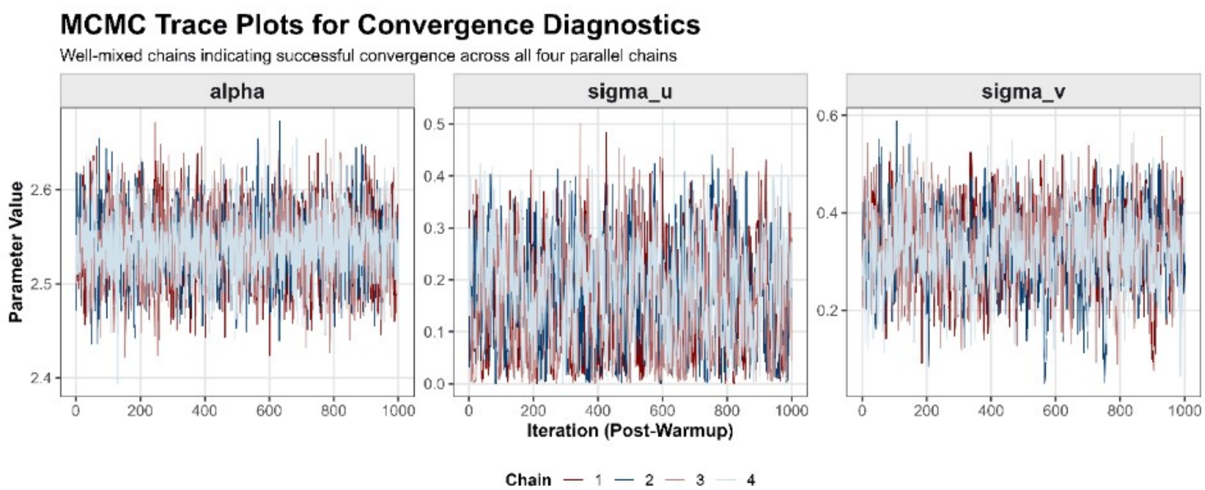
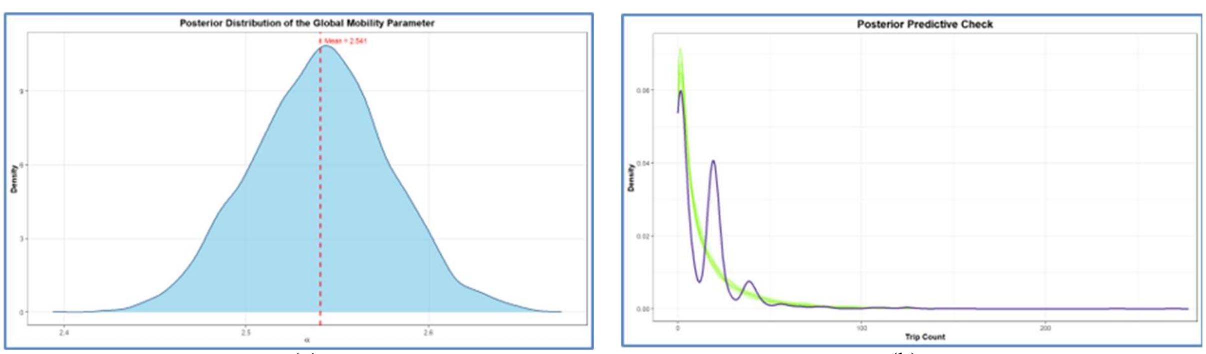


Figure 2. MCMC Traceplots for Convergence Diagnostics

The posterior mean of the global mobility parameter (β_0) is 2.541, indicating a stable estimate of the baseline trip intensity. The posterior distribution, as shown in **Figure 3(a)**, is nearly symmetric with minimal uncertainty indicating strong inferential stability (McElreath, 2020). Moreover, the model performance is tested by Posterior Predictive Checking (PPC). This means that the model may capture the overdispersion and zero-inflation of the OD matrix (Gelman et al., 2020) as seen in **Figure 3(b)** where the replicated distribution (y^{rep}) is very near to the actual data (y).



(a) (b)

Figure 3. (a) Posterior distribution of the global mobility parameter, (b) Posterior predictive check

Once the model adequacy was confirmed, the spatial effects were estimated using the IAR prior. The posterior estimates of the spatial standard deviation parameters (σu = 0.000117 and σv = 0.001817) indicate that the estimated spatial random effects remain stable under the selected Exp(2) prior specification. **Figures 4(a)** and **4(b)** present the

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

posterior mean origin and destination spatial effects, revealing clear spatial heterogeneity across the hexagonal zones. To facilitate geographical interpretation, **Figure 4(c)** summarizes the relative hotspot and coldspot areas identified from the posterior mean spatial effects using the mean ± one standard deviation criterion. The results indicate that passenger movements are not randomly distributed. Instead, neighbouring hexagonal cells tend to exhibit similar movement patterns, reflecting the spatial dependence captured by the IAR prior. (Banerjee et al., 2014; Donegan, 2022; Morris et al., 2019)

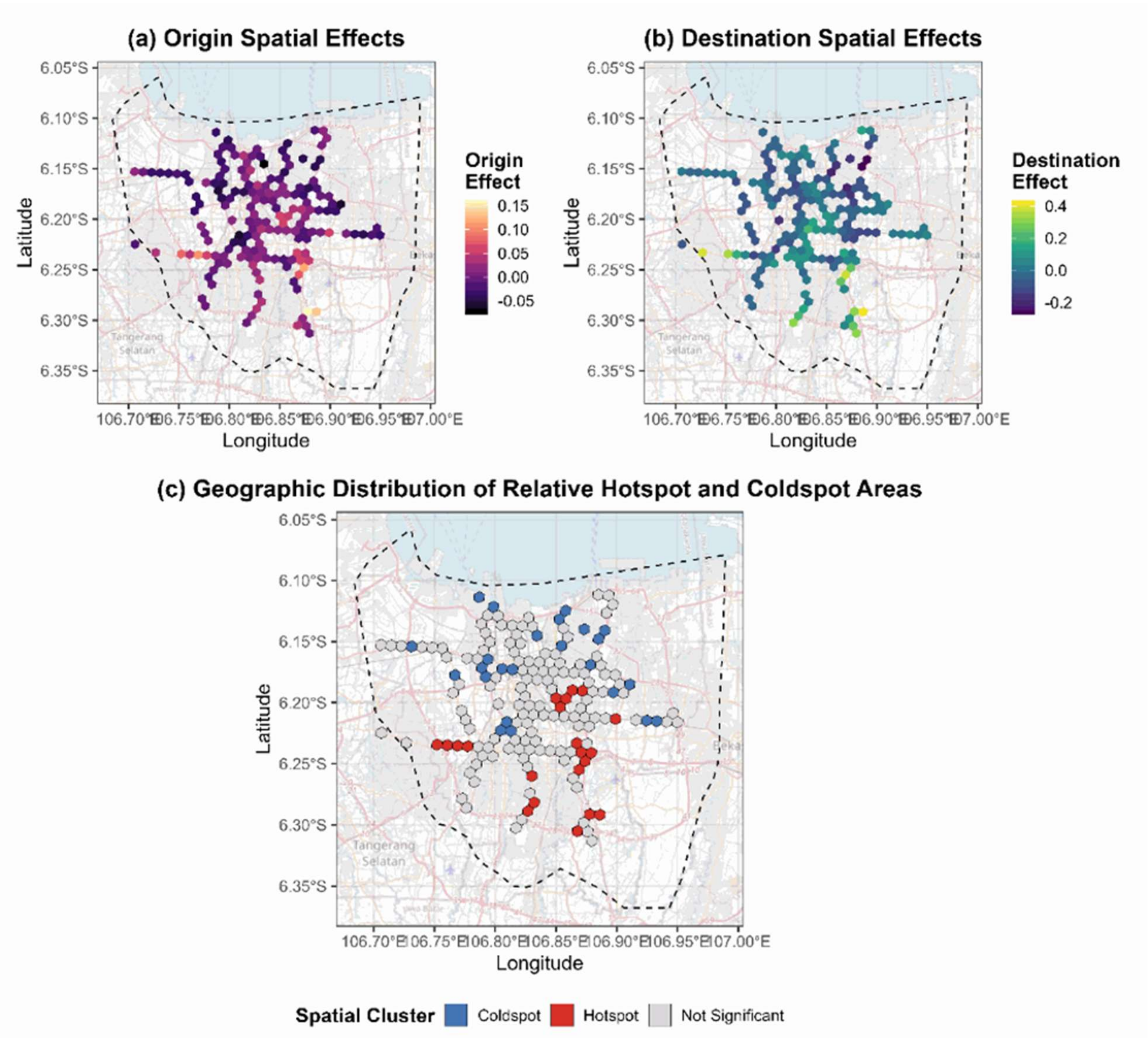
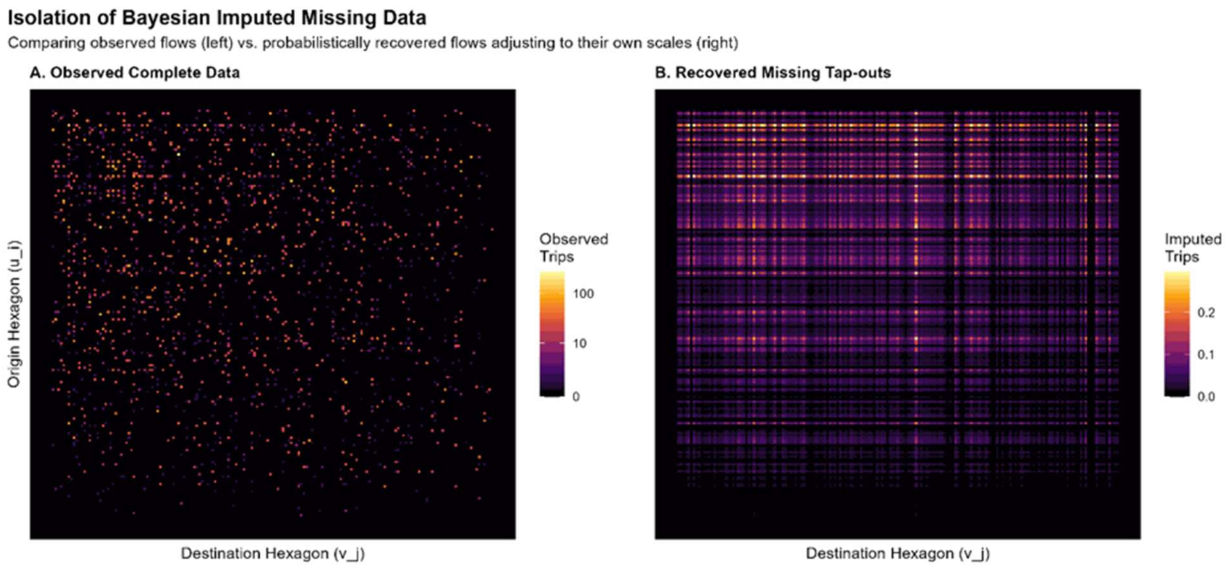


Figure 4. (a) Origin spatial effects, (b) destination spatial effects, and (c) Geographic Distribution of Hotspot and Coldspot Areas

Finally, the validated model is used to recreate missing flows in the sparse OD matrix. The reconstructed matrix, as shown in **Figure 5**, exhibits a much more organised mobility pattern than the initial sparse data. The model demonstrates the methodological potential to estimate missing tap-out flows and reconstruct a more comprehensive OD representation for mobility analysis by incorporating spatial information through the IAR prior (Blume et al., 2022; Gordon et al., 2018).

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



(a) (b)

Figure 5. (a) Original sparse OD matrix, (b) reconstructed OD matrix

C. Model Performance Evaluation and Benchmarking

Because the missing tap-out observations are naturally censored, their actual destination flows are unknown. Consequently, reconstruction errors such as RMSE or MAE cannot be computed directly, since these metrics require known reference values. The main Spatial IAR model is contrasted against an independent (non-spatial) Negative Binomial model as a benchmark for assessing the robustness of the proposed spatial specification. The assessment is based on the predictive information criteria, namely, the Widely Applicable Information Criterion (WAIC), Leave-One-Out Cross-Validation (LOOIC) and MCMC convergence diagnostics shown in **Table 2** (Vehtari et al., 2017, 2021).

Table 2. Model Performance and Convergence Diagnostics

Model	WAIC	LOOIC	Max R-hat
Spatial IAR (Primary)	12635.11	12635.77	1.015
Independent (Baseline)	12658.3	12658.3	1.012

Table 2 shows that the Spatial IAR model has lower WAIC and LOOIC values than the independent model, suggesting better predictive performance. Although the difference between the two models is relatively small, the Spatial IAR model consistently performed better after accounting for spatial dependence among neighbouring zones.

Both models achieved MCMC convergence. The Spatial IAR model has a maximum $\hat{R}$ The value is 1.015 and the independent model has a maximum $\hat{R}$ The value is 1.012. These values are still below the recommended convergence threshold ($\hat{R} < 1.05$). Therefore, the posterior estimates can be considered reliable. WAIC and LOOIC can then be used to compare the predictive performance of the two (Vehtari et al., 2017, 2021)

The performance of the Spatial IAR model shows that adding a spatial dependence structure through the Intrinsic Autoregressive prior provides additional information that cannot be captured by a model with independent random effects (Donegan, 2022). Based on these results, spatial relationships between zones contribute to better predictive performance. Therefore, the Spatial IAR model is more suitable for modelling sparse Origin–Destination data that exhibit spatial dependence.

D. Methodological Implications for Transportation Systems

The results of this study should be interpreted primarily as a methodological evaluation of the proposed Bayesian IAR framework using a synthetic transportation transaction dataset. The reconstructed OD matrix demonstrates how the proposed framework can be used to handle missing destination information and incorporate spatial dependence when analysing incomplete transportation data. However, because the dataset used in this study is synthetic and does not represent actual Transjakarta passenger transactions, the reconstructed mobility patterns and geographical clusters should not be interpreted as empirical representations of real passenger behaviour or as direct evidence for transportation

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

policy decisions. The usefulness of synthetic mobility data depends on its representativeness and task-specific utility, which should be carefully evaluated before extending methodological results to real-world applications (Kapp et al., 2024).

The methodological findings nevertheless demonstrate the potential applicability of the proposed framework to real-world transportation data, such as automated fare collection (AFC) or smart-card transaction data, where incomplete alighting information and spatially structured travel patterns may occur. AFC and smart-card data have been widely recognised as valuable sources for analysing public transport demand, passenger movements, and transportation planning (Pelletier et al., 2011). However, OD estimation methods based on incomplete fare-collection data require empirical evaluation and validation before being used to draw conclusions about actual travel behaviour (Alsger et al., 2016).

In practical applications, the framework could potentially support analyses related to passenger movement patterns, network demand, and spatial mobility after its performance has been validated using real transportation transaction data. Therefore, the present study should be regarded as a proof-of-concept and methodological validation using a controlled synthetic dataset. Future research should evaluate and validate the proposed framework using real AFC data before drawing practical conclusions regarding fleet allocation, service frequency adjustments, mobility hotspots, Transit-Oriented Development, or multimodal transportation integration.

IV. CONCLUSION

This paper offers a spatial hierarchical Bayesian model for the reconstruction of a sparse Origin-Destination (OD) matrix using a synthetic transportation transaction dataset, based on a Negative Binomial distribution and an Intrinsic Autoregressive (IAR) prior on a hexagonal grid. PPC examination indicates that the model is capable of reproducing the major features of the observed data, including overdispersion and zero-inflation. The comparison with the Independent baseline showed that the selected Spatial IAR model achieved lower WAIC and LOOIC values of 12635.11 and 12635.77, compared with 12658.30 and 12658.30 for the Independent model, while both models achieved satisfactory MCMC convergence, with maximum R-hat values of 1.015 and 1.012, respectively. The prior sensitivity analysis further indicated that the main results remained relatively stable under the alternative exponential prior specifications considered, with the Exp(2.0) specification producing the lowest WAIC and LOOIC values among the examined priors. The reconstructed spatial effects and hotspot analysis show spatial dependence across neighbouring hexagonal zones within the synthetic dataset, demonstrating that the proposed model provides a suitable framework for reconstructing incomplete trip interactions while incorporating spatial structure. However, because the dataset used in this study is synthetic, the findings should be interpreted primarily as a methodological evaluation rather than as empirical evidence of actual Transjakarta passenger mobility patterns. Future work may validate the proposed framework using real transportation transaction or automated fare collection data and incorporate physical route characteristics and spatio-temporal factors to further improve the reconstruction of urban mobility patterns.

ACKNOWLEDGMENTS

The author would like to offer his sincere thanks to everyone who has supported him and given helpful input throughout the research and production of this work.

REFERENCES

Alsger, A., Assemi, B., Mesbah, M., & Ferreira, L. (2016). Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C: Emerging Technologies*, *68*, 490–506. https://doi.org/10.1016/j.trc.2016.05.004

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17115

Betancourt, M. (2020). *Towards a principled Bayesian workflow*. Betanalpha.

Blume, S. O. P., Corman, F., & Sansavini, G. (2022). Bayesian origin-destination estimation in networked transit systems using nodal in- and outflow counts. *Transportation Research Part B: Methodological*, *161*, 60–94. https://doi.org/10.1016/j.trb.2022.04.006

Brodsky, I. (2018). *H3: Uber's hexagonal hierarchical spatial index*. Uber Engineering Blog.

©Department of Statistics, Universitas Negeri Padang
DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Comber, A., Harris, P., Bratkova, K., Phe, H. H., Kieu, M., Bui, Q. T., Nguyen, T. T. H., Wanjau, E., & Malleson, N. (2022). Handling the MAUP: methods for identifying appropriate scales of aggregation based on measures on spatial and non-spatial variance. *AGILE GIScience Ser., 3, 30.*, *3*, 1–5. https://doi.org/10.5194/agile-giss-3-30-2022

Dikisahkan. (2023). *TransJakarta Transportation Transaction. Retrieved from https://www.kaggle.com/dikisahkan/transjakarta-transportation-transaction*. Kaggle.

Donegan, C. (2022). *Building spatial conditional autoregressive (CAR) models in the Stan programming language*. https://doi.org/10.31219/osf.io/3ey65

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., & Surprenant-Legault, J. (2014). New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. *Transportation*, *41*(1), 193–210. https://doi.org/10.1007/s11116-013-9508-z

Fiechter, J., Santora, J. A., & Chavez, F. (2020). Krill hotspot formation and phenology in the California Current Ecosystem. *Geophysical Research Letters*, *47*(13), e2020GL088039. https://doi.org/10.1029/2020GL088039

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian Workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). *Bayesian Workflow*.

Gordon, J. B., Koutsopoulos, H. N., & Wilson, N. H. M. (2018). Estimation of population origin–interchange–destination flows on multimodal transit networks. *Transportation Research Part C: Emerging Technologies*, *90*, 350–365. https://doi.org/10.1016/j.trc.2018.03.007

Ibraeva, A., Correia, G. H. de A., Silva, C., & Antunes, A. P. (2020). Transit-oriented development: A review of research achievements and challenges. *Transportation Research Part A: Policy and Practice*, *132*, 110–130. https://doi.org/10.1016/j.tra.2019.10.018

Kapp, A., Hansmeyer, J., & Mihaljević, H. (2024). Generative models for synthetic urban mobility data: A systematic literature review. *ACM Computing Surveys*, *56*(4), 93. https://doi.org/10.1145/3610224

Krisztin, T., & Piribauer, P. (2021). Modelling European regional FDI flows using a Bayesian spatial Poisson interaction model. *The Annals of Regional Science*, *67*(3), 593–616. https://doi.org/10.1007/s00168-021-01058-x

LeSage, J. P., & Pace, R. K. (2008). Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, *48*(5), 941–967. https://doi.org/10.1111/j.1467-9787.2008.00573.x

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data,* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

McElreath, R. (2020). *Statistical Rethinking* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., & DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. *Spatial and Spatio-Temporal Epidemiology*, *31*, 100301. https://doi.org/10.1016/j.sste.2019.100301

Ni, L., & Zhang, D. (2024). Using a spatial autoregressive model with spatial autoregressive disturbances to investigate origin-destination trip flows. *PLOS ONE*, *19*(6), e0305932. https://doi.org/10.1371/journal.pone.0305932

©Department of Statistics, Universitas Negeri Padang
DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557–568. https://doi.org/10.1016/j.trc.2010.12.003

Shalit, N., Fire, M., & Ben-Elia, E. (2023). A supervised machine learning model for imputing missing boarding stops in smart card data. *Public Transport*, *15*(2), 287–319. https://doi.org/10.1007/s12469-022-00309-0

Sussman, A. L., Gardner, B., Adams, E. M., Salas, L., Kenow, K. P., Luukkonen, D. R., Monfils, M. J., Mueller, W. P., Williams, K. A., Leduc-Lapierre, M., & Zipkin, E. F. (2019). A comparative analysis of common methods to identify waterbird hotspots. *Methods in Ecology and Evolution*, *10*(9), 1454–1468. https://doi.org/10.1111/2041-210X.13209

Tessema, Z. T., Tesema, G. A., Ahern, S., & Earnest, A. (2023). A Systematic Review of Areal Units and Adjacency Used in Bayesian Spatial and Spatio-Temporal Conditional Autoregressive Models in Health Research. *International Journal of Environmental Research and Public Health*, *20*(13), 6277. https://doi.org/10.3390/ijerph20136277

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Zhuang, D., Wang, S., Koutsopoulos, H., & Zhao, J. (2022). Uncertainty Quantification of Sparse Travel Demand Prediction with Spatial-Temporal Graph Neural Networks. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4639–4647. https://doi.org/10.1145/3534678.3539093

©Department of Statistics, Universitas Negeri Padang
DOI: https://doi.org/10.24036/ujsds/vol4-iss3/508