Lab Report

Title: A comparison of spatiotemporal Bayesian models for uber pickup prediction

Notice: Dr. Bryan Runck

Author: Yaxuan Zhang

Date: 12/19/2023

Project Repository: https://github.com/YaxuanSeanZhang/Bayesian-Uber-Pickup

Google Drive Link:

Time Spent: 60 hours

Abstract

This project explores spatial-temporal traffic patterns in New York City using Uber pickup data, addressing spatial, temporal, and spatiotemporal trends and interactions. Bayesian models are developed to predict Uber pickup events at the county level, incorporating zone characteristics, spatial dependencies, and temporal dependencies. The project involves data pulling, processing, exploratory analysis, modeling, and evaluation, with input data from Kaggle and NHGIS sources including Uber pickup points and zone characteristics. The workflow involves data formatting, spatial joining, and aggregation for model input. Four Bayesian models are designed, including a base model, a spatial model incorporating neighboring effects, a temporal model considering temporal dependencies, and a combined spatial-temporal model. Model results demonstrate the impact of zone characteristics and spatial and temporal dependencies on Uber pickup frequency. Evaluation metrics such as Mean Absolute Deviance and Akaike Information Criterion indicate that models incorporating spatial information have better interpretability and goodness of fit. The study emphasizes the significance of considering spatiotemporal dependencies in traffic modeling for predictive accuracy and pattern understanding.

Problem Statement

This project aims to understand the spatial-temporal pattern of traffic patterns in NYC by using Uber Pickups data in New York City. The project will answer a few following questions:

- What is the spatial, temporal, and spatial-temporal traffic trend in NYC?

- How to model uber traffic across space and time considering spatial-temporal dependency/interaction?

In this project, I will design and implement three Bayesian models to predict the number of Uber pickup events at county levels using Bayesian posterior. The project splits into four main components: 1) Read Data; 2) Data Processing; 3) Data Modeling; and 4) Model Evaluation.

Table 1. Major Components of the Project

	Requirement	**Defined As**	**(Spatial) Data**	**Attribute Data**	**Dataset**	**Preparation**
1	Read Data	Raw input dataset from Kaggle and NHGIS	Pickup location and geo-boundary	population, employed population, housing unit	Kaggle; NHGIS	Request data and read data from local
2	Data Processing	Format data	County	population, employed population, housing unit	Processed data	Format data, spatial join, data summary
3	Data Modeling	Four Bayesian models	County; Spatial Neighbor	population, employed population, and housing unit	Processed training data	Define and train three models separately
4	Model Evaluation	Evaluate model performance and accuracy	County	Number of Pickup Events	Processed test data	measurements for model accuracy

Input Data

The data used in this project includes two main sources. 1) Uber picks up data points from the Kaggle website. 2) the spatial boundaries and zone characteristics requested on the NHGIS website. Specifically, each Uber pick-up data point includes a pickup location as a point of latitude and longitude and a pickup time with precision at the second level. For spatial boundaries and zone characteristics data, I sent a data request to the NHGIS server, both at the county and sub-county levels, including the total population, employed population, and housing unit variables from the ACS 5-year dataset.

Table 2. Input Data

	Title	**Purpose in Analysis**	**Link to Source**
1	Uber Pickup Points - training	Raw point data as training sets for Bayesian modeling, using data in April, 2014	https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city
2	Uber Pickup Points - test	Raw point data as test sets of model validation, using data in May, 2014	https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city
3	Boundary data – County	The analysis unit in this project	https://data2.nhgis.org/main
4	Zone Characteristics (pop, housing unit, employment)	The predictors in the Bayesian models	https://data2.nhgis.org/main

Methods

The workflow of this project is shown in Figure 1. The data is formatted in order to fit into the model. After spatially joining Uber pickup points data with clipped county boundary data, the data are aggregated as the mean value of pickup frequencies for each county at an hourly level. Moreover, the base pickup frequency in each county is summarized as the mean value of pickup frequencies from 2 am to 5 am. Spatial weight is generated by the *arcpy* function *Polygon Neighbors* and reformatted by the *nx* package. Temporal weight is generated by the *numpy* package. All the above data together with the county characteristics data are used for the following data modeling.

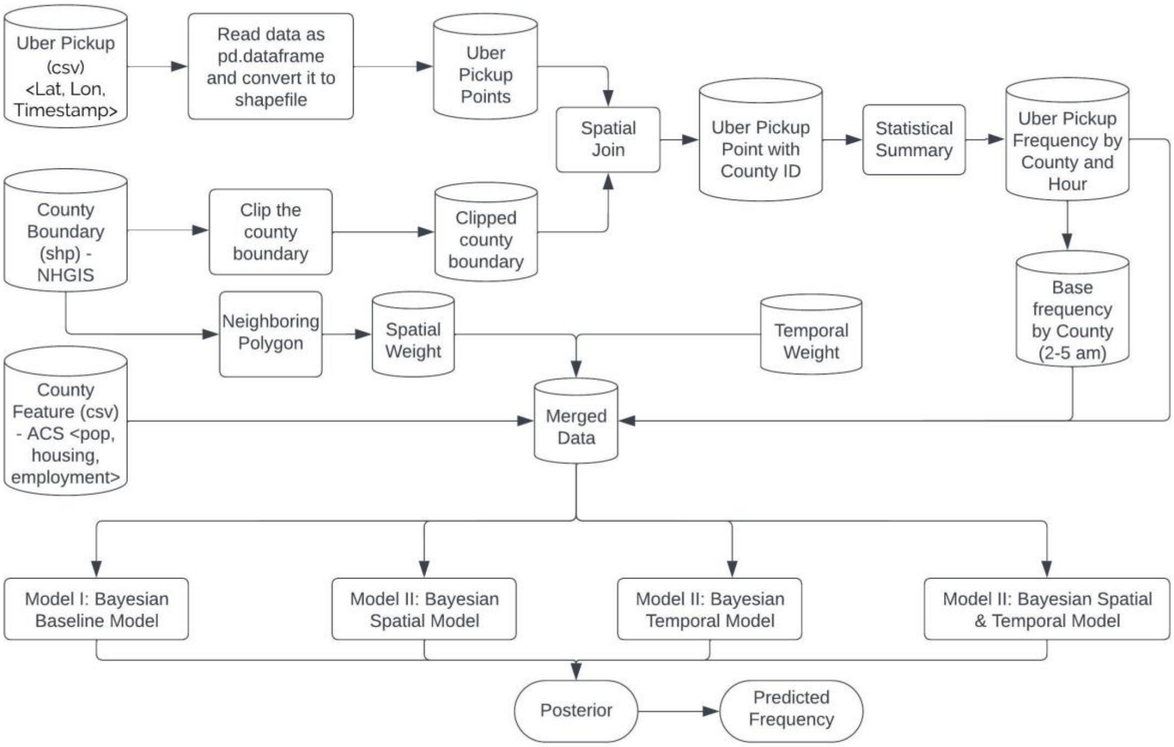
Uber Pickup (csv) <Lat, Lon, Timestamp>
Read data as pd.dataframe and convert it to shapefile
Uber Pickup Points
Spatial Join
Uber Pickup Point with County ID
Statistical Summary
Uber Pickup Frequency by County and Hour
County Boundary (shp) - NHGIS
Clip the county boundary
Clipped county boundary
Base frequency by County (2-5 am)
Neighboring Polygon
Spatial Weight
Temporal Weight
County Feature (csv) - ACS <pop, housing, employment>
Merged Data
Model I: Bayesian Baseline Model
Model II: Bayesian Spatial Model
Model II: Bayesian Temporal Model
Model II: Bayesian Spatial & Temporal Model
Posterior
Predicted Frequency

Figure 1. Project flow diagram

Hierarchical Bayesian models are applied. At the first level, the number of pickups (Y_{it}) follows a Poisson distribution, where n_i is the base pickup frequency for each county and R_{it} is the pickup rates of each county i at a t-th time interval.

$$Y_{it} \sim Poisson(n_i R_{it})$$

At the second level, four different Bayesian models will be implemented and compared:

1) Base Model (only with random bias)

The base model is designed as a log link of a combined effect of zonal characteristics predictors. X_i represents a series of zonal characteristics, e.g., population, housing unit, and employment. α represents the intercept. ε_{it} represents the random bias county i at a t-th time interval. ε_{it} follows a normal distribution with $\mu = 0$.

$$log(R_{it}) = \alpha + \beta X_i + \varepsilon_{it}$$

2) Spatial Bayesian Model (spatial neighboring effect)

The spatial Bayesian model adds a spatial effect term, borrowing information from its neighboring counties. The neighboring counties are defined as the counties that share boundaries. Thus, the value in the spatial weight (W) is 1 for two neighboring counties, otherwise is 0. The spatial effect term (θ_i) follows a conditional autoregression distribution.

$$log(R_{it}) = \alpha + \beta X_i + \theta_i + \varepsilon_{it}$$

$$\theta_i \sim \mathcal{N}(\mu_i, \sigma_\theta^2)$$

$$\mu_i \sim CAR(W, \sigma_\mu^2)$$

3) Temporal Bayesian Model (temporal neighboring effect)

The temporal Bayesian model adds a temporal effect term, borrowing information from its neighboring time interval. The neighboring time intervals are defined as the adjacent hour, e.g., 7 am and 8 am. The value in the temporal weight (U) is 1 for two neighboring time intervals, otherwise is 0. The temporal effect term (φ_t) follows a conditional autoregression distribution.

$$log(R_{it}) = \alpha + \beta X_i + \varphi_t + \varepsilon_{it}$$

$$\varphi_t \sim \mathcal{N}\left(\nu_t, \sigma_\varphi^2\right)$$

$$\nu_t \sim CAR(U, \sigma_\nu^2)$$

4) Spatial & Temporal Bayesian Model (spatial & temporal neighboring effect)

The spatial & temporal Bayesian model adds a spatial effect term and a temporal effect term as illustrated above.

$$log(R_{it}) = \alpha + \beta X_i + \theta_i + \varphi_t + \varepsilon_{it}$$

$$\theta_i \sim \mathcal{N}(\mu_i, \sigma_\theta^2)$$

$$\mu_i \sim CAR(W, \sigma_\mu^2)$$

$$\varphi_t \sim \mathcal{N}(\nu_t, \sigma_\varphi^2)$$

$$\nu_t \sim CAR(U, \sigma_\nu^2)$$

Results

Before fitting data into models, some exploratory analysis was done to discover the temporal, spatial, and spatiotemporal patterns of Uber pickups. As Figure 2 shows, there are two peaks around 7 am and 5 pm respectively, corresponding to the regular commuting time throughout the day. Moreover, the volume is largest in the evening, and it starts to decrease after 10 pm.

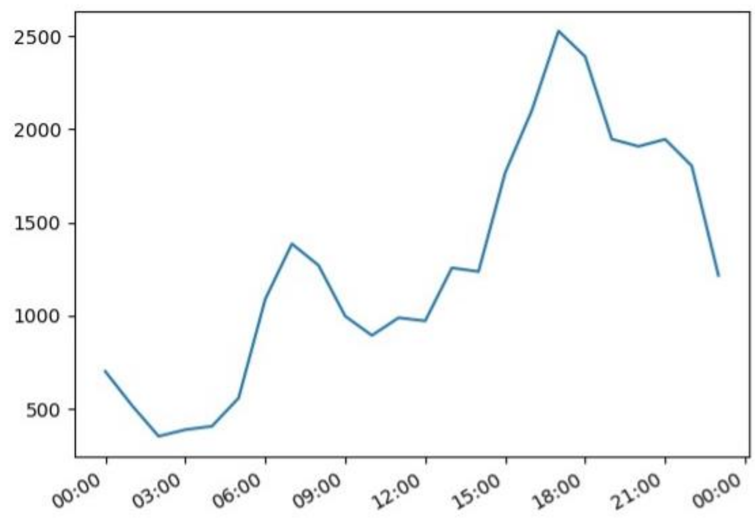

Figure 2 Temporal Patterns of Uber Pickups

In terms of spatial patterns (Figure 3), the volume is notably large within the vicinity of New York City, encompassing areas such as Manhattan, Queens, Brooklyn, the Bronx, and Hudson. The volume gradually diminishes as one moves outward from this central area. As for the spatiotemporal patterns, Figure 4 shows four spatial distribution maps at different times of the day. The overall distribution pattern is similar to the spatial pattern and volume near the New York City area increases significantly at peak hours.

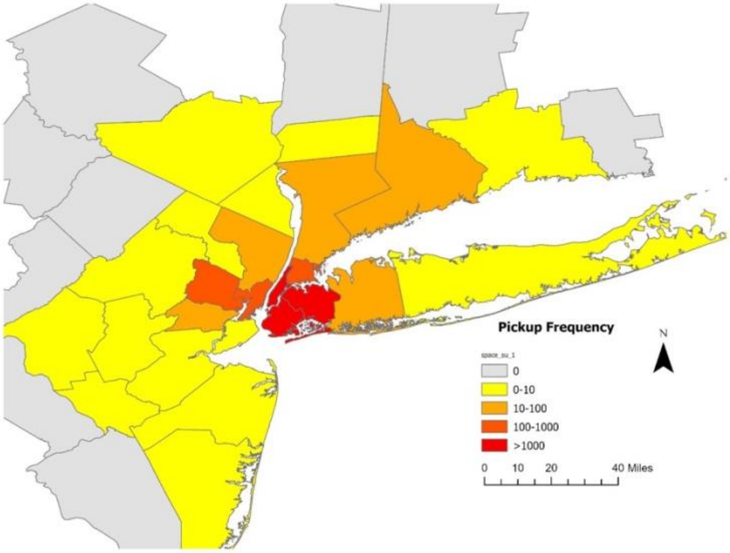

Figure 3 Spatial Patterns of Uber Pickups

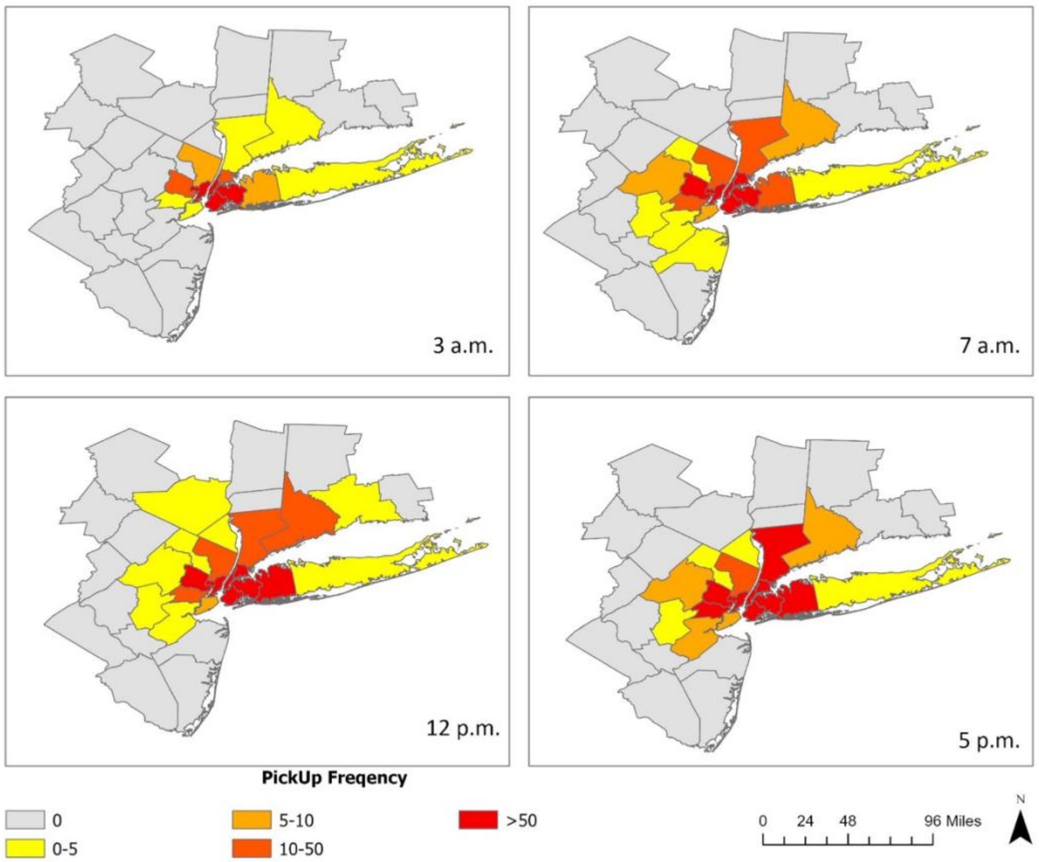

Figure 4 Spatiotemporal Patterns of Uber Pickups

The results of the Model I Bayesian base model are shown in Figure 5. PyMC3 samples from the posterior distribution using Markov Chain Monte Carlo (MCMC) methods. The trace object contains samples from the posterior distribution. It represents a record of the algorithm's exploration of the parameter space. Figure 5a shows the trace plot of the coefficient beta. Based on the posterior distribution (Figure 5b), the estimated probability of the influence of zone characteristics has a large uncertainty. Specifically, beta 0 (population) and beta 1 (employed population) cross both negative and positive values, which means their influences on Uber pickup frequency are not significant. Beta 2 shows that the housing unit has a positive association with Uber pickup frequency.

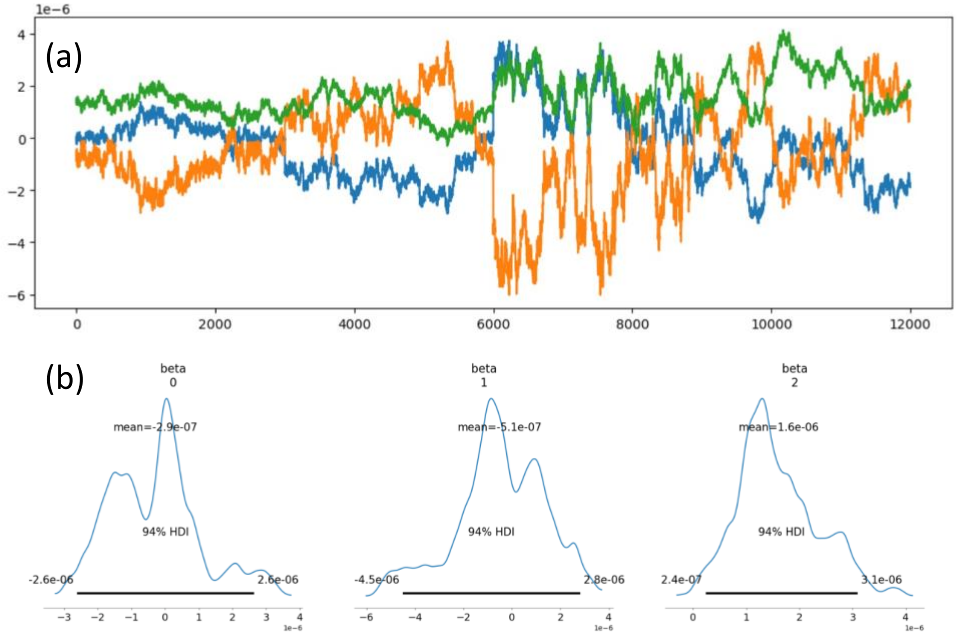

Figure 5 Model I: (a) trace plot of beta; (b) posterior of beta

The results of the Model II Bayesian spatial model are shown in Figure 6. After adding a spatial term, we can clearly see that the posterior distributions of coefficient beta are with less variance and uncertainty. As Figure 6 (a) shows, the population is negatively related to the Uber pickup frequency. On the contrary, the employed population and housing units are positively related to the Uber pickup frequency. Figure 6 (b) shows the mean of the posterior distribution of spatial term theta regarding different locations/counties. The posterior values of theta for the areas outside the NYC center area tend to be positive (orange and red colors).

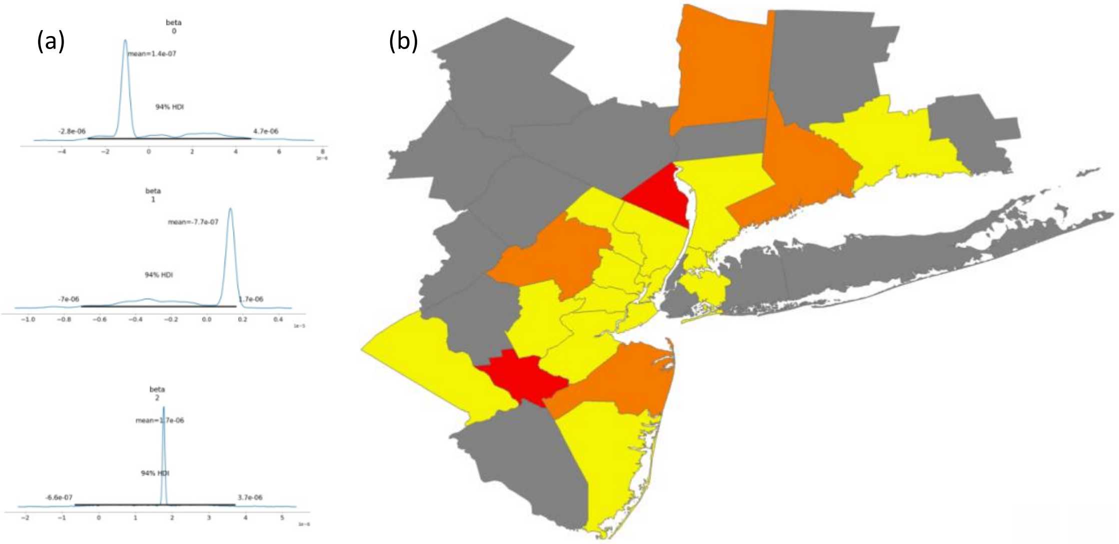

Figure 6 Model II: (a) posterior of coefficient beta; (b) posterior of spatial term theta

The results of the Model III Bayesian temporal model are shown in Figure 7. After adding a temporal, the posterior distributions of coefficient beta don't change a lot compared to the base model. As Figure 6 (b) shows, the posterior distributions of temporal term phi are consistent with the temporal patterns. For example, the phi from 2 am to 5 am are below 0 which corresponds to the low volume around that time.

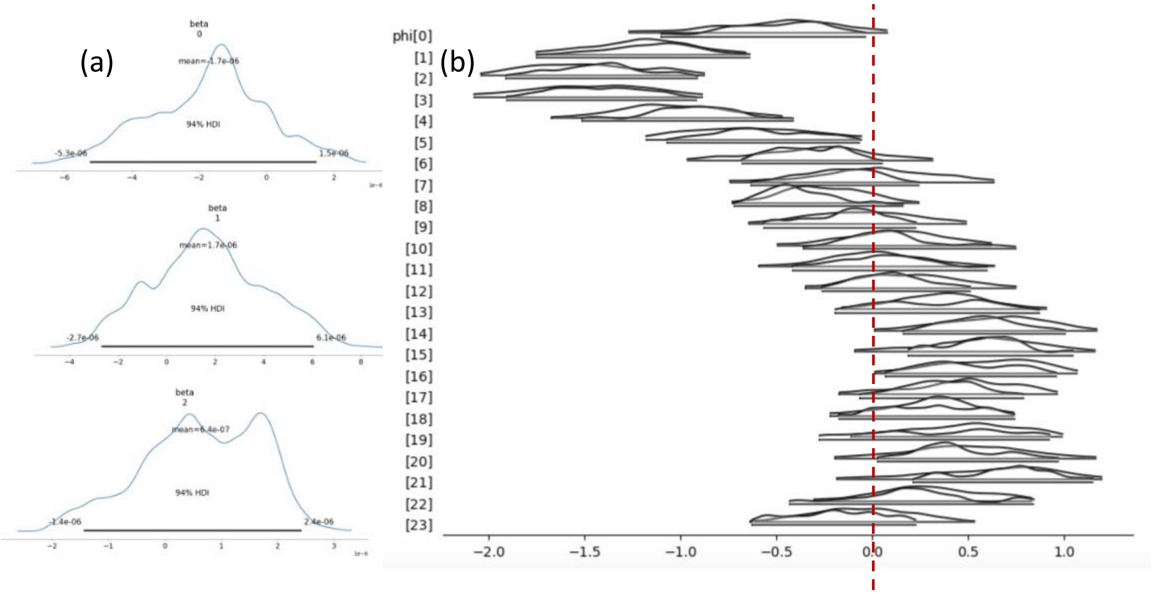

Figure 7 Model III: (a) posterior of coefficient beta; (b) posterior of temporal term phi

The results of the Model IV Bayesian spatial & temporal model are shown in Figure 8. The posterior distribution of phi for each hour exhibits a bimodal pattern, characterized by two peaks—one centered around 0, and the other either in the negative or positive range. This can be explained by both spatial and temporal patterns of Uber pickups. For example, phi [16] represents the posterior of phi at 5 pm. The right side of the posterior clusters represents the NYC area and surrounding area that has positive autoregressive effects. The posterior points clustered around 0 likely correspond to the outward area, as illustrated by the grey region in Figure 2, since they always have no pickup volume.

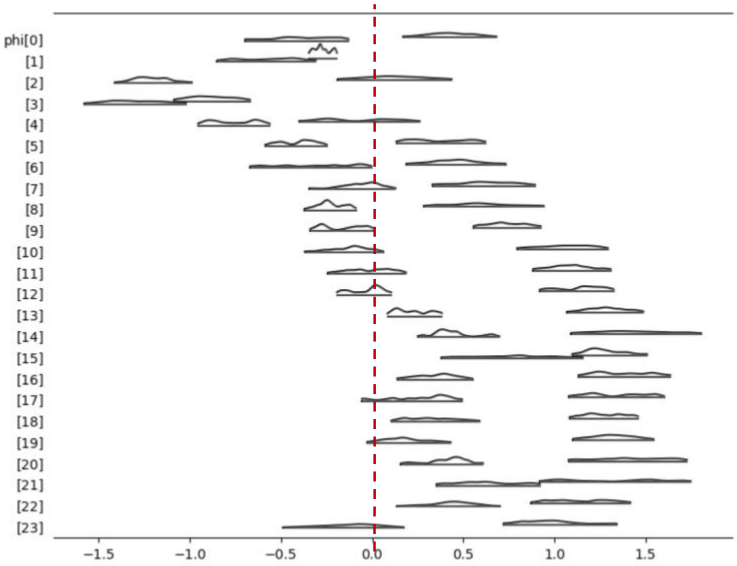

Figure 8 Model IV: posterior of temporal term phi

Results Verification

To evaluate the overall model fit and predictive performance, Mean Absolute Deviance (MAD) and Akaike Information Criterion (AIC) are used.

$$MAD = \frac{1}{N}\sum |Y^{pred} - Y^{obs}|$$

The MAD is the sum of the absolute difference between the observed value and the predicted value of each county.

$$AIC = -2 * \log likelihood + 2k$$

The AIC is a statistical measure used for model selection, balancing the goodness of fit against the complexity of the model. It penalizes models with more parameters, providing a numerical score that aids in choosing the most appropriate model for a given dataset.

Table 3 shows the evaluation results for the four models. Overall, the MAD values of the four models are similar. For AIC, Model II and Model IV have significantly smaller AIC values, which indicates that incorporating spatial neighboring information into models can better balance goodness of fit and model complexity.

Table 3. Model Comparison

	MAD	AIC
Model I: base model	109.86	1185.94
Model II: spatial model	110.29	638.01
Model III: temporal model	109.83	1218.37
Model IV: spatial & temporal model	109.83	320.89

Discussion and Conclusion

In this project, four Bayesian models are explored and compared using the Pymc3 package in Python. By adding spatial and temporal terms into the models, the spatial and temporal interactions among neighboring counties have been explored. By adding spatial terms, the location-specific biases can be diminished (Model II). Model IV suggests that we should not ignore the spatiotemporal dependency when building the model.

References

Davidson-Pilon, C. (2015). Bayesian methods for hackers: probabilistic programming and Bayesian inference. *Addison-Wesley Professional*.

Dong, N., Huang, H., Lee, J., Gao, M., & Abdel-Aty, M. (2016). Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. *Accident Analysis & Prevention*, 92, 256-264.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2, e55.

Wang, Z., Yue, Y., He, B., Nie, K., Tu, W., Du, Q., & Li, Q. (2021). A Bayesian spatio-temporal model to analyzing the stability of patterns of population distribution in an urban space using mobile phone data. *International Journal of Geographical Information Science*, 35(1), 116-134.