Scalable Spatiotemporal Modeling for Bicycle Count Prediction

Rishikesh Yadav[1], Alexandra M. Schmidt[2], Aurélie Labbe[1], Pratheepa Jeganathan[3], and Luis F. Miranda-Moreno[4]

[1]Department of Decision Sciences, HEC Montréal
[2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University
[3]Department of Mathematics and Statistics, McMaster University
[4]Department of Civil Engineering, McGill University

1 Motivation and Related Work

Understanding and modeling bicycle count data over space and time is essential for improving urban mobility, optimizing transportation systems, and supporting sustainable transportation initiatives. As cities increasingly adopt bicycle-sharing programs, robust and flexible models are needed for effective infrastructure planning, resource allocation, and enhancing the biking experience. However, bicycle count data pose several challenges. These data are inherently count-based and typically follow non-Gaussian distributions, such as Poisson or Negative Binomial, which often exhibit overdispersion, complicating traditional statistical methods. Moreover, strong non-stationarity exists, with temporal fluctuations driven by seasonal trends and weekly patterns [e.g., weekdays vs. weekends; 18], as well as spatial variations influenced by differences in urban environments. Additionally, spatiotemporal dependencies arise as bicycle usage at one location influences nearby locations over time. Another key issue is missing data, which may result from sensor malfunctions or incomplete observations, significantly affecting prediction accuracy. The high dimensionality of bicycle count data, due to the large number of locations where counts are recorded and time intervals, further require scalable and computationally efficient modeling approaches. Incorporating external covariates, such as weather conditions, special events, and infrastructure changes, can enhance predictive performance but introduces additional complexity.

Most existing approaches to modeling bicycle count data rely primarily on available covariates without explicitly accounting for spatiotemporal dependencies. Methods such as generalized linear models (GLMs) [6] offer interpretability but struggle to capture complex nonlinear relationships between predictors and bicycle counts. More recently, machine learning (ML) and deep learning (DL) algorithms have been employed for transportation data modeling, particularly in bicycle volume prediction [5, 8]. While ML and DL models effectively capture intricate nonlinearities, their lack of interpretability makes them challenging to use in transportation planning and decision-making. Additionally, there are no straightforward methods for quantifying uncertainty in ML and DL models, which often rely on bootstrapping and ensemble techniques for uncertainty estimation.

A third category of models, statistically-driven spatiotemporal models [1], offers a middle ground, balancing interpretability with flexibility. These models introduce explicit spatiotemporal dependence structures using random effects, making them particularly well-suited for transportation data applications where both interpretability and spatiotemporal prediction are essential. However, such models remain relatively underexplored for bicycle demand modeling compared to regression-based statistical techniques and ML/DL.

Among these spatiotemporal models, significant work has been done on time-varying parameter models, akin to state-space models, to capture complex temporal dependencies [16, 9, 10]. While these have been applied in other transportation contexts, their direct adaptation for bicycle count data remains an open area of research. Dynamic linear models (DLMs) [17] are a well-established class of state-space models capable of modeling non-stationary temporal trends, seasonality, and exogenous covariate effects in a unified setting. [13] recently extended such models by proposing dynamic generalized linear models (DGLMs) for site-specific human mobility counts forecasting. Their approach applies DGLMs separately to each site for future predictions without explicitly incorporating spatial dependencies between sites. These models can be further extended by incorporating spatial random effects [4] for explicit spatiotemporal modeling using the intrinsic autoregressive prior. Although this method has been applied to traffic crash frequency data, it could also be adapted for bicycle data. However, most existing approaches do not provide a fully general and computationally efficient framework or tools, such as `R` packages, for spatiotemporal inference and prediction in bicycle count data, highlighting an exciting research opportunity in this area.

2 Our Contributions

Building upon existing spatiotemporal modeling frameworks, we propose a novel computationally efficient and scalable spatiotemporal modeling approach that integrates dynamic generalized linear models (DGLMs) [17] with classical spatiotemporal statistical models [1]. Motivated by the pattern and properties of the data, our model includes the following components:

- Spatiotemporally varying intercepts to capture overall spatiotemporal trends,
- Temporally varying coefficients for purely temporal (spatially constant) covariates, allowing dynamic temporal effects on bicycle counts,
- Temporally varying coefficients for seasonal harmonics to account for periodic behaviors, such as weekly seasonality, and
- Fixed coefficients for selected exogenous covariates.

These components are combined additively in the linear predictor, which is linked to the log-mean of the Poisson distribution. Given this linear predictor, bicycle count data is assumed to follow an independent Poisson distribution. To capture spatial dependence, we include latent Gaussian processes with a Matérn correlation function. These spatial effects evolve smoothly with time through a random walk model. As the spatial correlations defined by the Matérn function involve floating operations of $O(n^3)$, we employ an approximation technique that connects the Matérn field to Gaussian Markov random fields (GMRFs) [14] through the Stochastic Partial Differential Equation (SPDE) approach [11], ensuring computational efficiency and scalability for higher spatial dimensions.

Our model naturally performs missing value imputation, spatial interpolation at unobserved locations, forecasting, and spatiotemporal prediction. Additionally, it allows for interpretation and inference of the model components, offering predictive performance comparable to complex ML/DL models while remaining interpretable. We adopt a Bayesian inference approach for our model, making uncertainty quantification straightforward. In addition, our model provides the estimation of average annual daily bicyclists (AADB) [12] which is a widely used metric in cycling studies providing a measure of cycling activity over time. For the long-term monitoring sites where sensors are installed for longer time, AADB maybe estimated directly. However for short-term sites where data are collected for only a few days or months, AADB estimation requires interpolation and missing value imputations. Traditional interpolation methods rely on long-term site data, which introduce errors due to their inability to capture dynamic trends and seasonality [2]. In contrast, our Bayesian hierarchical model explicitly incorporates these spatiotemporal patterns, significantly improving the accuracy and uncertainty quantification of AADB estimates.

While inference when the response variable follows a Gaussian distributio is straightforward using conjugacy, Bayesian inference for Poisson response is challenging due to the lack of closed-form full conditionals. To address this, we develop an efficient Markov chain Monte Carlo (MCMC) sampler that leverages gradient and observed Hessian information, using the preconditioned Metropolis-adjusted Langevin algorithm (pMALA) [7] to efficiently sample from complex posterior distributions. Furthermore, our MCMC sampler supports various GLM families and we implement our modeling framework in the `R` package `sparseDGLM`, making it applicable to a broad range of transportation datasets.

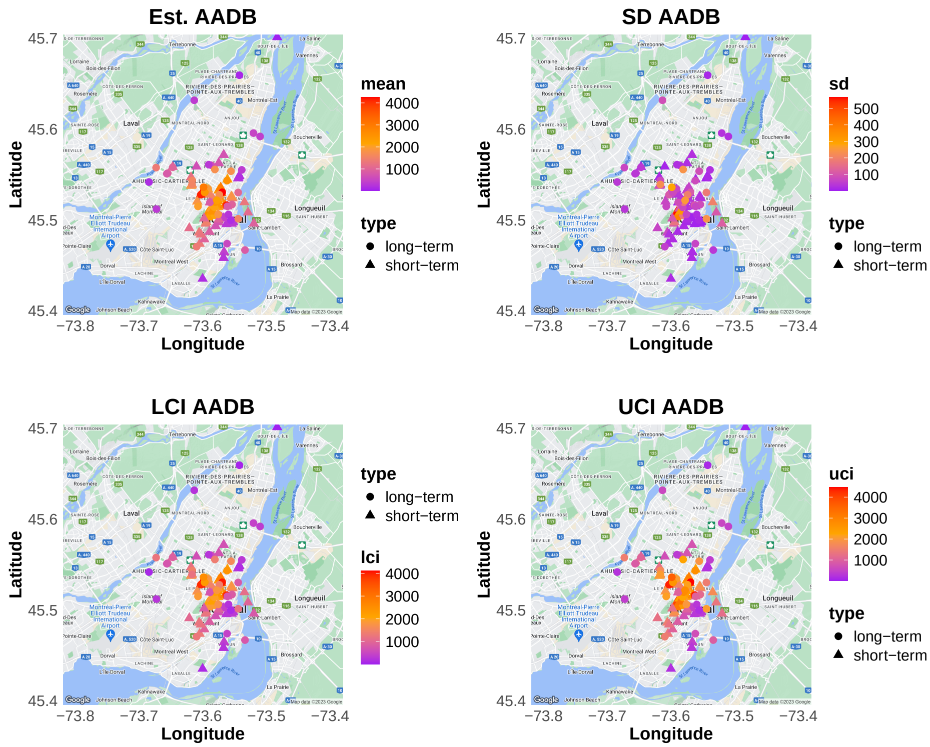

Figure 1: Estimated AADB (top left), standard errors (top right), lower (bottom left), and upper (bottom right) 95% credible intervals for the AADB estimates for both long-term and short-term sites based on the best fitted model.

3 Results

We conduct a detailed synthetic data experiments to evaluate the performance and scalability of our MCMC sampler and to assess the accuracy of the sparse SPDE-based approach. Specifically, we compare the effectiveness in spatial interpolation, forecasting, and spatiotemporal prediction. The synthetic data experiments demonstrate the efficiency and scalability of our inference approach as well as the sparse SPDE approximation of continuous Matérn random field, thus allowing us to fit our model in higher spatial dimensions.

We also apply our model to a set of Montreal's daily bicyclist data which includes 53 long-term monitoring sites and 93 short-term sites where data availability is extremely limited, often spanning only two to three days. We compare the predictive performance of our proposed model against simple Bayesian GLMs with Poisson response, BKTR [9], a complex Bayesian spatiotemporal model, and BayesNF [15], a deep learning-based framework that combines Gaussian processes and neural networks. Our model achieves comparable predictive performance with strong computational efficiency while maintaining interpretability across all model components, making it well-suited for real-world transportation applications.

The set of 93 short-time sites with more than 99% missing values are treated as completely unobserved and we perform spatial interpolation using the best fitted model. AADB is estimated for both long-term and short-term sites, along with 95% credible intervals for each site based on posterior samples, obtained using our MCMC sampler. Figure 1 showcase the estimated AADB along with their standard errors and credible intervals for both long-term and short-term sites. The results align with empirical expectations, showing higher bicycle demand near Mont-Royal park and major universities, likely due to a high concentration of students and recreational users in these areas. In contrast, sites on the outskirts of Montreal exhibit significantly lower bicycle counts. This pattern is also evident at short-term monitoring sites, where bicycle use is intuitively lower, showcasing the consistency of our model with observed trends.

In summary, our proposed model bridges the gap between the interpretability of traditional statistical models and the ability to capture non-stationarities, similar to deep learning approaches, while offering flexibility in modeling bicycle count data. Our model can also be applied to other transportation datasets that exhibit similar patterns to bicycle count data, allowing users to directly benefit from our developed `R` package, `spasreDGLM`. Finally, several future research avenues exist in order to extend the proposed model. A natural direction is the incorporation of the road network structure [3], which could further enhance prediction accuracy and provide deeper insights into urban cycling behavior.

References

[1] Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical Modeling and Analysis for Spatial Data*. Second edition. CRC Press.

[2] Beitel, D., McNee, S. and Miranda-Moreno, L. F. (2017) Quality measure of short-duration bicycle counts. *Transportation Research Record* **2644**(1), 64–71.

[3] Borovitskiy, V., Azangulov, I., Terenin, A., Mostowsky, P., Deisenroth, M. and Durrande, N. (2021) Matérn Gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics*, pp. 2593–2601.

[4] Buddhavarapu, P., Bansal, P. and Prozzi, J. A. (2021) A new spatial count data model with time-varying parameters. *Transportation Research Part B: Methodological* **150**, 566–586.

[5] Das, S. and Tsapakis, I. (2020) Interpretable machine learning approach in estimating traffic volume on low-volume roadways. *International Journal of Transportation Science and Technology* **9**(1), 76–88.

[6] Ermagun, A., Lindsey, G. and Hadden Loh, T. (2018) Bicycle, pedestrian, and mixed-mode trail traffic: a performance assessment of demand models. *Landscape and Urban Planning* **177**, 92–102.

[7] Girolami, M. and Calderhead, B. (2011) Riemann manifold Langevin and hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **73**(2), 123–214.

[8] Islam, S. (2016) *Estimation of Annual Average Daily Traffic (AADT) and Missing Hourly Volume Using Artificial Intelligence*. Master's thesis, Clemson University.

[9] Lanthier, J., Lei, M., Sun, L. and Labbe, A. (2023) BKTR: An efficient spatiotemporally varying coefficient regression package in R and Python .

[10] Lei, M., Labbe, A. and Sun, L. (2024) Scalable spatiotemporally varying coefficient modeling with Bayesian kernelized tensor regression. *Bayesian Analysis* **1**(1), 1–29.

[11] Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

[12] Nosal, T., Miranda-Moreno, L. F. and Krstulic, Z. (2014) Incorporating weather: Comparative analysis of annual average daily bicyclist traffic estimation methods. *Transportation Research Record* **2468**(1), 100–110.

[13] Pherwani, P., Hass, N. and Yanchenko, A. K. (2024) Spatiotemporal modeling and forecasting at scale with dynamic generalized linear models. *arXiv preprint arXiv:2410.07161* .

[14] Rue, H. and Held, L. (2005) *Gaussian Markov random fields: theory and applications*. CRC press.

[15] Saad, F., Burnim, J., Carroll, C., Patton, B., Köster, U., A. Saurous, R. and Hoffman, M. (2024) Scalable spatiotemporal prediction with Bayesian neural fields. *Nature Communications* **15**(1), 7942.

[16] Wang, X. and Ni, W. (2016) An improved particle filter and its application to an ins/gps integrated navigation system in a serious noisy scenario. *Measurement Science and Technology* **27**(9), 095005.

[17] West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. Second edition. Springer.

[18] Zhao, W. (2016) *Modeling Seasonal and Weather Impacts on Cycling Count*. Ph.D. thesis.