Predicting cycling flows in cities without cycling data

Eduardo B. Falbel[1], Lucas Meyer de Freitas[2], Kay W. Axhausen[2], Fabio Kon[1], Raphael Y. de Camargo[3]

[1]Department of Computer Science – University of São Paulo
São Paulo - SP - Brazil

[2]Institute for Transport Planning and Systems – ETH Zurich
Zurich, Switzerland

[3]Center for Mathematics, Computing, and Cognition – Federal University of ABC
Santo André – SP – Brazil

eduardo.falbel@usp.br

Abstract. *Cycling is a potential tool to mitigate many of the problems faced by urban populations today. Encouraging the use of bicycles as a legitimate mobility tool, however, demands adequate knowledge of current mobility patterns, such as locations of trip generation and attraction. Unfortunately, cities usually do not gather enough data to adequately understand cycling demand. We propose models based on spatial econometrics and gradient boosted regression trees which can be trained with data from cities with mature cycling cultures and then applied to cities still in their cycling infancy to supply city officials with a better estimate of potential future OD matrices. We perform a case study in the Boston Metropolitan Area and show results comparing both types of models.*

1. Introduction

Cycling's low CO_2-equivalent life-cycle emissions when compared to other forms of transport [Brand et al. 2021] and its potential use for short trips [Wang and Zhou 2017], thus replacing car use, present a strategy to combat climate change and help mitigate traffic congestion in large cities. Due to these and other reasons, city officials have been looking to encourage bike ridership and evidence shows that cycling sharing schemes can play a significant role in increasing cycling rates [Félix et al. 2020].

Deciding where to build these interventions, however, is quite tricky, as potential cycling demand is unknown. One piece of information that is particularly relevant is the Origin-Destination (OD) matrix, the acquisition of which is an issue that has been at the forefront of transport engineering for decades [Ortuzar and Willumsen 2011]. With existing methods, one usually needs some sample of trip/flow data from the region of interest to estimate an OD matrix [Ortuzar and Willumsen 2011, Bera and Rao 2011]. However, cities without an established cycling culture presumably do not have such data to begin with and attempts to collect it would either be too costly [Ortuzar and Willumsen 2011] or not yield enough information due to low ridership.

We propose models that can be trained on a region with a mature cycling culture- and then be used to predict flows/estimate an OD matrix for a region of interest which only has contextual data available, allowing transport planners to get the information they need to make better decisions. Our main models are based on gradient boosted decision

trees (GBDT) and spatial econometric interaction models (SEIM); the former is a state-of-the-art machine learning model which is widely used in both academia and industry, while the latter can be considered to be an 'upgrade' to the standard gravity model, a staple of transportation research. This work builds on previous research done by our group [Kon et al. 2021], which first analyzed the data we used and laid the groundwork for the project. The contributions of this paper are twofold:

- This is the first bike flow prediction model tested on spatial out-of-sample regions (meaning regions not present in the training data and completely disjoint from it), and
- it is the first application of spatial econometric interaction models for bike-flow modelling and spatial out-of-sample prediction.

The rest of the paper is organized as follows: in Section 2, we take a deeper look into one of the main approaches used in this paper, spatial econometric interaction models. In Section 3, we present a literature review focusing on analysis and prediction of flows in bike-sharing systems and the use of spatial econometric interaction models. We then move on to describe the methodology we used to carry out this research in Section 4, followed by the presentation of our results in Section 5. We discuss our results and present our conclusions in Section 6.

This undergraduate research work was the result of a collaboration with the Institute for Transport Planning and Systems at ETH Zurich. An extended version of this article was accepted to the 103rd Transportation Research Board Annual Meeting and presented on January 10th, 2024 in Washington D.C., USA.

2. Spatial Econometric Interaction Models

Spatial interaction models (SIMs) have been used extensively when modelling mobility flows, usually as the *trip distribution* step in the traditional 4-step model [Ortuzar and Willumsen 2011]. The most well-known of these is the gravity model, based on Newton's law of gravitation, specified by Equation 1. T_{ij} denotes trips from region i to region j, O_i and D_j are sets of variables measured at the origin and destination regions, respectively, $f(c_{ij})$ is a function of the generalized trip costs, and α is a generic balancing factor.

$$T_{ij} = \alpha O_i D_j f(c_{ij}) \tag{1}$$

Generally, estimation of the gravity model is done with Maximum Likelihood Estimation (MLE) on its log-linearized form (Equation 2).

$$\begin{aligned} \log(T_{ij}) &= \log(\alpha O_i D_j f(c_{ij})) \\ &= \log\alpha + \log O_i + \log D_j + \log f(c_{ij}) \end{aligned} \tag{2}$$

We can 'rebrand' this equation to make it look more like a regression specification for which one would use MLE, as shown in Equation 3 [LeSage and Thomas-Agnan 2014]. Now, α is the intercept and $\beta_i, i = 1, 2, 3$ are the coefficients we are trying to estimate.

$$y_{ij} = \alpha + \beta_1 O_i + \beta_2 D_j + \beta_3 f(c_{ij}) + \varepsilon \tag{3}$$

A major issue with this model, however, is that it assumes independence between flows, a premise that has been shown not to hold [LeSage and Pace 2008]. Due to this limitation, [LeSage and Pace 2008] have proposed a new class of SIMs named Spatial Econometric Interaction Models (SEIMs), which attempt to address the issue of spatial dependence between flows. One specification of these models, which we refer to as the LAG model, is given by Equation 4.

$$y = \rho_o W_o y + \rho_d W_d y + \rho_w W_w y + \beta_d X_d + \beta_o X_o + \gamma g + \alpha + \varepsilon \tag{4}$$

One can see that it is structurally very similar to the 'rebranded' log-linear form of the gravity model, except for the addition of the first three autoregressive terms in the right-hand side of the equation ($\rho_o W_o y, \rho_d W_d y, \rho_w W_w y$), composed of the dependent variable y, a spatial coefficient ρ which we attempt to estimate, and a spatial weights matrix $W_i, i = d, o, w$. The predictor we crafted, described in Section 4.3, is based on this underlying model.

3. Related Work

The rise of bike-sharing systems has motivated many researchers to attempt to predict cycling flows, usually for the purposes of system optimization [Jiang 2022]. Most of the literature focuses on the modelling of docked systems [Jiang 2022], in which there are predetermined, physical stations where users can retrieve and deposit bicycles. The goal then is usually optimizing for the current system [Zhou and Huang 2018, Chai et al. 2018] or for immediate-future improvements [Wang and Zhou 2017]. [Guidon et al. 2020] attempted to extrapolate mobility patterns from one city to another by training a spatial econometric model in Zurich and testing the model's predictions in Bern. However, the authors were only capable of modelling trip generation (i.e., demand) as opposed to the complete flow (generation and attraction), meaning they could not take into account spatial autocorrelation between flows and construct a full OD matrix from their prediction data.

Only a few studies have applied spatial econometric interaction models to flows associated with mobility/commuting. Most notably, [Ni et al. 2018] applied it to flows abstracted from mobile phone data in Hangzhou, China; [Schatzmann et al. 2019] used a simplified version of the model which did not include all three spatial weights matrices simultaneously on public transport commuting data in Switzerland; [Kerkman et al. 2017] used a multilevel approach to model public transport flows in the Netherlands and used a SEIM as the upper level model for one of these approaches; [Dargel 2021] modeled home-to-work commuting flows in Paris with SEIMs. To the best of the authors' knowledge, none has tried using SEIMs to model cycling flows or has attempted to test the predictive capabilities of these models on spatial-out-of-sample data.

4. Methodology

We begin this section by discussing our data and how it is structured, followed by the chosen data split for training and testing. Finally, we discuss one of the models used during this project in depth.

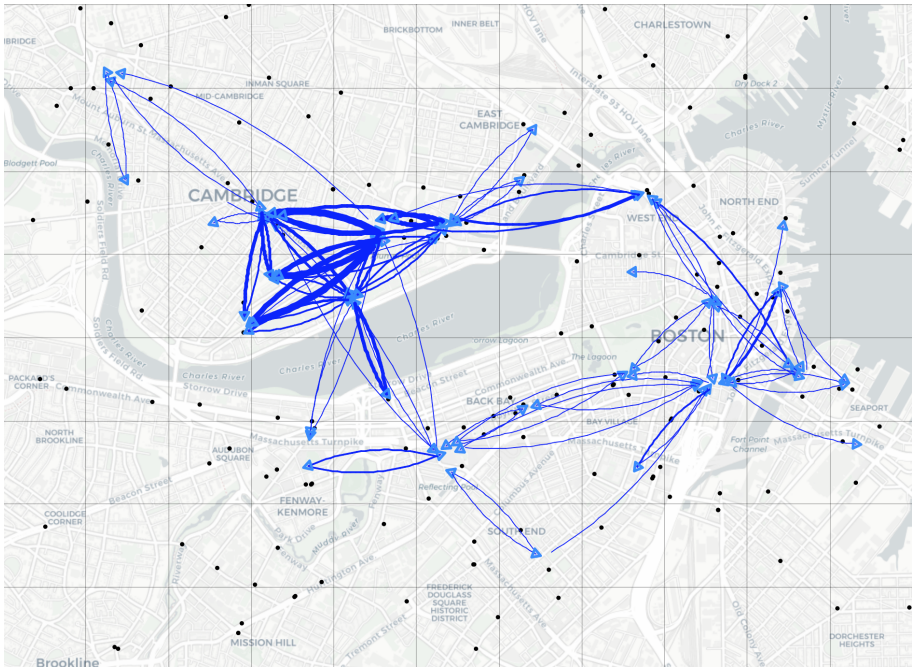

Figure 1. Flows belonging to the 1st quartile in terms of trip volume.

$$\text{Cycling Infrastructure Ratio} = \frac{length(\text{Cycling route} \cap \text{Cycling infrastructure})}{length(\text{Cycling route})} \quad (5)$$

4.1. Data sources and fusion

The base dataset we used pertains to trips made using the Boston Blue Bikes bike-sharing system between April 2018 and March 2019. This service uses fixed stations for the pick-up and drop-off of the bikes, the location of which can be seen in Figure 1.

The process for abstracting the cycling trips into flows is based on the use of a regular grid of cells for trip aggregation, since it can be used with any type of cycling trip data, be it from station-based or dockless BSSs or GPS data from tracking surveys. We then assign each trip the start and end grid cells in which its start and end stations are located, respectively, and aggregate trips with the same origin and destination to form our flows. Figure 1 shows the most substantial flows during the time period specified above. We also remove grid cells without stations in them to try and remove some of the implicit bias in the dataset, as trips cannot start or end in these cells.

We then enrich that dataframe with the distance between cells, POI data collected from Google and socioeconomic data from the US census to each cell. Since we are using cells and not census tracts, we proportionally distribute each feature derived from the census based on the area of the cell each census tract occupies. Finally, we incorporate cycling infrastructure data by means of the 'cycling infrastructure ratio' (Equation 5).

4.2. Training and Testing

Unlike standard machine learning training and testing procedure, randomly splitting the data and performing some k-fold cross validation is not feasible for accurately gauging

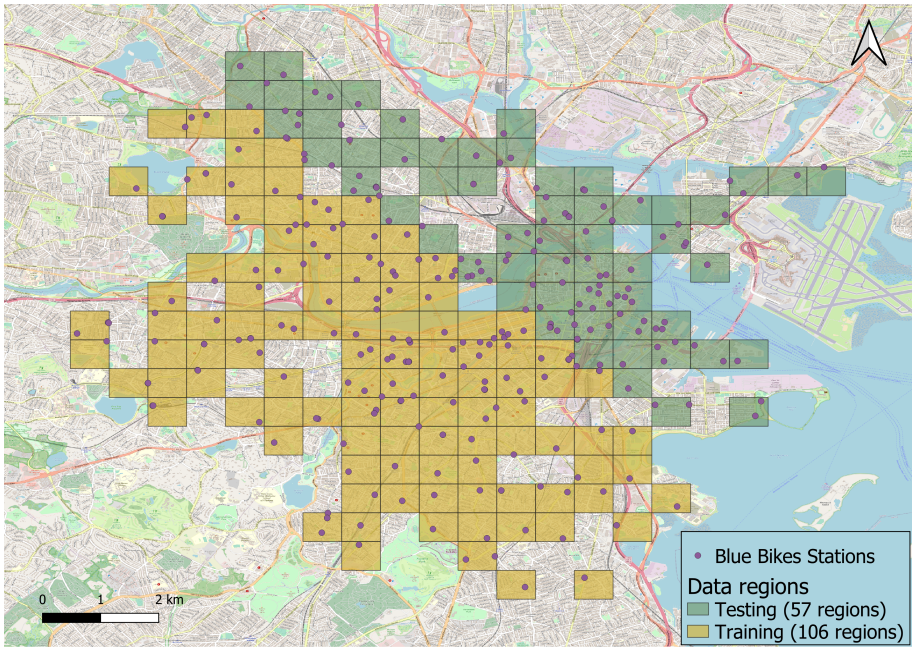

Figure 2. Training and testing split for the Boston metropolitan area

the models' predictive capabilities. This is because we rely on the explicit spatial structure of the data to determine the neighbourhood structure, and thus, the spatial weights matrices for the spatial econometric model. This means that we must manually and carefully split the dataset into regions that at least somewhat resemble each other (regarding flow volumes and patterns) as well as maintain some cohesive spatial structure. Basing ourselves on these principles, we chose the split in Figure 2, in which the yellow shaded regions constitute the training set and the green shaded ones, the testing set.

After splitting the data, we trained the models with the yellow-shaded regions in Figure 2 and tested with the green-shaded regions in Figure 2. It should be noted that the training and testing sets are considered completely disjoint from each other (meaning regions at the border of these sets do not have neighbors across the border). This is done to simulate the case in which the model is trained with data from one city and tested on data from a completely different city.

4.3. Predictors

In total, we tested 3 different predictors on the Boston bike-sharing system data for out-of-sample prediction. They are:

- CatboostRegressor (CB)
- Aspatial predictor (A)
- LAG trend-corrected predictor (TC)

As can be seen from Equation 4, the LAG model specification cannot directly be used for out-of-sample prediction, since it relies explicitly on the dependent variable. At the time of writing, there were no dedicated out-of-sample predictors for spatial econometric interaction models that we could use. Consequently, we implemented our own out-of-sample predictor[1], based on the models' expected value, described in

[1] `https://github.com/EduFalbel/seim`

[LeSage and Thomas-Agnan 2014] (Equation 10). To arrive at that specification, we start by condensing Equation 4 into Equation 6.

$$y = \rho_d W_d y + \rho_o W_o y + \rho_w W_w y + \delta Z + \varepsilon, \tag{6}$$

where $\delta = \left[\beta_d \beta_o \gamma \alpha\right]$ and $Z = \left[X_d X_o g \iota_{n^2}\right]$, ι_{n^2} is an $n^2 \times 1$ (the number of OD pairs) vector of ones.

We then subtract both sides by the spatial lags of the dependent variable and factor out the dependent variable y, resulting in Equation 7.

$$(I - \rho_d W_d - \rho_o W_o - \rho_w W_w)y = \delta Z + \varepsilon. \tag{7}$$

Now multiplying by the inverse of the spatial factor:

$$\begin{aligned}(I - \rho_d W_d - \rho_o W_o - \rho_w W_w)^{-1}(I - \rho_d W_d - \rho_o W_o - \rho_w W_w)y = \\ = (I - \rho_d W_d - \rho_o W_o - \rho_w W_w)^{-1}(\delta Z + \varepsilon),\end{aligned} \tag{8}$$

is finally equal to

$$y = (I - \rho_d W_d - \rho_o W_o - \rho_w W_w)^{-1}(\delta Z + \varepsilon). \tag{9}$$

We then craft our predictor as Equation 10, in accordance with [LeSage and Thomas-Agnan 2014].

$$\hat{y} = (I - \hat{\rho}_d W_d - \hat{\rho}_o W_o - \hat{\rho}_w W_w)^{-1}\hat{\delta} Z, \tag{10}$$

Keeping with the convention set forth in [Goulard et al. 2017] and [Dargel 2021], we will henceforth refer to Equation 10 above as the 'trend-corrected' predictor.

$$\hat{y} = \hat{\beta}_d X_d + \hat{\beta}_o X_o + \hat{\gamma} g + \hat{\alpha} + \varepsilon \tag{11}$$

Unlike the LAG model, the aspatial model specification in Equation 11 can be directly used for prediction (dependent variable is only present on the left-hand side of the model equation), so the terms 'model' and 'predictor' will be used interchangeably when referring to that specification. For the machine learning predictor, we decided on using a gradient-boosted trees model, namely Catboost, as it comes with good defaults 'out-of-the-box' [Bentéjac et al. 2020] and has demonstrated high performance in our research.

5. Results

We start by examining the Root Mean Squared Error (RMSE) and then dive into other metrics we used to gauge the quality of the models' predictive abilities, namely the predicted vs. observed flow quartiles and spatial dependency analysis.

Table 1. RMSE scores for the models' In- and Out-of-sample predictions.

Model	Out-of-sample
Aspatial	268.204
Catboost	255.014
Trend-corrected	229.366

5.1. Prediction accuracy

The first metric we used to compare the predictive power of the models is the Root Mean Squared Error. Table 1 shows this metric for each model.

What is very interesting in Table 1 is that the SLA econometric model actually beat the machine learning model when it came to the out-of-sample prediction. We believe this is a valued contribution, since the latter models are directly interpretable as shown previously and, thus, provide more useful information to transport planners as opposed to a black-box machine learning algorithm. Even though there have been advances in Explainable Machine Learning with the use of Shapley values [Lundberg and Lee 2017], the SEI models allow for a much clearer interpretation of the influence of spatial interaction effects on the cycling flows.

5.2. Trip quartiles

For this analysis we split the flows into quartiles based on the number of trips in each flow, such that each quartile contains 25% of all trips. We then created tables of the predicted versus observed flow quartiles for each of the models, which can be seen in tables 2 to 4. The quartiles are numbered in descending order of trips, meaning quartile 0 has the most substantial flows and quartile 3 has the least substantial flows (a lot of which are 'null'/0 flows). The desired outcome for the tested models would be to maximize the main diagonal, since that indicates correctly predicted flow quartiles, while minimizing the upper-right and lower-left corners, which points to *egregious* errors. That is, when the model predicted one of the least substantial flows to be one of the most substantial and vice-versa.

We believe this analysis to be one of the most important we conducted, since it is reasonable to assume that, when deciding where to build cycling infrastructure, transport officials will focus on the handful of regions with the majority of trips.

Table 2. Aspatial model quartiles.

Predicted	0	1	2	3	All
Observed					
0	4	9	13	11	37
1	2	8	37	29	76
2	4	9	56	105	174
3	2	6	86	2868	2962
All	12	32	192	3013	3249

Table 3. Catboost quartiles

Predicted	0	1	2	3	All
Observed					
0	6	14	12	5	37
1	4	13	33	26	76
2	5	11	65	93	174
3	1	13	108	2840	2962
All	16	51	218	2964	3249

We can observe that the trend-corrected predictor had the highest shares of correctly predicted tier 0, tier 1, and tier 2 flows (Table 4), while it was the Catboost followed

Table 4. Trend-corrected model quartiles.

Predicted	0	1	2	3	All
Observed					
0	12	11	10	4	37
1	9	25	28	14	76
2	12	22	70	70	174
3	7	33	163	2759	2962
All	40	91	271	2847	3249

by the Aspatial predictors that had the lowest share of tier 3 flows incorrectly predicted as tier 0 (tables 2 and 3, respectively).

5.3. Spatial dependence analysis

Another way to compare the quality of the models' predictions is to measure the presence of spatial dependence within their residuals. We do so by creating what are called Moran scatterplots for the tests (Figure 3), where the residuals are plotted against their spatial lag for each of the spatial weights matrices W_d, W_o, W_w. What we are then interested in for each graph is the angle of the linear fit and, in essence, the flatter it is, the better the model was able to account for the spatial dependence.

When examining the residuals of the test, all three of the models display similar levels of spatial dependence, as can be seen by the comparably angled trendlines in Figure 3. This was unexpected since the catboost model was not designed with the intent of dealing with spatial dependence, unlike the trend-corrected predictor, and even managed to edged it out.

6. Discussion and Conclusions

The catboost and trend-corrected predictors had similar results for the analyses made; in the quartile analysis, the trend-corrected fared better when it came to predicting the quartiles of the most substantial flows, while the catboost was less error-prone on the other end of the spectrum. The TC predictor came out on top in terms of RMSE by about a 10% margin, but the catboost model proved to be best when it came to mitigating spatial dependence within the model residuals. It is unclear why the trend-corrected predictor fails to live up to the expectation of erradicating spatial dependence and, despite this poor performance in this metric, still managed a decent one in all other analyses. It seems that the choice for out-of-sample prediction is not 'cut and dried' and, at least for this study, is dependent on whether one wants to correctly identify the highest number of most substantial flows or minimize the number of incorrectly identified ones. For the former goal, the trend-corrected predictor was better and for the latter, the catboost model.

We attempted to predict cycling flows in regions not present in the training set to test whether cities with bountiful cycling data could be leveraged to help planners in cities without such data availability. We found that both spatial econometric interaction models as well as gradient boosted regression trees offer improvements over traditional transport models such as the gravity model (represented by its log-linear version in this study) for flow prediction and are accurate enough so as to provide useful information about regions' potential cycling flows. The OD matrix produced by these models can be

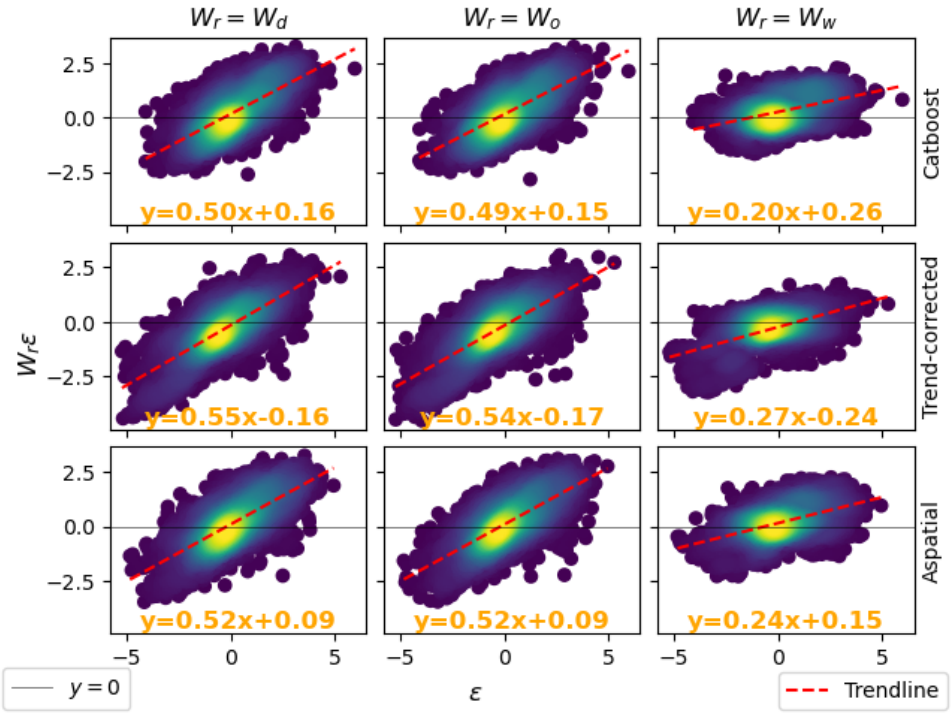

Figure 3. Moran scatter plot of residuals for out-of-sample prediction test.

used as is by planners or even serve as inputs for existing research which aims at aiding cycling planning. This data can also be fed into an existing routing algorithm (Google, Graphhopper, etc.) to compute the actual routes cyclists could take in these flows, thus allowing for more precise interventions when it comes to building cycling infrastructure.

7. Code availability

To enable reproducibility, all the code used for this research is available at `https://gitlab.com/interscity/bike-science` and `https://github.com/EduFalbel/seim`.

References

Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967.

Bera, S. and Rao, K. V. K. (2011). Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport Trasporti Europei*, (49):2–23.

Brand, C., Dons, E., Anaya-Boig, E., Avila-Palencia, I., Clark, A., de Nazelle, A., Gascon, M., Gaupp-Berghausen, M., Gerike, R., Götschi, T., Iacorossi, F., Kahlmeier, S., Laeremans, M., Nieuwenhuijsen, M. J., Orjuela, J. P., Racioppi, F., Raser, E., Rojas-Rueda, D., Standaert, A., Stigell, E., Sulikova, S., Wegener, S., and Panis, L. I. (2021). The climate change mitigation effects of daily active travel in cities. *Transportation Research Part D: Transport and Environment*, 93:102764.

Chai, D., Wang, L., and Yang, Q. (2018). Bike flow prediction with multi-graph convolutional networks. In *Proceedings of the 26th ACM SIGSPATIAL International*

Conference on Advances in Geographic Information Systems, SIGSPATIAL '18, page 397–400, New York, New York. Association for Computing Machinery.

Dargel, L. (2021). Revisiting estimation methods for spatial econometric interaction models. *Journal of Spatial Econometrics*, 2(10).

Félix, R., Cambra, P., and Moura, F. (2020). Build it and give 'em bikes, and they will come: The effects of cycling infrastructure and bike-sharing system in lisbon. *Case Studies on Transport Policy*, 8(2):672–682.

Goulard, M., Laurent, T., and Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: optimal and almost optimal strategies. *Spatial Economic Analysis*, 12(2-3):304–325.

Guidon, S., Reck, D. J., and Axhausen, K. (2020). Expanding a(n) (electric) bicycle-sharing system to a new city: Prediction of demand with spatial regression and random forests. *Journal of Transport Geography*, 84:102692.

Jiang, W. (2022). Bike sharing usage prediction with deep learning: a survey. *Neural Computing and Applications*, 34(18):15369–15385.

Kerkman, K., Martens, K., and Meurs, H. (2017). A multilevel spatial interaction model of transit flows incorporating spatial and network autocorrelation. *Journal of Transport Geography*, 60:155–166.

Kon, F., Ferreira, É. C., de Souza, H. A., Duarte, F., Santi, P., and Ratti, C. (2021). Abstracting mobility flows from bike-sharing systems. *Public Transport*, 14(3):545–581.

LeSage, J. P. and Pace, R. K. (2008). Spatial econometric modeling of origin-destination flows*. *Journal of Regional Science*, 48(5):941–967.

LeSage, J. P. and Thomas-Agnan, C. (2014). Interpreting spatial econometric origin-destination flow models. *Journal of Regional Science*, 55(2):188–208.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Ni, L., Wang, X. C., and Chen, X. M. (2018). A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. *Transportation Research Part C: Emerging Technologies*, 86:510–526.

Ortuzar, J. d. D. and Willumsen, L. G. (2011). *Modelling Transport*. Wiley, 4th edition.

Schatzmann, T., Sarlas, G., and Axhausen, K. W. (2019). Spatial modelling of origin-destination commuting flows in Switzerland. In *98th Annual Meeting of the Transportation Research Board (TRB 2019)*. Transportation Research Board.

Wang, M. and Zhou, X. (2017). Bike-sharing systems and congestion: Evidence from US cities. *Journal of Transport Geography*, 65:147–154.

Zhou, Y. and Huang, Y. (2018). Context aware flow prediction of bike sharing systems. In *2018 IEEE International Conference on Big Data (Big Data)*.

This research was supported by FAPESP, proc. 14/50937-1, 21/09285-4, and 22/06328-7.