Swiss Federal Institute of Technology

Institute for Transport Planning and Systems

Urban Bike Flow Prediction: Spatial Econometrics vs. Explainable Machine Learning

Final Scientific Report pertaining to Bolsa de Estágio de Pesquisa no Exterior, funded by Fundação de Amparo à Pesquisa do Estado de São Paulo.

FAPESP #2022/06328-7

Student : Eduardo Bobrow Falbel
External advisor: Kay W. Axhausen
Advisor : Fabio Kon
Co-advisor: Raphael Camargo

Zurich, August 29, 2023

General Project Information

- Project title:

  Urban Bike Flow Prediction:
  Spatial Econometrics vs. Explainable Machine Learning

- Author:

  Eduardo Bobrow Falbel

- Host institution:

  Institute for Transport Planning and Systems, Swiss Federal Institute of Technology

- Advisor, co-advisor and external supervisor:

  Fabio Kon

  Raphael Camargo

  Kay W. Axhausen

- Research project number:

  2022/06328-7

- Validity period:

  23/08/2022 - 23/12/2022

- Period covered by this scientific report:

  23/08/2022 - 28/02/2023

Contents

1 Project summary

Our goal throughout the BEPE fellowship was to implement a Spatial Econometric Interaction Model to predict cycling flows within an urban center and then compare its results against the ones of the machine learning models we had been using.

2 Summary of the work carried out

During the fellowship's validity period, we managed to use the previously obtained Boston bike-sharing data from the Blue Bikes[1] service alongside census, Google POI and elevation data in order to train/fit various models. These were comprised of a host of spatial interaction models and a Catboost boosted regression trees machine learning model. We then proceeded to test the predictive capabilities of all such models with both in-sample and out-of-sample tests. The former gave us an idea of how well the models fit to the data, while the latter, how much they were able to generalize in terms of the data domain. We then analyzed all of the prediction in order to try and understand each of the models' peculiarities, such as bias, variance, and spatial dependence within the residuals.

After the period of the Fellowship had expired, FAPESP also cleared me to stay with the group for another 7 months. During this period I focused on a new project, a diagnostic tool for the state of traffic simulations' calibration, while also working on the article for the previous research, which we have submitted to the Transportation Research Board Annual Conference.

3 Accomplishments throughout the period

The main accomplishment pertaining exclusively to the BEPE period was the production of a scientific article which was submitted to one of the biggest transportation conferences in the world, the TRB Annual Meeting. We will mostly touch on this result during this report. Besides this, however, we also produced a diagnostic tool[1] during the special allowance period after the fellowship's completion which is also extremely relevant and is already in use by the IVT research group.

3.1 Introduction

Cycling has the potential to mitigate critical problems faced by urban populations worldwide: its low CO_2-equivalent life-cycle emissions when compared to other forms of trans-

[1] https://www.bluebikes.com/

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

port [1] present a strategy to combat climate change; as a means of active travel, it promotes healthy living for its users and has the added benefit of lowering costs associated with healthcare [2]; finally, its use for short trips can replace car use and thus help mitigate traffic congestion in large cities [3]. Due to these and other reasons, city officials have been looking to encourage bike ridership. There are of course many ways to do that, such as educational and promotional campaigns and policy changes (e.g. congestion charging and alternate-day travel schemes, a practice common in Latin-American countries). Evidence shows that the most effective measures are qualitative improvements aiming at improving the subjective safety of cycling infrastructure [4, 5]. Nonetheless, cycling sharing schemes can play a significant role in increasing cycling rates as well [6].

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 Matrix (ODM), the acquisition of which is an issue that has been at the forefront of transport engineering for decades [7]. ODMs provide knowledge on all flows made between their Traffic Analysis Zones (TAZs), that is, the amount of trips from each physical region to every other region. With existing methods, however, one usually needs some sample of trip/flow data from the region of interest, such as travel surveys, link counts, bike-sharing data, etc., to estimate an ODM [7, 8]. Cities without an established cycling culture however, presumably do not have such data to begin with and attempts to collect it would either be too costly [7] or not yield enough information due to low ridership.

We propose models that can be trained on a region with a mature cycling culture, i.e., a region with available ridership and contextual (Points of Interest and socidodemographic) data, and then be used to predict flows/estimate an ODM for a region of interest which only has contextual data available, allowing transport planners to easily 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 works builds on [9], 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
- 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 will 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 cycling OD estimation, 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.

3.2 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 [7]. The most well-known of these is the gravity model, based on Newton's law of gravitation, specified by Equation 3.1. 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.

$$T_{ij} = \alpha O_i D_j f(c_{ij}) \tag{3.1}$$

This model also appears very frequently in one of its constrained variants, wherein we assign the totals for either or both origin and/or destination interactions (referred to as singly and doubly constrained variants, respectively), which we enforce with balancing factors [10]. Generally, estimation of the gravity model is done with Maximum Likelihood Estimation (MLE) on its log-linearized form (Equation 3.2).

$$\begin{aligned} \log\left(T_{ij}\right) &= \log\left(\alpha O_i D_j f(c_{ij})\right) \\ &= \log \alpha + \log O_i + \log D_j + \log f(c_{ij}) \end{aligned} \tag{3.2}$$

We can 'rebrand' this equation to make it look more like a regression specification for which one would use MLE, as shown in Equation 3.3 [11]. Now, α is the intercept and $\beta_i, i = 1, 2, 3$ are the coefficients we are trying to estimate.

$$y_{ij} = \alpha + \beta_1 O_i + \beta_2 D_j + \beta_3 f(c_{ij}) + \varepsilon \tag{3.3}$$

A major issue with this model, however, is that it assumes independence between flows [12], a premise that has been shown not to hold. Due to this limitation, [12] 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'll refer to as the LAG model, is given by Equation 3.4.

$$y = \rho_o W_o y + \rho_d W_d y + \rho_w W_w y + \beta_d X_d + \beta_o X_o + \gamma g + \alpha + \varepsilon \tag{3.4}$$

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 terms in the right-hand side of the equation ($\rho_o W_o y, \rho_d W_d y, \rho_w W_w y$), which are there to mitigate origin dependence, destination dependence, and origin-destination dependence, respectively. Each of these terms is composed of the dependent variable y, a spatial coefficient ρ which we attempt to estimate, and a spatial weights matrix $W_i, i = d, o, w$.

3.3 Related Work

Traditionally, ODM estimation is derived from information provided by surveys and roadside interviews, however their high cost was a motivator for the use of techniques that relied on data from automatic loop detectors [8], although recent developments allowed researchers to use mobile phone and GPS data as well. For the specific case of static ODM estimation, some of the most common methods employed are based on models such as the Gravity model (Equation 3.1) or Path Flow Estimation (PFE) models. Yet even with the increasing push from city officials to encourage cycling and other forms of sustainable mobility, few scholars have tackled the problem of Cycling ODM estimation from this angle; the most notable examples being [13] and [14], both of which incorporated PFE in their respective estimation procedures and were based on bike link count data. There are practical issues with these approaches, such as the availability of the needed data, as both surveys and automated equipment are expensive to conduct and install, respectively, and in cities still in their cycling infancy ridership is presumably low, meaning these data collection methods would not yield representative information.

The rise of bike-sharing systems has also motivated many researchers to attempt to predict cycling flows, usually for the purposes of system optimization [15]. Most of the literature focuses on the modelling of docked systems [15], in which there are predetermined, physical stations where users can retrieve and deposit bicycles. For this type of system, there are usually two main scenarios: forecasting the flow between individual stations or clusters of stations for some short time-frame [16, 17] for system rebalancing and predicting flows from and to possible new stations to determine the best location for such a station [18, 3].

Dockless bike-sharing systems, however, force scholars to take a different approach, which usually involves some sort of spatial aggregation such as rectangular grids [19, 20]. Inadvertently, the problem of bike-flow prediction in this case is the exact same as of ODM estimation and doing so based on this kind of data intuitively seems more accurate than the last method, since one already has all base flows after the initial process of spatial aggregation. Nonetheless, none of these studies have attempted to build a model and test its predictive capabilities on regions which are not present in the training sample.

[21] 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 tried applying spatial econometric interaction models to flows associated with mobility/commuting. Most notably, [22] applied it to flows abstracted from mobile phone data in Hangzhou, China; [23] used a simplified version of the model which did not include all three spatial weights matrices simultaneously on public transport (PT) commuting data in Switzerland; [24] used a multilevel approach to model PT flows in the Netherlands and used a SEIM as the upper level model for one of these approaches;

[25] modeled home-to-work commuting flows in Paris with SEIMs. Besides those, SEIMs have also been used to model migration flows in the US [12] and [26] use it to forecast commodity flows between Spanish regions. 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.

3.4 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 will discuss one of the models used during this project in depth.

3.4.1 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, whose location can be seen in Figure 3.1.

The process for abstracting the cycling trips into flows is based on the use of a regular grid of cells for trip aggregation. We chose this aggregation method because 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, since it only needs start- and endpoints. Also, the use of a regular grid instead of existing census tracts means that all TAZs have the exact same area and are "agnostic" in the sense that we can have the same kind of grid for whichever city/region we choose to model, which will facilitate future generalization efforts. For the Boston case study, we start off by creating a 20 X 20 regular grid of the city's metropolitan area, giving each square cell a side of about 650 meters. 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 3.1 shows the most substantial flows which, together, contain 25% of all trips made during the time period specified above. Since we are dealing with a station-based system in this case we 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.

From the steps above we obtain a table such that every row corresponds to a particular flow (origin-destination dyad with associated volume). It should be noted that flows are unique. We first enrich that dataframe with the distance between cells, calculated using the haversine method [27] between their centroids. We establish the distance for flows which start and end in the same cell, which we will henceforth refer to as intrazonal as opposed to interzonal, to be zero. We then add 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. Meaning, if a cell intersects 80% of a tract's area, we 'give' that cell 80% of the given population in the tract, for example. Finally, we incorporate cycling infrastructure data by means of the 'cycling infrastructure

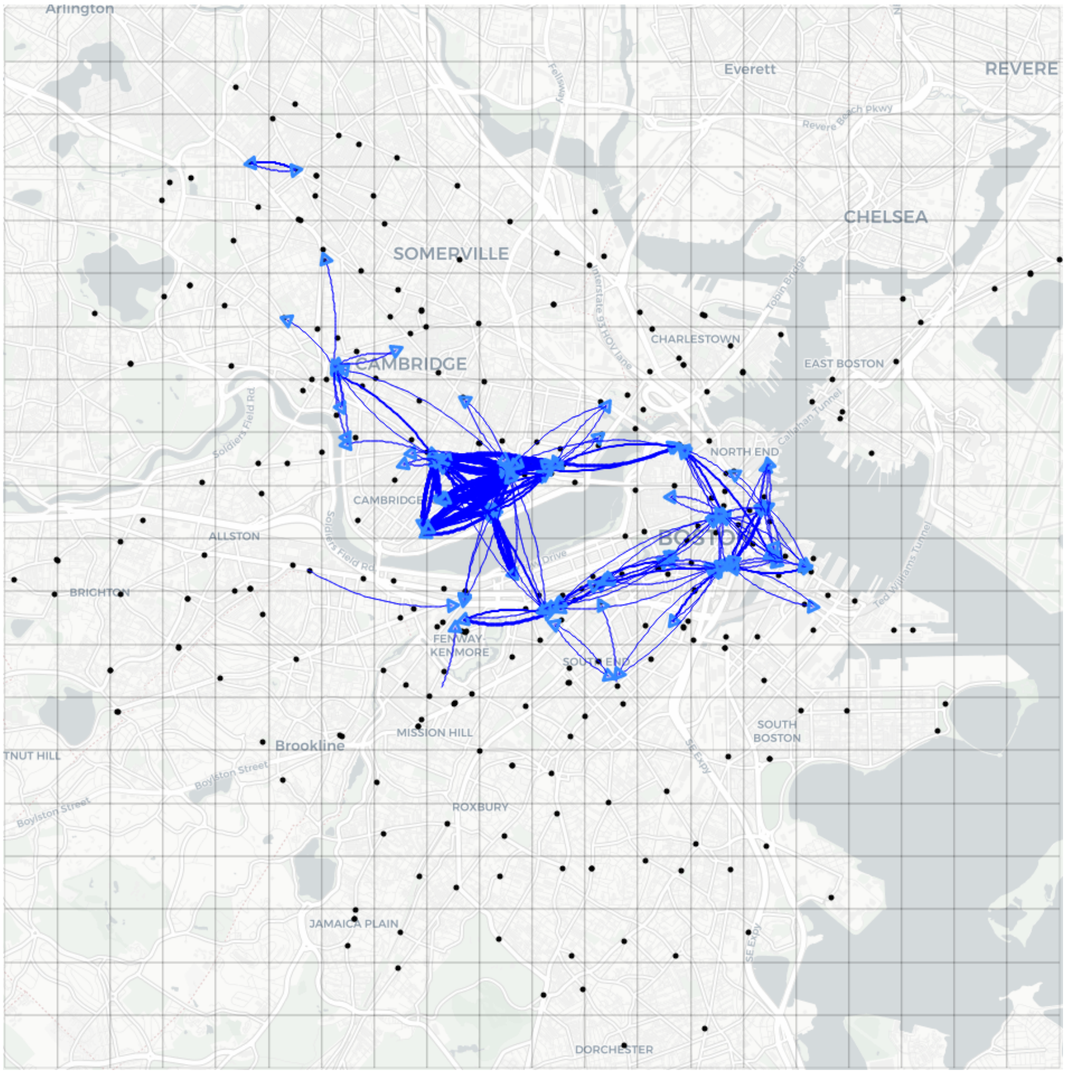

Figure 3.1: Flows belonging to the 1st quartile in terms of trip volume.

$$\text{Cycling Infrastructure Ratio} = \frac{length(\text{Cycling route} \cap \text{Cycling infrastructure})}{length(\text{Cycling route})} \tag{3.5}$$

ratio'. This is a number between 0 and 1 calculated by first geographically intersecting the 'cycling route' (obtained through the GraphHopper API) with the existing cycling infrastructure given by the city and then dividing the length of the result by the length of the route (Equation 3.5). This gives us the fraction of the route between cells that is covered by infrastructure such as bike lanes and paths.

For our models, we used the following features:

- University POIs
- Food POIs
- Mean income per capita
- Total population
- Distance (between cells' centroids)
- Ratio of cycling route covered by cycling infrastructure

All of the features above were log-transformed to comply with the theoretical basis of these models and standardized (subtracted the mean, then divided by the standard deviation) so that the estimated coefficients had similar orders of magnitude. We also applied the log function to the cycling flows so that they would better resemble a Normal distribution.

3.4.2 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 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 models. 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. For example, there should not be any "holes" such that a nearest neighbour is too far away from a region of interest. Basing ourselves on these principles, we chose the split in Figure 3.2, 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/estimated their parameter coefficients with the training data and performed two prediction tests: in-sample and out-of-sample. For the in-sample test, the models were trained with the yellow-shaded regions in Figure

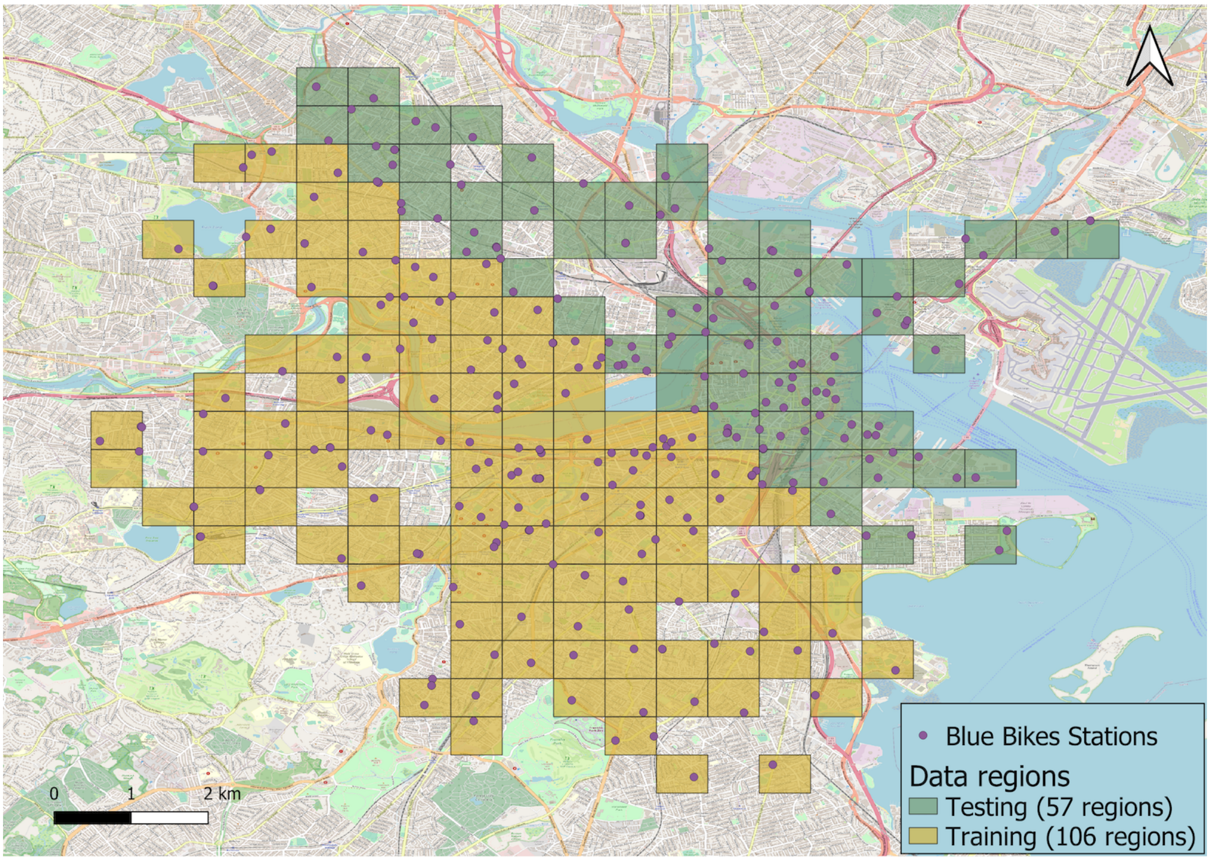

Figure 3.2: Training and testing split for the Boston metropolitan area

3.2 and tested on these same regions. In the out-of-sample test, the models were also trained with the yellow-shaded regions, but this time were tested with the green-shaded regions in Figure 3.2. 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. This way, we could compare the different characteristics of each predictor in terms of over-fitting and generalization potential.

3.4.3 Predictors

As can be seen from Equation 3.4, 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, based on the models' expected value, described in [11] (Equation 3.11). To arrive at that specification, we start by condensing Equation 3.4 into Equation 3.6.

$$y = \rho_d W_d y + \rho_o W_o y + \rho_w W_w y + \delta Z + \varepsilon, \tag{3.6}$$

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, resulting in Equation 3.7.

$$y - \rho_d W_d y - \rho_o W_o y - \rho_w W_w y = \delta Z + \varepsilon, \tag{3.7}$$

which is equal to Equation 3.8.

$$(I - \rho_d W_d - \rho_o W_o - \rho_w W_w) y = \delta Z + \varepsilon. \tag{3.8}$$

Now multiplying by the inverse:

$$\begin{aligned} (I - \rho_d W_d - \rho_o W_o - \rho_w W_w)^{-1} (I - \rho_d W_d - \rho_o W_o - \rho_w W_w) y = \\ = (I - \rho_d W_d - \rho_o W_o - \rho_w W_w)^{-1} (\delta Z + \varepsilon), \end{aligned} \tag{3.9}$$

is finally equal to

$$y = (I - \rho_d W_d - \rho_o W_o - \rho_w W_w)^{-1} (\delta Z + \varepsilon). \tag{3.10}$$

We then craft our predictor as Equation 3.11, in accordance with [11].

$$\hat{y} = (I - \hat{\rho}_d W_d - \hat{\rho}_o W_o - \hat{\rho}_w W_w)^{-1} \hat{\delta} Z, \tag{3.11}$$

[28] describe various predictors for spatial econometric models which [25] used to implement some of their interaction counterparts. Keeping with the convention set forth

in [28] and [25], we will henceforth refer to Equation 3.11 above as the 'trend-corrected' predictor. We will compare the performance of this predictor with three others: the 'aspatial' model (Equation 3.13), which is the log-linear version of the gravity model (it can also be interpreted as a restricted version of the LAG model wherein we set $\rho_d = \rho_o = \rho_w = 0$); a Catboost gradient boosted tree regressor; and the interaction version of the 'trend-signal-noise' predictor [29, 30], as implemented in the R 'spflow' package [25]. It should be noted that the trend-corrected and trend-signal-noise predictors both have the same underlying model - LAG interaction model (Equation 3.4), meaning their estimated coefficients are also the same, but differ in regards to the specification used for prediction. This is important, since the trend-signal-noise predictor is not suitable for out-of-sample prediction (where the testing set is spatially disjoint from the training set), since it depends explicitly on the values of the dependent variable (which are unknown in the case of an out-of-sample test), as can be seen in Equation 3.12. As such, it was only used for in-sample prediction.

$$\hat{y} = \hat{\rho}_o W_o y + \hat{\rho}_d W_d y + \hat{\rho}_w W_w y + \hat{\delta} Z \tag{3.12}$$

$$\hat{y} = \hat{\beta}_d X_d + \hat{\beta}_o X_o + \hat{\gamma} g + \hat{\alpha} + \varepsilon \tag{3.13}$$

Unlike the LAG model, the aspatial model specification in Equation 3.13 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. All interaction models (aspatial and LAG were trained using the 'spflow' package), while the Catboost model was trained and tested using the 'catboost' Python package. Testing of the aspatial and trend-signal-noise predictors was done with the spflow package as well, while testing of the trend-corrected predictor was done using the Python package we have developed[2].

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' [31]. The only hyperparameter we tuned was the trees' depth, which we set to one. This was based on the notion that 'stumps' can provide better generalization capabilities than other tree depths [32].

Finally, we proceeded to analyze the predictions for each model and compare them against each other. In total, we tested 4 different predictors on the Boston bike-sharing system data for in-sample prediction and 3 for out-of-sample prediction, since the trend-signal predictor cannot be used for such purposes, as mentioned previously. They are:

- CatboostRegressor (CB)
- Aspatial predictor (A)
- LAG trend-corrected predictor (TC)
- LAG trend-signal-noise predictor [28] (TS)

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

We used the same weights matrix specification for all the spatial econometric interaction models: a row-standardized weights matrix without distance decay based on 8-nearest neighbors. We chose this variant as the grid nature of the regions meant that most would be contiguous with 8 other cells, however had we chosen to base the neighborhood on contiguity, some regions would end up without neighbours (due to discontinuities which can be seen in Figure 3.2).

3.5 Results

We'll start by examining the Root Mean Squared Error (RMSE) and then dive into various other metrics we used to gauge the quality of the models' predictive abilities. We performed some more in-depth analyses of the models' predictions, such as investigating the predicted vs. observed flow quartiles and some spatial dependency analysis. We calculated these metrics and performed these analyses for both the in-sample and out-of-sample tests so that we could get some more insights into the predictions of each model.

3.5.1 Prediction accuracy

A common metric to evaluate the prediction accuracy of statistical and machine learning models is the Root Mean Squared Error (Equation 3.14). Table 3.1 shows this metric for each model, calculated for both the in-sample and out-of-sample prediction tests (except for the trend-signal model, since it cannot be used for out-of-sample prediction, as mentioned earlier).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{3.14}$$

Table 3.1: RMSE scores for the models' In- and Out-of-sample predictions.

Model	In-sample	Out-of-sample
Aspatial	192.542	268.204
Catboost	174.022	255.014
Trend-signal	181.611	-
Trend-corrected	190.693	229.366

The Catboost model had the lowest - therefore, the best - RMSE score of any of the models in the in-sample test, followed by the trend-signal, trend-corrected, and aspatial predictors, respectively (3.1). What is very interesting, however, 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 [33], the SEI models allow for a much clearer interpretation of the influence of spatial interaction effects on the cycling flows.

3.5.2 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 in both types of tests (in- and out-of-sample), which can be seen in Tables 3.2 to 3.5. 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 3.2: In-sample analysis for Trend-signal model.

Predicted	0	1	2	3	All
Observed					
0	28	22	2	0	52
1	7	90	77	3	177
2	1	35	306	178	520
3	0	4	98	10385	10487
All	36	151	483	10566	11236

Table 3.3: Catboost analysis

	In-sample					Out-of-sample				
Predicted	0	1	2	3	All	0	1	2	3	All
Observed										
0	28	20	4	0	52	6	14	12	5	37
1	38	71	61	7	177	4	13	33	26	76
2	36	141	242	101	520	5	11	65	93	174
3	6	115	681	9685	10487	1	13	108	2840	2962
All	108	347	988	9793	11236	16	51	218	2964	3249

Starting with the in-sample test, we can see that the Catboost and Trend-signal models (Tables 3.3 and 3.2, respectively) had the highest share of correctly predicted tier 0 flows (the flows with the highest volume), followed by the trend-corrected predictor,

Table 3.4: Trend-corrected model analysis.

	In-sample					Out-of-sample				
Predicted	0	1	2	3	All	0	1	2	3	All
Observed										
0	11	19	22	0	52	12	11	10	4	37
1	34	30	101	12	177	9	25	28	14	76
2	43	102	230	145	520	12	22	70	70	174
3	25	210	890	9362	10487	7	33	163	2759	2962
All	113	361	1243	9519	11236	40	91	271	2847	3249

Table 3.5: Aspatial model analysis.

	In-sample					Out-of-sample				
Predicted	0	1	2	3	All	0	1	2	3	All
Observed										
0	9	21	21	1	52	4	9	13	11	37
1	16	47	80	34	177	2	8	37	29	76
2	23	86	220	191	520	4	9	56	105	174
3	15	136	770	9566	10487	2	6	86	2868	2962
All	63	290	1091	9792	11236	12	32	192	3013	3249

and finally the aspatial predictor. On the other end of the spectrum, however, we can see that the Trend-signal predictor fared better than even the Catboost one in regards to incorrectly predicting observed tier 3 flows (lowest volume) as tier 0 flows. This quality could be the most relevant, since one could consider it worse to spend money on incredibly small flows than to not spend it on substantial ones. For the rest of the in-sample test, we generally have $TS > CB > A > TC$ in order of best to worst, meaning the trend-signal predictor was best able to fit the training data, while the trend-corrected predictor was the worst of the bunch.

In the out-of-sample test (Table 3.4), we can observe that the trend-corrected predictor had the highest shares of correctly predicted tier 0, tier 1, and tier 2 flows, while it was the Catboost followed by the Aspatial predictors that had the lowest share of tier 3 flows incorrectly predicted as tier 0 (Tables 3.3 and 3.5, respectively). However, this last observation seems to be due to the models' general tendency to underpredict flows in this test, as opposed to some other inherent quality.

3.5.3 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 in- and out-of-sample tests (Images Figs. 3.3 and 3.4, respectively), 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. For the in-sample test (Figure 3.3), we can see how well the Trend-signal predictor performs, displaying almost no spatial dependence within the residuals. Surprisingly, however, the next best model in that regard is the Catboost one, followed by the Trend-corrected and Aspatial predictors in shared third place. This was unexpected, not only because of how well the Trend-signal predictor managed to deal with spatial dependence, but of how the Trend-corrected failed to do the same, event though they are based on the same underlying spatial lag model. This exemplifies the trade-off between these predictors: the former allows for a much better fit, but is not capable of performing predictions for spatial units not in the training sample.

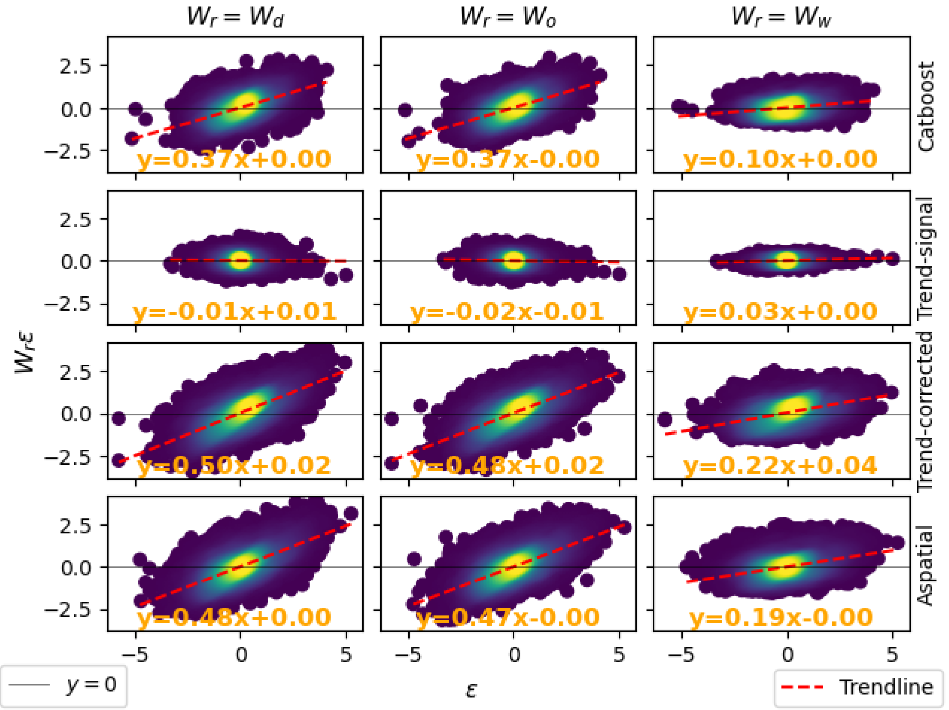

Figure 3.3: Moran scatter plot of residuals for in-sample prediction test.

When examining the residuals of the out-of-sample prediction test, however, all three of the models display similar levels of spatial dependence, as can be seem by the comparably angled trendlines in Figure 3.4. Even then, however, the catboost model was the best out of all three, even though it was not designed with the intent of dealing with spatial dependence, unlike the trend-corrected.

3.6 Discussion and Conclusions

The in-sample tests revealed the unrivaled predictive capabilities of the trend-signal predictor as evidenced mainly by the quartile analysis, in which this predictor was the only

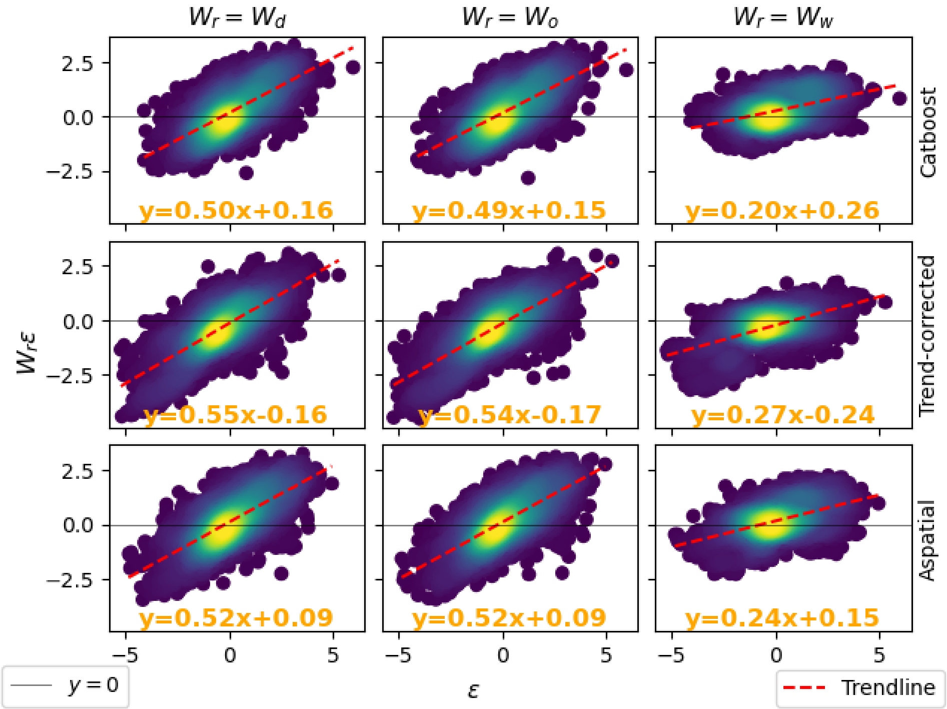

Figure 3.4: Moran scatter plot of residuals for out-of-sample prediction test.

one which did not mistakenly assign any of the least substantial flows as one of the most substantial. This quality might be the most appreciated among city officials, since they can have a higher degree of certainty that new infrastructure will be fully utilized. When analysing the Moran scatterplots, this test also seemingly confirmed that spatial dependence does indeed play an important role within cycling flows and that this predictor was the one most suitable to handle it. Considering both of these results, it would be safe to say that, for the particular train-test-split used, the assumptions governing the underlying spatial econometric interaction model appear to hold. Conversely, the trend-corrected predictor was one of, if not the worse performer, demonstrating that, for these tests, the ability to perform spatial-out-of-sample predictions incurs the loss of optimality properties which greatly increase the in-sample predictive potential of the underlying model.

When it came to the out-of-sample test, the catboost and trend-corrected predictors were similar for the comparisons; 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, which might be more relevant in the context of public policies, since it would mean avoiding potentially unnecessary spending of taxpayer money. The former 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 performing similarly to how the trend-signal model did in the in-sample test and, despite this poor showing in this metric, still managed a decent showing in all other analyses. Finally, for the presented case study, it would seem that the trend-signal model is the most appropriate when the focus is purely on in-sample prediction, while the choice for out-of-sample 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 was better and the catboost, for the latter. The log-linear gravity model trailed the other models in almost all comparisons in both in-sample and out-of-sample tests leading us to believe that its main, and possibly only, advantage is in terms of ease of use and interpretation.

This work presents a starting point for bike-sharing systems based out-of-sample cycling flow prediction. However, more work can be done, especially when it comes to the predictive power of the spatial econometric interaction models for out-of-sample tests. For example, the use of more optimal out-of-sample predictors, akin to the ones in [28]. Next, our base data coming from a fixed-station bike-sharing system means that it has inherent bias, since the locations of those stations were not determined by some underlying organic characteristic of trip production and demand, but were chosen by people based on their notions about whether having such a station would be best for the company. Thus, it is impossible for the resulting O-D matrix to faithfully reproduce 'ideal' trip patterns (that is, those that would be done if infinite bicycles were available everywhere). Finally, to completely measure the generalization capabilities of the models, one would need to conduct tests using data from various cities. That is, training the models with data from one or more cities and then testing their predictions on cities not present in the training set.

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 used as is by planners or even serve as inputs for existing research which aims at aiding cycling planning. This data can be fed into an approach like jittering [34] or some routing algorithm (Google, Graphhopper, etc.) to try and calculate the actual routes cyclists would take in these flows, thus allowing for more precise interventions when it comes to building cycling infrastructure, for example.

4 Description and evaluation of the Institutional Support

FAPESP's financial support was indispensable for the completion of this project. Besides that, the "Converse com a FAPESP" service was extremely helpful when dealing with the funding's bureaucreacy.

The Institute for Transport Planning and Systems' financial, computing, and human resources made the student's stay with the group not only possible, but extremely productive and enjoyable.

The Computer Science Department of the University of São Paulo was also very supportive, providing access to remote computing services which the student could use to train and test the developed models.

5 Participation in scientific events

During the period from May 10-12, I participated in the Swiss Transportation Research Conference[1], presenting the work carried out throughout the Fellowship period, as well as being chair of one of the presentation sessions.

6 Planned activities for the next period

Unfortunately, since we had to develop so much tooling from scratch for the spatial econometric interaction models, we were unable to test the models in cities besides Boston. However, the work done on the simulation diagnostic tool opened up new research avenues which we are also keen on exploring. Our main goals for the next period are to work on the feedback received from the TRB Annual Conference reviewers which should come in the beginning of October so that we can later submit the improved work to the Journal of Transport Geography, where it will hopefully be published and yield another paper. Besides this, the IVT group has shown interest in extending our academic collaboration, and so we are also aiming to leverage the diagnostic tool we created to research auto-calibration tools for the MATSim traffic simulation software[1]. Our goal with this part of the research is to produce a paper based on a comprehensive survey of the current methods/algorithms as well as a comparison of these when applied to the IVT's upcoming E-Bike City[2] model.

[1] `https://www.strc.ch/2023.php`

[1] https://www.matsim.org/

[2] `https://ebikecity.baug.ethz.ch/en/`

Bibliography

[1] BRAND, C. et al. The climate change mitigation effects of daily active travel in cities. *Transportation Research Part D: Transport and Environment*, Elsevier BV, v. 93, p. 102764, 2021. Disponível em: <https://doi.org/10.1016/j.trd.2021.102764>.

[2] GöTSCHI, T. Costs and benefits of bicycling investments in Portland, Oregon. *Journal of Physical Activity and Health*, Human Kinetics, v. 8, n. s1, p. S49–S58, 2011. Disponível em: <https://doi.org/10.1123/jpah.8.s1.s49>.

[3] WANG, M.; ZHOU, X. Bike-sharing systems and congestion: Evidence from US cities. *Journal of Transport Geography*, Elsevier BV, v. 65, p. 147–154, 2017. Disponível em: <https://doi.org/10.1016/j.jtrangeo.2017.10.022>.

[4] GOEL, R. et al. Cycling behaviour in 17 countries across 6 continents: levels of cycling, who cycles, for what purpose, and how far? *Transport Reviews*, Informa UK Limited, v. 42, n. 1, p. 58–81, 2021. Disponível em: <https://doi.org/10.1080/01441647.2021.1915898>.

[5] NELLO-DEAKIN, S. Environmental determinants of cycling: Not seeing the forest for the trees? *Journal of Transport Geography*, Elsevier BV, v. 85, p. 102704, 2020. Disponível em: <https://doi.org/10.1016/j.jtrangeo.2020.102704>.

[6] FÉLIX, R.; CAMBRA, P.; MOURA, F. 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*, Elsevier BV, v. 8, n. 2, p. 672–682, 2020. Disponível em: <https://doi.org/10.1016/j.cstp.2020.03.002>.

[7] ORTUZAR, J. d. D.; WILLUMSEN, L. G. *Modelling Transport*. 4. ed. [S.l.]: Wiley, 2011.

[8] BERA, S.; RAO, K. V. K. Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport Trasporti Europei*, n. 49, p. 2–23, 2011. Disponível em: <https://ideas.repec.org/a/sot/journl/y2011i49p2-23.html>.

[9] KON, F. et al. Abstracting mobility flows from bike-sharing systems. *Public Transport*, Springer Science and Business Media LLC, v. 14, n. 3, p. 545–581, 2021. Disponível em: <https://doi.org/10.1007/s12469-020-00259-5>.

[10] GRIFFITH, D. A.; FISCHER, M. M. Constrained variants of the gravity model and spatial dependence: model specification and estimation issues. *Journal of Geographical Systems*, Springer Science and Business Media LLC, v. 15, n. 3, p. 291–317, 2013. Disponível em: <https://doi.org/10.1007/s10109-013-0182-7>.

[11] LESAGE, J. P.; THOMAS-AGNAN, C. Interpreting spatial econometric origin-destination flow models. *Journal of Regional Science*, Wiley, v. 55, n. 2, p. 188–208, 2014. Disponível em: <https://doi.org/10.1111/jors.12114>.

[12] LESAGE, J. P.; PACE, R. K. Spatial econometric modeling of origin-destination flows*. *Journal of Regional Science*, v. 48, n. 5, p. 941–967, 2008. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9787.2008.00573.x>.

[13] RYU, S. A bicycle origin–destination matrix estimation based on a two-stage procedure. *Sustainability*, MDPI AG, v. 12, n. 7, p. 2951, 2020. Disponível em: <https://doi.org/10.3390/su12072951>.

[14] LIU, Z.; SONG, Z. *A Constraint-Based Bicycle Origin-Destination Estimation Procedure*. [S.l.], 2019.

[15] JIANG, W. Bike sharing usage prediction with deep learning: a survey. *Neural Computing and Applications*, Springer Science and Business Media LLC, v. 34, n. 18, p. 15369–15385, 2022. Disponível em: <https://doi.org/10.1007/s00521-022-07380-5>.

[16] ZHOU, Y.; HUANG, Y. Context aware flow prediction of bike sharing systems. In: *2018 IEEE International Conference on Big Data (Big Data)*. [S.l.: s.n.], 2018.

[17] CHAI, D.; WANG, L.; YANG, Q. Bike flow prediction with multi-graph convolutional networks. In: *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, New York: Association for Computing Machinery, 2018. (SIGSPATIAL '18), p. 397–400. Disponível em: <https://doi.org/10.1145/3274895.3274896>.

[18] BEAIRSTO, J. et al. Identifying locations for new bike-sharing stations in glasgow: an analysis of spatial equity and demand factors. *Annals of GIS*, Informa UK Limited, v. 28, n. 2, p. 111–126, 2021. Disponível em: <https://doi.org/10.1080/19475683.2021.1936172>.

[19] XU, C.; JI, J.; LIU, P. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Research Part C: Emerging Technologies*, v. 95, p. 47–60, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0968090X18306764>.

[20] LI, Y.; SHUAI, B. Origin and destination forecasting on dockless shared bicycle in a hybrid deep-learning algorithms. *Multimedia Tools and Applications*, Springer Science and Business Media LLC, v. 79, n. 7-8, p. 5269–5280, 2018. Disponível em: <https://doi.org/10.1007/s11042-018-6374-x>.

[21] GUIDON, S.; RECK, D. J.; AXHAUSEN, K. Expanding a(n) (electric) bicycle-sharing system to a new city: Prediction of demand with spatial regression and random forests. *Journal of Transport Geography*, Elsevier BV, v. 84, p. 102692, 2020. Disponível em: <https://doi.org/10.1016/j.jtrangeo.2020.102692>.

[22] NI, L.; WANG, X. C.; CHEN, X. M. A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. *Transportation Research Part C: Emerging Technologies*, v. 86, p. 510–526, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0968090X17303601>.

[23] SCHATZMANN, T.; SARLAS, G.; AXHAUSEN, K. W. Spatial modelling of origin-destination commuting flows in Switzerland. In: *98th Annual Meeting of the Transportation Research Board (TRB 2019)*. Transportation Research Board, 2019. Disponível em: <http://hdl.handle.net/20.500.11850/283769>.

[24] KERKMAN, K.; MARTENS, K.; MEURS, H. A multilevel spatial interaction model of transit flows incorporating spatial and network autocorrelation. *Journal of Transport Geography*, Elsevier BV, v. 60, p. 155–166, 2017. Disponível em: <https://doi.org/10.1016/j.jtrangeo.2017.02.016>.

[25] DARGEL, L. Revisiting estimation methods for spatial econometric interaction models. *Journal of Spatial Econometrics*, Springer Science and Business Media LLC, v. 2, n. 10, 2021. Disponível em: <https://doi.org/10.1007/s43071-021-00016-1>.

[26] LESAGE, J. P.; LLANO-VERDURAS, C. Forecasting spatially dependent origin and destination commodity flows. *Empirical Economics*, Springer Science and Business Media LLC, v. 47, n. 4, p. 1543–1562, 2014. Disponível em: <https://doi.org/10.1007/s00181-013-0786-2>.

[27] SINNOTT, R. W. Virtues of the haversine. *Sky and Telescope*, v. 68, n. 2, p. 158–159, 1984.

[28] GOULARD, M.; LAURENT, T.; THOMAS-AGNAN, C. About predictions in spatial autoregressive models: optimal and almost optimal strategies. *Spatial Economic Analysis*, Informa UK Limited, v. 12, n. 2-3, p. 304–325, 2017. Disponível em: <https://doi.org/10.1080/17421772.2017.1300679>.

[29] HAINING, R. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, 1990. Disponível em: <https://doi.org/10.1017/cbo9780511623356>.

[30] BIVAND, R. Spatial econometrics functions in r: Classes and methods. *Journal of Geographical Systems*, Springer Science and Business Media LLC, v. 4, n. 4, p. 405–421, 2002. Disponível em: <https://doi.org/10.1007/s101090300096>.

[31] BENTÉJAC, C.; CSöRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, Springer Science and Business Media LLC, v. 54, n. 3, p. 1937–1967, 2020. Disponível em: <https://doi.org/10.1007/s10462-020-09896-5>.

[32] JAMES, G. et al. *An Introduction to Statistical Learning: With Applications in R*. [S.l.]: Springer Publishing Company, Incorporated, 2014.

[33] LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777.

[34] LOVELACE, R.; FéLIX, R.; CARLINO, D. Jittering: A computationally efficient method for generating realistic route networks from origin-destination data. *Findings*, Findings Press, p. 33873, 2022.