Quantifying the effect of factors on bike-sharing demand: A regression model with spatially varying coefficients (21-00928)

Xudong Wang[a], Zhanhong Cheng[a], Martin Trépanier[b], Lijun Sun[a,†]

[a] Department of Civil Engineering, McGill University, Montreal, Quebec H3A 0C3, Canada

[b] Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Montreal, Quebec, H3T 1J4, Canada

Introduction

- As an emerging mobility service, bike-sharing has become increasingly popular around the world. A critical question in planning and designing bike-sharing services is to know **how different land-use and built environment factors affect bike-sharing demand.**
- Most existing research investigated this problem from a holistic view by regression models, where the regression coefficients are the same for all bicycle stations. **However, a global regression model essentially ignores the local spatial effects of different factors.**
- In this paper, we develop a **regression model with spatially varying coefficients** to investigate how land use attributes, social-demographic, and transportation infrastructure affect the bike-sharing demand at different stations. **The regression coefficients in the model are station-specific and regularized by a graph structure that encourages nearby stations to have similar coefficients.**
- We apply the model to the station-level bike-sharing demand data from the BIXI service in Montreal, Canada. We find that the obtained regression coefficients demonstrate clear spatially-varying patterns.

Bike-sharing Demand and Influential Factors

Bike-sharing demand

- BIXI, the first large-scale bike-sharing system in North America, is located in Montreal, Canada. In 2019, there are **615 stations** and 5.6 million trip records in the BIXI system.
- For most stations, bicycle demand peaks at two periods in a weekday—**6:00-10:00 am** and **3:00-7:00 pm**. The spatial distributions of bicycle demand are different in these two periods.
- Fig. 1 shows the average hourly departure demand distributions for morning and afternoon peaks, where each BIXI station is represented by a catchment area with 250 meters radius, and **Thiessen polygon** is used to determine the boundary between overlapping catchment areas

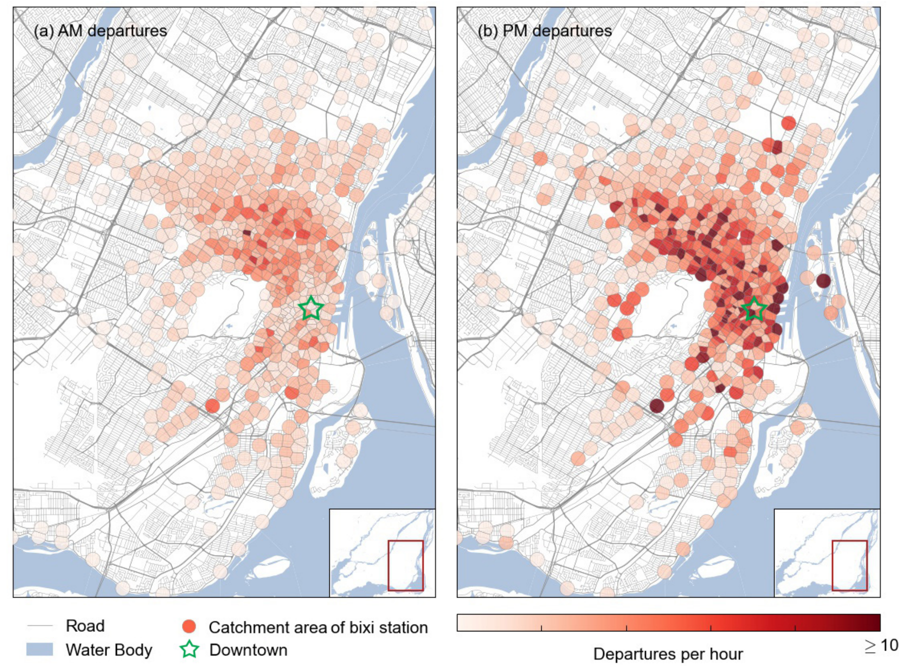

Fig. 1 Average hourly departures for BIXI stations in Montreal (left: morning peak, right: afternoon peak).

Influential factors

TABLE 1 : Summary of independent variables (aggregated by catchment areas).

Factors	Description	Min	Max	Mean
Population	Total residential population.	0	4042	1073.88
POI	The number of commercial POI (e.g., wholesale trade).	0	89	14.94
Park	The proportion of parks in the catchment area.	0	1	0.08
University	Binary variable for university.	0	1	0.013
Road length	Total length of roads (in meters).	133.27	4824.40	1934.93
Walkscore	A measure of walkability on a scale from 0 to 100.	7	99	79.28
Metro	Binary variable for metro station.	0	1	0.13
Bus route	log(number of bus routes + 1).	0	3.3	1.281
Cycle path	The proportion of roads with cycle path.	0	1	0.227567
Capacity	The number of docks of a BIXI station.	11	105	23.1

Regression with Graph regularization

Linear regression model

- For station i ($i \in \{1, \ldots, N\}$), y_i denotes its bike-sharing demand, which is the dependent variable. Denote a vector of independent variables by $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{im}]^\top$ for station i, where m is the number of factors. The regression model with spatially varying coefficients for all stations is described as:

$$\min \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_i\right)^2. \tag{1}$$

where $\boldsymbol{\beta}_i = [\beta_{i0}, \beta_{i1}, \ldots, \beta_{im}]^\top$ is a coefficient vector for station i, ε_i is the error term.

- If coefficient vectors $\boldsymbol{\beta}_i$ are the same for all stations, Eq. (1) is the least square problem. However, we study the factors' effect at a station-level and assume coefficient vectors are varying over stations. In this situation, we can find many sets of $\boldsymbol{\beta}_i$ that minimizes Eq. (1) (to zero), which is not our expected.

Linear regression with graph regularization

- A fundamental assumption in modeling the spatial effects is that **nearby stations have similar coefficients.** Following this assumption, we introduce a graph structure into the linear regression model.
- The linear regression model for each station with a graph regularization term, which penalizes the difference between $\boldsymbol{\beta}_i$ in adjacent nodes, is proposed in:

$$\min \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_i\right)^2 + \lambda \sum_{(i,j)\in\mathcal{E}} w_{ij} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2, \quad \lambda \geq 0, \tag{2}$$

where the parameter λ balances the regression error and the difference between coefficients in adjacent nodes, w_{ij} is the weight of the edge (i, j), which is a decaying function of distance.

- Eq. (2) is a convex optimization problem and can be efficiently solved by the commonly-used optimization software, such as CVXPY. For large-scale problems, the Alternating Direction Method of Multipliers (ADMM) can be used to solve the problem in a distributed and scalable manner.
- Hyper-parameter tuning The regression with graph regularization model has three hyper-parameters, λ determines the intensity of the graph regularization term, K is the number of neighbors of a node, and α controls the weight decaying of edges. We use 10-fold cross-validation to search the optimal values of hyper-parameters.

Applications

- Spatially varying coefficients The solution $\boldsymbol{\beta}_i$ for each station i in Eq. (2) will be station-based parameters, which describe the heterogeneous effects of influential factors on bike-sharing demand.
- Prediction new stations For a new station p ($p \notin \{1, \ldots, N\}$) with unknown bike-sharing demand, we can estimate its factor coefficients $\boldsymbol{\beta}_p$ by interpolating the coefficients $\boldsymbol{\beta}^\star$ from known stations and then achieve demand prediction of station p. Let Nei(p) denote the K nearest stations of station p, and we want to find a $\boldsymbol{\beta}_p$ that minimizes the difference between the coefficients of neighbors, leading to the following optimization:

$$\min \sum_{i\in \mathrm{Nei}(p)} w_{pi} \|\boldsymbol{\beta}_p - \boldsymbol{\beta}_i^\star\|_2^2. \tag{3}$$

Results

The effect of factors on bike-sharing demand

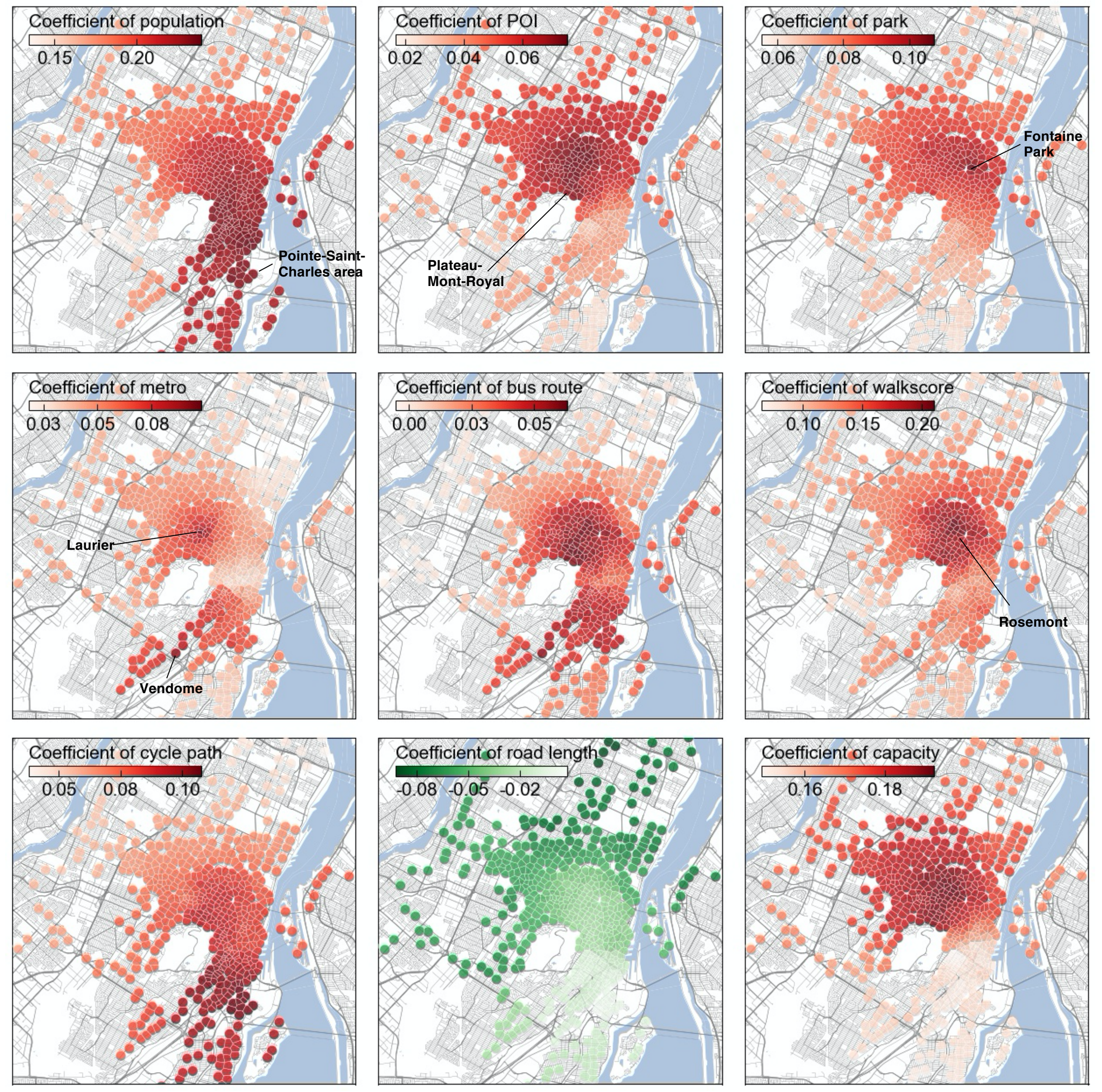

Fig. 2 The factor coefficients distribution of morning peak with optimal parameters $\lambda = 2$, $\alpha = 1$ and $K = 4$.

Out-of-sample prediction

TABLE 2 : The RMSE and R^2 in the prediction problem.

	AM departures		PM departures	
	RMSE×10	R^2	RMSE×10	R^2
Linear regression	1.041 (0.117)	0.300 (0.066)	1.016 (0.096)	0.494 (0.070)
Random forest	0.970 (0.119)	0.392 (0.075)	1.071 (0.147)	0.443 (0.080)
SVM regression	0.995 (0.126)	0.362 (0.070)	1.053 (0.140)	0.461 (0.076)
KNN	1.032 (0.110)	0.310 (0.072)	1.124 (0.152)	0.390 (0.061)
Nearest neighbors average	0.938 (0.122)	0.431 (0.077)	1.209 (0.169)	0.294 (0.083)
Regression kriging	**0.809 (0.101)**	**0.567 (0.050)**	0.956 (0.106)	0.535 (0.078)
GWR	0.854 (0.103)	0.514 (0.081)	0.959 (0.100)	0.525 (0.121)
Graph regularization (circle buffer)	0.871 (0.111)	0.510 (0.063)	1.014 (0.107)	0.498 (0.068)
Graph regularization (proposed)	0.831 (0.104)	0.554 (0.051)	**0.949 (0.096)**	**0.560 (0.059)**

Acknowledgement This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Mitacs Canada, and the Canada Foundation for Innovation (CFI). X. Wang would like to thank Fonds de recherche du Quebec - Nature et technologies (FRQNT) for providing scholarship. We also thank BIXI Montreal for providing data.