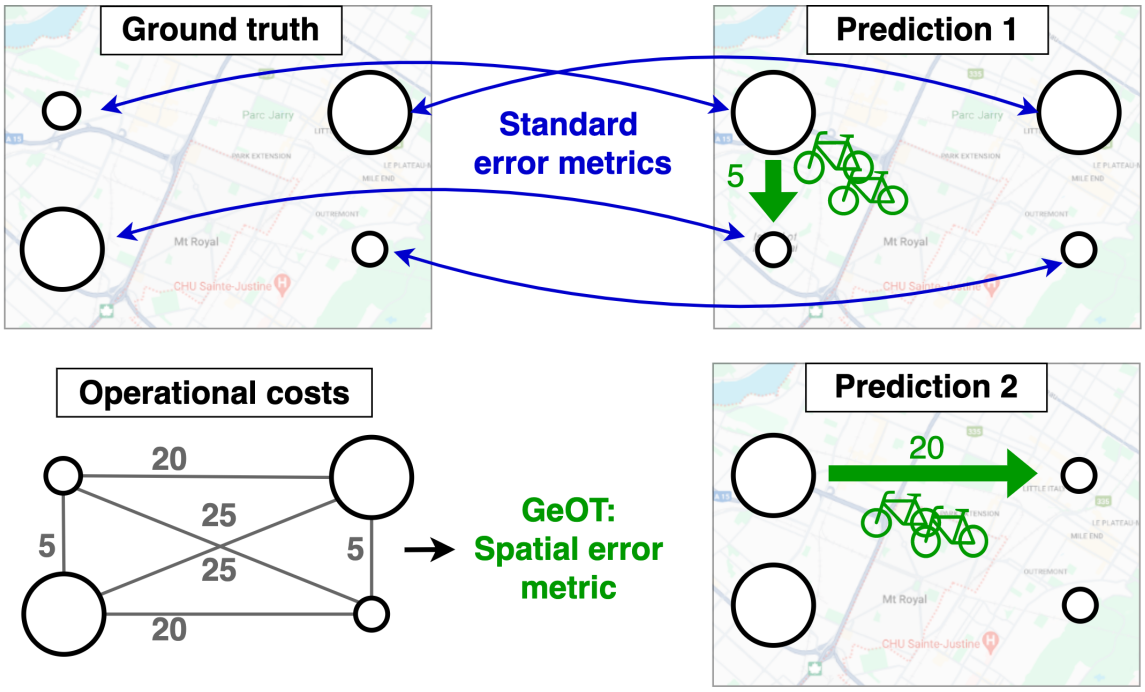


Bridging prediction and optimization in on-demand transportation systems with Optimal Transport

Nina Wiedemann

Théo Uscidda

Martin Raubal

STRC Conference Paper 2025 **April 28, 2025**

STRC | **25th Swiss Transport Research Conference**
Monte Verità / Ascona, May 14-16, 2025

Bridging prediction and optimization in on-demand transportation systems with Optimal Transport

Nina Wiedemann
Chair of Geoinformation Engineering
ETH Zurich
nwiedemann@ethz.ch

Théo Uscidda
CREST-ENSAE
Paris, France
theo.uscidda@gmail.com

Martin Raubal
Chair of Geoinformation Engineering
ETH Zurich
mraubal@ethz.ch

April 28, 2025

Abstract

Predictions for on-demand transportation services are oftentimes motivated by the possibility to enhance operational efficiency. For example, bike-sharing demand prediction aids in relocation planning. However, the prediction accuracy is usually evaluated with standard metrics such as the root-mean-squared-error (RMSE), which fall short in assessing the value of predictions for downstream tasks. Since standard metrics treat spatial locations independently, they disregard the costs stemming from the spatial displacement of the predicted demand, such as relocation costs. We put forward Optimal Transport (OT) as a spatial evaluation metric and loss function to bridge the gap between prediction and optimization in transport applications. The proposed framework, GeOT, evaluates prediction models by quantifying the transport costs associated with their prediction errors. Through case studies on bike sharing data, we show that 1) OT better captures spatial costs than existing metrics, 2) OT enhances comparability across spatial and temporal scales, and 3) using OT as a loss function effectively reduces spatial costs. The method is broadly applicable to spatiotemporal prediction tasks, and we provide an open-source Python package for seamless adoption (`https://github.com/mie-lab/geospatialOT`)

Keywords

GeoAI; spatio-temporal modelling; evaluation framework

Preferred citation

Wiedemann, N., T. Uscidda and M. Raubal (2025) Bridging prediction and optimization in on-demand transportation systems with Optimal Transport, paper presented at the *25th Swiss Transport Research Conference (STRC 2025)*, Ascona, May 2025.

Acknowledgments

We would like to thank Thomas Spanninger for the fruitful discussions and valuable feedback. Furthermore, many thanks to Thomas Klein and Jannis Born for reviewing a prior version of this paper. This project is part of the E-Bike City project funded by the Department of Civil and Environmental Engineering (D-BAUG) at ETH Zurich and the Swiss Federal Office of Energy (BFE).

Contents

1 Introduction

The transport sector accounts for 20% of CO_2-emissions worldwide (Statista, 2023) and thus plays a key role in climate action. One possible avenue to reducing emissions is the adoption of on-demand services such as (autonomous) car sharing, which was shown to effectively reduce car ownership (Mishra *et al.*, 2015; Martin and Shaheen, 2011; Liao *et al.*, 2020). There are two main research avenues to improving on-demand transport services: Prediction (Nguyen *et al.*, 2018), e.g., estimating the number of shared cars/bicycles that will be picked up in the next hour, and optimization, e.g., computing the most efficient way to re-distribute bikes/cars. Importantly, good predictions only lead to a reduction of emissions if the system is optimized with respect to the predicted demand.

Meanwhile, machine learning (ML) research in geographic information sciences (GIS) or transportation usually treats prediction as a standalone problem, ignoring its role in downstream tasks (Yan and Wang, 2022). Consider the example of forecasting bike sharing demand per hour and per station. Usually, a time series prediction model such as an LSTM is trained on the data and the prediction quality is evaluated via the mean squared error (MSE) or mean absolute percentage error (MAPE) (Hulot *et al.*, 2018; Brahimi *et al.*, 2022; Shin *et al.*, 2020; Ma and Faye, 2022), since evaluating the resulting CO_2-efficiency or business costs is cumbersome. Crucially, such metrics only quantify the error per time step and station, but ignore the spatial distribution of residuals and their implications in a production setting involving relocation costs. Critically, these costs depend on the distance between erroneous predictions and can be viewed either as a *resource* relocation or as a *user* relocation that is necessary due to prediction errors.

We propose to leverage Optimal Transport (OT) to approximate and minimize relocation costs and thereby the involved emissions. OT provides methods to measure the disparity between two (probability) distributions, which can be leveraged as an evaluation framework comparing the real and predicted spatial distribution in any spatiotemporal prediction task, such as estimating bike sharing demand, traffic congestion or charging station occupancy. Moreover, we demonstrate how the relocation costs can be directly minimized with an OT-based loss function. Our framework, named GeOT, is based on partial OT (Guittet, 2002; Piccoli and Rossi, 2014; Maas *et al.*, 2015) and provides important tools to researchers and industry working with spatiotemporal data to achieve actual advances in resource management and operational efficiency with ML methods.

2 Methods

Optimal transport (OT) is a mathematical framework for comparing probability distributions (Santambrogio, 2015) and has recently become increasingly influential in the field of machine learning (Peyré *et al.*, 2019). Solving an OT problem involves finding the most cost-effective way to transport a source distribution μ to a target distribution ν. In this work, we investigate the use of this distance as an evaluation metric and loss function for geospatial prediction problems. Let $\mathbf{x}_1, \dots \mathbf{x}_n$ be the spatial locations, and let $\mathbf{o} = (\mathbf{o}_1, \dots, \mathbf{o}_n)$ be the true observations and $\hat{\mathbf{o}} = (\hat{\mathbf{o}}_1, \dots, \hat{\mathbf{o}}_n)$ the predicted observations at these locations. For instance, $\mathbf{o}_i$ could represent the demand for shared bicycles at the i-th bike sharing station, located at $\mathbf{x}_i$. For geospatial data, the locations $\mathbf{x}$ are usually two-dimensional, $\mathbf{x} \in \mathbb{R}^2$. The source and target distribution, μ and ν, are set to the discrete distribution of predicted spatial observations, $\mu = \sum_{i=1}^{n} \hat{\mathbf{o}}_i \delta_{\mathbf{x}_i}$, and of the true observations $\nu = \sum_{i=1}^{n} \mathbf{o}_i \delta_{\mathbf{x}_i}$. Additionally, let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a cost function, s.t. $c(\mathbf{x}_i, \mathbf{x}_j)$ measures the cost of moving a unit of mass from location $\mathbf{x}_i$ to location $\mathbf{x}_j$. $\mathbf{C} \coloneqq [c(\mathbf{x}_i, \mathbf{x}_j)]_{1 \le i \le n} \in \mathbb{R}^{n \times n}$ is called the cost matrix. The goal of OT is to transport μ onto ν through a coupling matrix $\mathbf{T} \in \mathrm{U}(\mathbf{p}, \mathbf{q}) \coloneqq \{\mathbf{T} \in \mathbb{R}_+^{n \times n} \mid \mathbf{T}1_n = \mathbf{o}, \mathbf{T}^\top 1_n = \hat{\mathbf{o}}\}$ while minimizing the cost of transportation quantified by c. Here, $\mathbf{T}_{ij}$ denotes the amount of mass transported from $\mathbf{x}_i$ to $\mathbf{x}_j$. In sum, OT aims to solve the following optimization problem $\min_{\mathbf{T} \in \mathrm{U}(\mathbf{p},\mathbf{q})} \sum_{i,j=1}^{n,n} \mathbf{T}_{ij}\mathbf{C}_{ij} \Leftrightarrow \min_{\mathbf{T} \in \mathrm{U}(\mathbf{p},\mathbf{q})} \langle \mathbf{T}, \mathbf{C} \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. A solution $\mathbf{T}^\star$ to this linear programming problem, which always exists, is called an OT coupling. We define the geospatial Optimal Transport (GeOT) error as the c-Wasserstein distance between the true and predicted distribution:

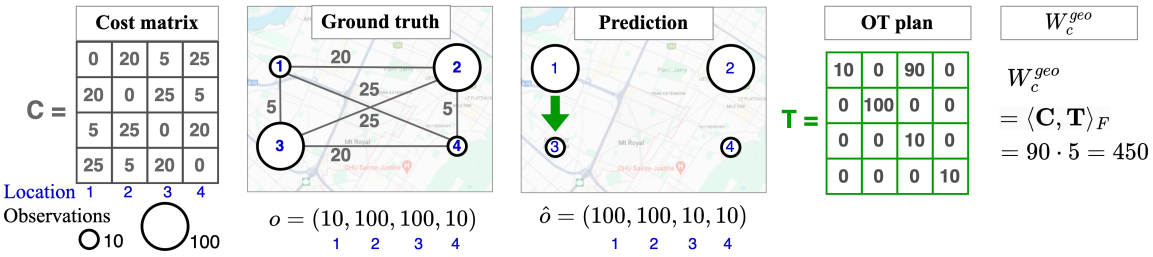
$$\mathrm{W}_c^{geo}(\mu, \nu) = \min_{\mathbf{T} \in \mathrm{U}(\mathbf{p},\mathbf{q})} \langle \mathbf{T}, \mathbf{C} \rangle = \sum_{i,j=1}^{n,n} \mathbf{T}_{ij}^\star \mathbf{C}_{ij}. \tag{1}$$

This value translates to the minimal cost necessary to align the predicted with the true spatial distribution of observations. In other words, $\mathrm{W}_c^{geo}(\mu, \nu)$ measures the total spatial costs to "undo" errors of the predictive model. For a visual explanation, see Figure 1.

2.1 Partial Optimal Transport

The standard OT formulation assumes equal total mass in both distributions, which is unrealistic in our case without normalization. Partial OT was introduced to address mass imbalance by assigning explicit costs to untransported mass using methods like "dummy

Figure 1: Quantifying spatial costs with Optimal Transport. Given a cost matrix C defined between location pairs, prediction errors are measured in terms of the minimal transport costs required to align the predictions with the true observations. In the example, a mass of 90 must be transported from location 1 to location 3 with cost 5, leading to an OT error of 450.



points"(Chapel *et al.*, 2020), "dustbin"(Sarlin *et al.*, 2020), or "waste vectors" (Guittet, 2002). Following Chapel *et al.* (2020), we add a dummy location $\mathbf{x}_{n+1}$ in both source and target measures. The mass at this dummy point is set to zero or the total mass difference ($|\sum_{i=1}^{n} \mathbf{o}_i - \sum_{i=1}^{n} \hat{\mathbf{o}}_i|$) respectively, dependent on whether the source or target distribution has larger mass. Formally, let $s = \min(\sum_{i=1}^{n} \mathbf{o}_i, \sum_{i=1}^{n} \hat{\mathbf{o}}_i)$, and we define $\mathbf{o}_{n+1} = \sum_{i=0}^{n} \hat{\mathbf{o}}_i - s$, and $\hat{\mathbf{o}}_{n+1} = \sum_{i=0}^{n} \mathbf{o}_i - s$. For example, if the sum of observations over all locations is 10 ($\sum_{i=1}^{n} \mathbf{o}_i = 10$), and the predicted total is 12, we add a dummy location with $\mathbf{o}_{n+1} = 2$ and $\hat{\mathbf{o}}_{n+1} = 0$. We denote the adapted measures including the dummy points as $\tilde{\mu}$ and $\tilde{\nu}$, which now have equal mass by design. The cost matrix is adapted to penalize the overshooting mass with a fixed cost ϕ:

$$\widetilde{\mathbf{C}}(\phi) = \begin{pmatrix} c_{11} & \dots & c_{1n} & \phi \\ \dots & \ddots & \dots & \dots \\ c_{n1} & \dots & c_{nn} & \phi \\ \phi & \dots & \phi & \phi \end{pmatrix}$$

As Chapel *et al.* (2020) show, partial OT corresponds to solving balanced OT on $\tilde{\mu}, \tilde{\nu}$ and $\widetilde{\mathbf{C}}$. Thus, we define:

$$\mathrm{W}_{c,\phi}^{geo} = \mathrm{W}_{\tilde{c}}(\tilde{\mu}, \tilde{\nu}) \text{ with } \tilde{\mu} = \sum_{i=1}^{n+1} \hat{\mathbf{o}}_i \delta_{\mathbf{x}_i},\ \tilde{\nu} = \sum_{i=1}^{n+1} \mathbf{o}_i \delta_{\mathbf{x}_i} \text{ and } \widetilde{\mathbf{C}} \text{ as the cost matrix} \tag{2}$$

The solution yields a transportation matrix that contains the flow of mass between locations, as well as the outflow or inflow dependent on the total mass difference. In our evaluation framework, $\mathrm{W}_{c,\phi}^{geo}$ combines the total prediction error and the spatial

distributional error, with ϕ controlling their emphasis: Higher ϕ puts more weight on the total error $\sum_{i=1}^{n} \mathbf{o}_i - \sum_{i=1}^{n} \hat{\mathbf{o}}_i$, while lower ϕ highlights the distributional error.

2.2 OT-based loss function based on Sinkhorn divergences

A natural progression for the OT error is its integration into the *training* of neural networks as a spatial loss function. However, W_c is non-differentiable with respect to its inputs, impeding its direct use as a loss function. One way to alleviate these challenges is to rely on entropic regularization (Cuturi, 2013). Introducing $H(\mathbf{T}) = \sum_{i,j=1}^{n,m} \mathbf{T}_{ij} \log(\mathbf{T}_{ij})$ and $\varepsilon > 0$, the Entropic OT problem between μ and ν is defined as

$$\mathrm{W}_{c,\varepsilon}(\mu, \nu) = \min_{\mathbf{T} \in \mathrm{U}(\mathbf{o}, \hat{\mathbf{o}})} \langle \mathbf{T}, \mathbf{C} \rangle - \varepsilon H(\mathbf{T}). \tag{3}$$

Sinkhorn's algorithm provides an iterative approach for finding a unique solution to the dual formulation of (3). By Danskin's Theorem, the uniqueness of the solution guarantees the differentiability of $\mathrm{W}_{c,\varepsilon}(\mu, \nu)$ with respect to its inputs, allowing its use as a loss function. To correct biases in this loss function it was proposed to center the Entropic OT objective (Genevay *et al.*, 2018; Feydy *et al.*, 2019; Pooladian *et al.*, 2022), defining the *Sinkhorn divergence* as follows:

$$\mathrm{S}_{c,\varepsilon}(\mu, \nu) = \mathrm{W}_{c,\varepsilon}(\mu, \nu) - \tfrac{1}{2}(\mathrm{W}_{c,\varepsilon}(\mu, \mu) + \mathrm{W}_{c,\varepsilon}(\nu, \nu)) \tag{4}$$

In practice, we use the implementation provided in the `geomloss` package (Feydy *et al.*, 2019), which employs the Sinkhorn algorithm.

3 Results

The GeOT framework is applicable to a wide range of application; basically all spatial prediction problems where the spatial distribution of the errors matter. As a real-world example of a spatio-temporal forecasting problem, we utilize bike sharing demand prediction in the following. In this case, W_c^{geo} can be interpreted as bike or user relocations that are necessary due to prediction errors. A public dataset is available from the BIXI bike sharing service in Montreal. The number of bike pickups at 458 stations is aggregated by hour and by station, following Hulot *et al.* (2018). A state-of-the-art time series

prediction model, N-HiTS (Challu *et al.*, 2022), is trained to predict the demand for the next five hours at any time point. The predictions are evaluated on a hundred time points from the test data period. For details on data preprocessing and model training, see Appendix A.

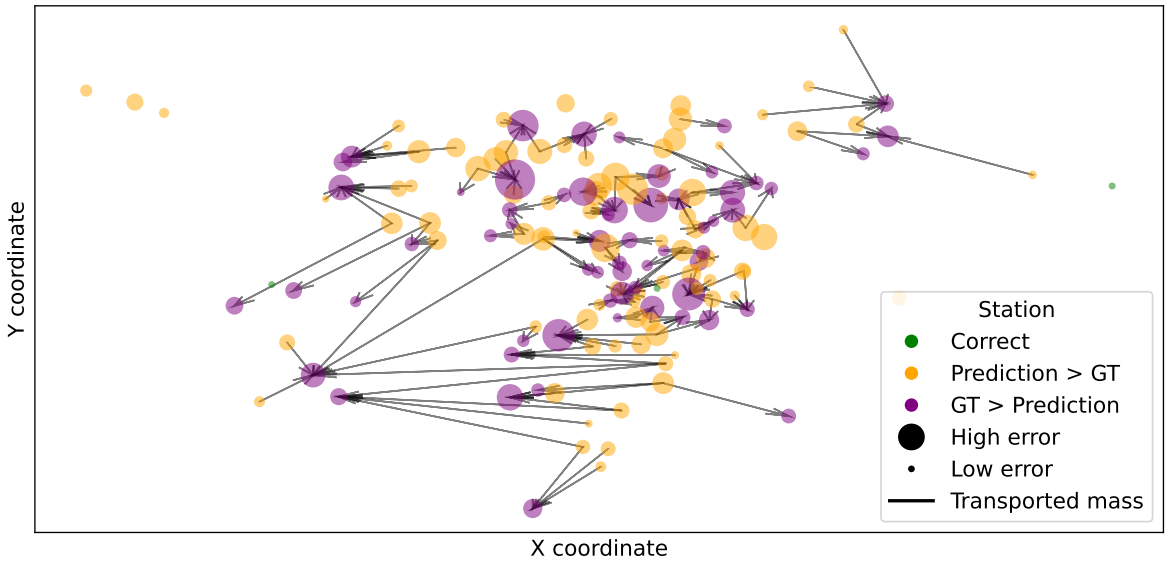
3.1 Evaluating bike sharing demand prediction with OT

First, we demonstrate the computation of the OT error using one example of bike-sharing demand predictions, for a single point in time. For visualization purposes, we subsample one third of the stations. Figure 2 shows the spatial distribution of the residuals at these stations, highlighting, for example, a few stations with significantly underestimated demand (big purple circles) or an overestimation of bike-sharing demand in the bottom-left (orange points). Calculating the OT error involves computing $\mathbf{T}^*$, the optimal transport matrix. We apply partial OT with $\phi = 0$, essentially computing the difference between both distributions without penalizing the total difference of their masses. The arrows in Figure 2 illustrate all nonzero cells of $\mathbf{T}^*$, representing all required redistribution of mass to align the predictions with the true observations. The length of the arrows corresponds to the transport cost, since $\mathbf{C}$ was set to the Euclidean distance between stations. In this example, most errors can be balanced out between neighboring stations, resulting in mass being relocated over short distances from prediction to ground truth. It is worth noting that a few errors are not balanced out since they are ignored through partial OT (see orange point in the bottom-left). The total spatial error corresponds to the sum of all arrow lengths when $\phi = 0$, here $\mathrm{W}_{c,0}^{geo} = \sum_{i,j=1}^{n} \widetilde{\mathbf{C}}_{ij} \mathbf{T}_{ij}^* = 58.91$.

To interpret this error, assume that relocating one bicycle over one kilometer costs \$5. $W_{c,0}^{geo}$ represents the total relocation kilometers required to match the real bike-sharing demand with the predicted supply (apart from their total difference). Thus, the error of this prediction model would cost the bike-sharing service $58.91 \cdot \$5 = \$294.55$ if they needed to fully rebalance their supply to meet future demand. The GeOT framework's output could be integrated into more complex analysis tools specific to the company, such as considering the option of collecting and redistributing multiple bicycles simultaneously.

A major strength of OT is its flexibility to incorporate any arbitrary cost function, without requirements on the function's smoothness or other properties. This enables tailored application-specific analyses, such as using map-matched distances or monetary costs.

Figure 2: Transport map as computed with the GeOT framework. The goodness of the prediction is measured in terms of the relocation costs necessary to align the predictions with the real observations. Here, the difference between real and predicted bike sharing demand is shown, where mass is transported from bike sharing stations with overestimated demand (orange) to stations where the demand was underestimated (purple). In the example, the total spatial costs are rather low since most errors are balanced out with nearby points.



3.2 Comparability across scales

Research on spatio-temporal data oftentimes aggregates data across both space and time, leading to incomparable outcomes due to the Modifiable Areal Unit Problem (MAUP). The choice of aggregation size and method influences results, as observed in various analytical (Gehlke and Biehl, 1934; Buzzelli, 2020) and predictive studies (Smolak *et al.*, 2021; Smith *et al.*, 2014). We argue that OT allows to compare results across scales and between different aggregation methods. Intuitively, aggregating the data in space decreases the error, since the clustered observations are less noisy. On the other hand, the *utility* of the predictions is lower when they are not available on a fine-grained per-location level. In the following, we demonstrate how the GeOT framework can quantify this trade-off for the bike sharing data. In bike sharing research, there is indeed a lack of comparability of previous work due to different aggregation schemes, ranging from single-station prediction (Yang *et al.*, 2016; Qiao *et al.*, 2021) to various clustering schemes (Hulot *et al.*, 2018; Shir *et al.*, 2023; Li and Zheng, 2020). To capture this variety, we also aggregate the bike sharing data with several methods, namely 1) grouping by

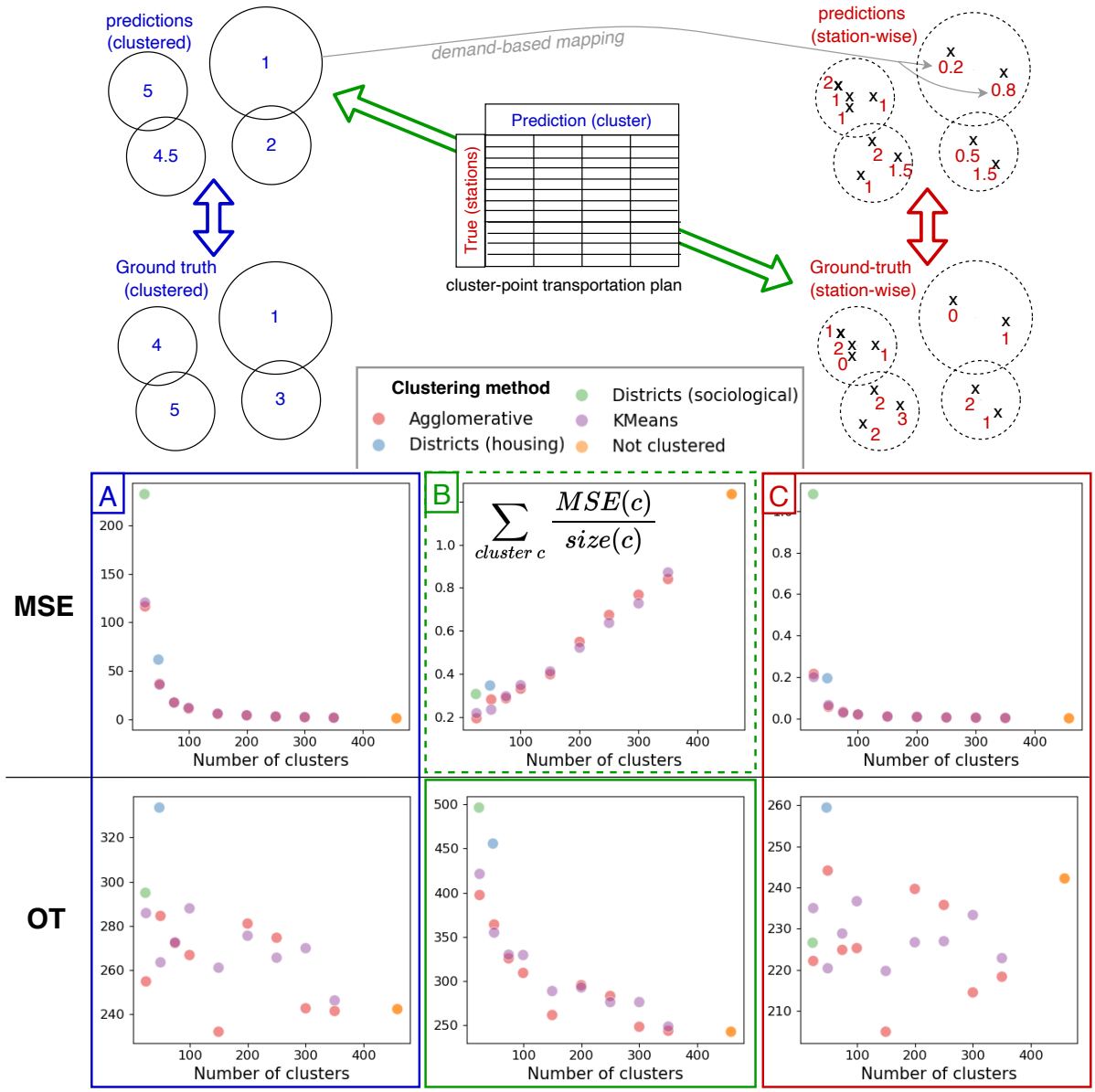
sociological or housing district[1], 2) clustering with the KMeans algorithm (varying k), and 3) clustering with hierarchical (Agglomerative) clustering using different cutoffs. The bike sharing demand of a cluster is the sum of the demand of all its associated stations. One model is trained per configuration, where again the N-HiTS time series prediction model is used. The results are evaluated on the same test time points as before.

As illustrated in Figure 3, we consider three evaluation methods: cluster-level comparison, evaluation of clustered predictions against point observations, and point-level errors. Cluster-level MSE, the standard approach taken in related work, decreases with more clusters (see Figure 3A) because each cluster contains fewer observations, typically resulting in lower errors. The OT error[2] offers a different perspective as it accounts for distances between cluster centers, which increase when fewer clusters are used. Moreover, OT enables comparisons between cluster-predictions and station-level observations, as shown in Figure 3 (green). In bike-sharing, for instance, cluster centers can be viewed as distribution hubs, and the OT error quantifies transport costs for redistributing bikes from hubs to stations. In this case, the OT error decreases with higher granularity (see Figure 3B), because it must redistribute the mass from the cluster centers to individual stations, which are further away from the hub if the cluster is larger. This insight enables balancing operational costs of additional hubs against reduced transport costs – an analysis not possible with the MSE, which can only compare samples of the same size. One way to account for the clusters in the MSE is normalizing the prediction error by the cluster size (see dotted line in Figure 3B). In this case we can observe that larger clusters seem are easier to predict, probably because they exhibit more regular patterns. Finally, point-level errors (Figure 3C) are computed by allocating cluster predictions to stations based on their relative demand in the training set. The trends are similar but more pronounced than in Figure 3A. The MSE declines with the granularity, whereas the OT error balances accuracy with spatial granularity and achieves a minimum at 150 clusters found with Agglomerative clustering. In summary, the GeOT framework provides a refined evaluation across scales, capturing both absolute errors and their operational implications. When models are trained at multiple scales, the GeOT framework helps to select the optimal scale and aggregation method for each use case.

[1] sociological districts from `https://www.donneesquebec.ca/recherche/dataset/vmtl-quartiers-sociologiques` and housing districts according to `https://www.donneesquebec.ca/recherche/dataset/vmtl-quartiers`

[2] Here, we use $\mathrm{W}^{geo}_{c,\phi}$ with ϕ equivalent to the 10%-quantile of $\mathbf{C}$, to model realistic business costs that arise mainly from redistribution but partly from a general over- or underestimation of bike sharing demand.

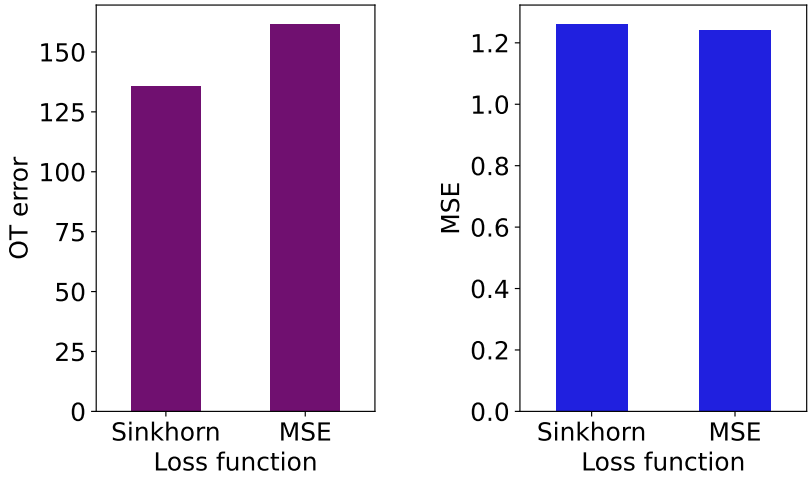
Figure 3: Comparing the prediction quality with MSE and OT error across spatial aggregation scales and clustering techniques. The MSE simply indicates higher errors for bigger clusters (A), or lower station-wise error with more data aggregation (B). Optimal Transport allows for asymmetric cost matrices (green) to compute the costs for transporting from prediction-*clusters* to the ground-truth-*points* (B). In addition, OT takes into account the distances between clusters, providing a refined perspective on the optimal aggregation scale (A and C).



3.3 Training models with an OT-based loss function

To demonstrate the effectiveness of the OT-based loss function (see subsection 2.2), we train the N-HiTS model with the Sinkhorn divergence as the loss function and compare to a standard MSE loss. The trained models are evaluated on test data in terms of the

Figure 4: Training with the Sinkhorn loss (an OT-based loss function) can effectively reduce the OT error between predictions and ground truth. This comes at minor increase of the MSE.



(a) OT error (b) MSE

MSE and the balanced OT error ($\phi = 0$). The cost matrix $\mathbf{C}$ was set to the Euclidean distance between stations in km. Figure 4 demonstrates that the OT error W^{geo} can be reduced when training with the Sinkhorn loss. This comes at a minor increase of the MSE, compared to training with a standard MSE loss. Thus, this experiment shows promising evidence that training with the Sinkhorn loss can improve the spatial distribution of the predictions.

4 Conclusion

This paper proposes to evaluate spatio-temporal predictions with Optimal Transport, highlighting its capacity to reflect reductions in operational costs within predictive methods. Our experiments demonstrate the value of OT for evaluating and training prediction models. The proposed framework is generally applicable to any prediction problem where the spatial distribution of the errors matters. A notable limitation is the computational demand of computing the OT distance and the Sinkhorn loss, particularly in cases involving numerous locations. The potential of OT in GIS and transportation extends further, such as its extension to the temporal dimension considering relocation across space *and* time.

5 References

Brahimi, N., H. Zhang, L. Dai and J. Zhang (2022) Modelling on car-sharing serial prediction based on machine learning and deep learning, *Complexity*, **2022** (1) 8843000.

Buzzelli, M. (2020) Modifiable Areal Unit Problem, *International encyclopedia of human geography*, 169.

Challu, C., K. G. Olivares, B. N. Oreshkin, F. Garza, M. Mergenthaler-Canseco and A. Dubrawski (2022) N-HiTS: Neural hierarchical interpolation for time series forecasting, *arXiv preprint arXiv:2201.12886*.

Chapel, L., M. Z. Alaya and G. Gasso (2020) Partial optimal tranport with applications on positive-unlabeled learning, *Advances in Neural Information Processing Systems*, **33**, 2903–2913.

Cuturi, M. (2013) Sinkhorn distances: Lightspeed computation of Optimal Transport, paper presented at the *Advances in Neural Information Processing Systems*, vol. 26.

Danskin, J. M. (1967) *The Theory of Max-Min and its Applications to Weapons Allocation Problems*, vol. 5, Springer.

Feydy, J., T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé and G. Peyré (2019) Interpolating between Optimal Transport and MMD using Sinkhorn divergences, paper presented at the *The 22nd International Conference on Artificial Intelligence and Statistics*, 2681–2690.

Gehlke, C. E. and K. Biehl (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association*, **29** (185A) 169–170.

Genevay, A., G. Peyré and M. Cuturi (2018) Learning generative models with sinkhorn divergences, paper presented at the *International Conference on Artificial Intelligence and Statistics*, 1608–1617.

Guittet, K. (2002) Extended Kantorovich norms: a tool for optimization, *Dissertation, INRIA*.

Herzen, J., F. Lässig, S. G. Piazzetta, T. Neuer, L. Tafti, G. Raille, T. Van Pottelbergh,

M. Pasieka, A. Skrodzki, N. Huguenin *et al.* (2022) Darts: User-friendly modern machine learning for time series, *The Journal of Machine Learning Research*, **23** (1) 5442–5447.

Hulot, P., D. Aloise and S. D. Jena (2018) Towards station-level demand prediction for effective rebalancing in bike-sharing systems, paper presented at the *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 378–386.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu (2017) Lightgbm: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems*, **30**.

Li, Y. and Y. Zheng (2020) Citywide bike usage prediction in a bike-sharing system, *IEEE Transactions on Knowledge and Data Engineering*, **32** (6) 1079–1091.

Liao, F., E. Molin, H. Timmermans and B. van Wee (2020) Carsharing: the impact of system characteristics on its potential to replace private car trips and reduce car ownership, *Transportation*, **47** (2) 935–970.

Ma, T.-Y. and S. Faye (2022) Multistep electric vehicle charging station occupancy prediction using hybrid LSTM neural networks, *Energy*, **244**, 123217.

Maas, J., M. Rumpf, C. Schönlieb and S. Simon (2015) A generalized model for Optimal Transport of images including dissipation and density modulation, *ESAIM: Mathematical Modelling and Numerical Analysis*, **49** (6) 1745–1769.

Martin, E. and S. Shaheen (2011) The impact of carsharing on household vehicle ownership, *Access Magazine*, **1** (38) 22–27.

Mishra, G. S., R. R. Clewlow, P. L. Mokhtarian and K. F. Widaman (2015) The effect of carsharing on vehicle holdings and travel behavior: A propensity score and causal mediation analysis of the San Francisco Bay area, *Research in Transportation Economics*, **52**, 46–55.

Nguyen, H., L.-M. Kieu, T. Wen and C. Cai (2018) Deep learning methods in transportation domain: a review, *IET Intelligent Transport Systems*, **12** (9) 998–1004.

Peyré, G., M. Cuturi *et al.* (2019) Computational Optimal Transport: With applications to data science, *Foundations and Trends® in Machine Learning*, **11** (5-6) 355–607.

Piccoli, B. and F. Rossi (2014) Generalized Wasserstein distance and its application to transport equations with source, *Archive for Rational Mechanics and Analysis*, **211**, 335–358.

Pooladian, A.-A., M. Cuturi and J. Niles-Weed (2022) Debiaser beware: Pitfalls of centering regularized transport maps.

Qiao, S., N. Han, J. Huang, K. Yue, R. Mao, H. Shu, Q. He and X. Wu (2021) A dynamic convolutional neural network based shared-bike demand forecasting model, *ACM Transactions on Intelligent Systems and Technology*, **12** (6) 1–24.

Santambrogio, F. (2015) Optimal transport for applied mathematicians, *Birkhäuser, NY*, **55** (58-63) 94.

Sarlin, P.-E., D. DeTone, T. Malisiewicz and A. Rabinovich (2020) Superglue: Learning feature matching with graph neural networks, paper presented at the *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Shin, D.-H., K. Chung and R. C. Park (2020) Prediction of traffic congestion based on LSTM through correction of missing temporal and spatial data, *IEEE Access*, **8**, 150784–150796.

Shir, B., J. Prakash Verma and P. Bhattacharya (2023) Mobility prediction for uneven distribution of bikes in bike sharing systems, *Concurrency and Computation: Practice and Experience*, **35** (2) e7465.

Sinkhorn, R. (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices, *Annals of Mathematical Statistics*, **35**, 876–879.

Smith, G., R. Wieser, J. Goulding and D. Barrack (2014) A refined limit on the predictability of human mobility, paper presented at the *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 88–94.

Smolak, K., K. Siła-Nowicka, J.-C. Delvenne, M. Wierzbiński and W. Rohm (2021) The impact of human mobility data scales and processing on movement predictability, *Scientific Reports*, **11** (1) 15177.

Statista (2023) Transportation emissions worldwide.

Yan, R. and S. Wang (2022) Integrating prediction with optimization: Models and applications in transportation management, *Multimodal Transportation*, **1** (3) 100018.

Yang, Z., J. Hu, Y. Shu, P. Cheng, J. Chen and T. Moscibroda (2016) Mobility modeling and prediction in bike-sharing systems, paper presented at the *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* , 165–178, Singapore Singapore, ISBN 978-1-4503-4269-8.

A Data and preprocessing

The bike sharing dataset was downloaded from Kaggle[3] and restricted to the period from 15th of April to 15th of November 2014, since the service is closed in winter, leading to large gaps in the time series across years. Only stations with missing coordinates or maintenance stations were removed.

We train an established time series prediction model, N-HiTS (Challu *et al.*, 2022), implemented in the `darts` library (Herzen *et al.*, 2022). The model was chosen since it outperformed other common approaches such as Exponential Smoothing, LightGBM (Ke *et al.*, 2017) or XGBoost in our initial experiments.

The model is trained for 100 epochs with early stopping. The learning rate was set to $1e^{-5}$. The time series was treated as multivariate data with one variable per bike sharing station or charging station. A lag of 24 is used to learn daily patterns, and the hour and weekday are provided as past covariates. The number of stacks in the N-HiTS model was set to 3. The number of output time steps corresponds to our forecast horizon of five time steps. For evaluation, we draw 100 samples from the test data (last 10% of the time series) and predict the next five time steps based on the respectively preceding time series, without re-training the model. For further implementation details, we refer to our source code.

[3] `https://www.kaggle.com/datasets/aubertsigouin/biximtl`