Active Learning for transport studies: the case of rare or sparse demand samples

Zhan Shen[1], Guido Cantelmo*[2], S. F. A. Batista[3], Mónica Menéndez[4], and Constantinos Antoniou[5]

[1]M.Sc., Technical University of Munich, Germany

[2]Assistant Professor, Technical University of Denmark, Denmark; * Corresponding author: guica@dtu.dk

[3]Dr., New York University Abu Dhabi, United Arab Emirates

[4]Prof., New York University Abu Dhabi, United Arab Emirates

[5]Prof., Technical University of Munich, Germany

SHORT SUMMARY

For two decades now, the rapid rise and wide diffusion of new technologies have enabled the generation of a massive amount data – commonly known as Big-Data. Big-Data have been widely adopted in nearly every field of transportation, from behavioural analysis, to traffic predictions, or model calibration. While the initial promise of Big-Data was to allow for a better understanding of the total population and their diversity, years of research prove that this is not always the case. The main limitation is that Big-Data are difficult to interpret. For instance, if we use mobile phone network data to create a demand matrix, the model will have a bias (penetration rate, service provider, multiple devices) which is difficult to measure.

In general, the data should be representative of the entire population and represent all different demand segments in a correct way. This is currently not the case for many so-called big-data sources and, anyway, it is not easy to assess the representativeness of a given data-set. This paper proposes using active learning to address this problem. Active learning models use machine learning algorithms to sample data from a population (or dataset) and create a small yet representative set of observations that encompass the main attributes of the entire population. To that end, we introduce an enhanced active learning algorithm that combines two models. Traditional Active Learning Techniques (such as Gaussian Processes) are used to sample supply-related data. A heuristic model based on the Branch and Bound algorithm is instead used to sample demand-related information. By combining the two models, the proposed approach can sample supply and demand data together. The model is used to sample origin-destination trips for bike-sharing from sparse demand matrices. The case study uses real world data from New York city, showing promising results.

Keywords: Active Learning, Big-Data, Data Collection, Mobility Demand, Optimization.

1. INTRODUCTION

Due to advances in computing and telecommunication technologies, immense volumes of data are constantly produced and collected – the so-called Big-Data (Mahajan et al., 2021). The spread of these new data, often coming from non-conventional sources (Chaniotakis, Antoniou, & Pereira, 2016), combined with the availability of affordable sensors (Cipriani, Gemma, Mannini, Carrese, & Crisalli, 2021) is changing the way we conduct mobility analyses and forecast travel demand

(Chen, Ma, Susilo, Liu, & Wang, 2016). In transportation research, Big-Data has been widely adopted in many fields, including behavioral analysis (Pereira, Rodrigues, Polisciuc, & Ben-Akiva, 2015), traffic modelling (Bramich, Menéndez, & Ambühl, 2022), analysis of network performance (Loder, Ambühl, Menendez, & Axhausen, 2019), model calibration (Cantelmo & Viti, 2019), fleet re-balancing (Lahoorpoor, Faroqi, Sadeghi-Niaraki, & Choi, 2019), and demand forecasting (Ma, Antoniou, & Toledo, 2020).

One promise of Big-Data was to allow a better understanding of the total population and its underlying mobility habits. However, there is still a large debate about how useful or representative this data is, especially when it comes to the mobility demand. That being said, assessing the representativeness of a given dataset is not a trivial task (Chen et al., 2016). For example, social media has been abundantly used to study travel behavior (Pereira et al., 2015; Chaniotakis et al., 2016). We now know that users aged between 45 and 55 use more Facebook than Twitter or Instagram (Singh, Halgamuge, & Moses, 2019). Therefore, different social media data sets will show different behavioral trends. Moreover, these statistics might change from country to country. While age distribution is – hopefully – easy to observe, other aspects such as political views or level of education might not be. The same problem applies to other data sources. For instance, if mobile phone network data are used to estimate the mobility demand, the model will have similar problems (depending on the penetration rate, service provider, type of subscription, among other factors) (Huang et al., 2018). In all these cases, Big-Data is only useful if the sample represents the entire population and all the different demand segments properly. This, however, is currently not the case for many of the so-called Big-Data sources (Chen et al., 2016).

In this context, this paper focuses on developing a methodology to sample a representative datasets for mobility analysis using Active Learning. The term Active learning refers to the subset of machine learning models that, starting from a non representative data-set, is able to interactively collect new data–points and build a small yet representative set of observations that encompasses the main attributes of the problem. More specifically, we propose an enhanced active learning algorithm that combines two models. On the one hand, traditional Active Learning Techniques (such as Gaussian Processes) are used to sample supply-related data. On the other hand, a heuristic model based on the Branch and Bound algorithm is used to sample demand-related information. By combining the two models, the proposed approach can sample supply and demand data together. In this paper, we use this model to sample origin-destination trips for bike-sharing in New York City.

The remainder of this paper is organized as follows. Section 2 describes the Active Learning model and the extension of the Branch-and-Bound algorithm. Section 3 presents the case study and discusses some preliminary results. Section 4 outlines the conclusions of this paper.

2. METHODOLOGY

Batista, Cantelmo, Menéndez, and Antoniou (2022) presented an Active Learning framework based on Gaussian Processes to sample travel patterns for mobility analysis, considering only supply-related data (i.e. the spatial distribution of intersections which are framed as the possible origin and destination points of travelers). The model was used to sample travel distances for the calibration of aggregated traffic models based on the Macroscopic Fundamental Diagram (MFD). Under the assumption that travel distances are uniformly distributed, results show that a small percentage (between 2% and 10%) of all possible travels in the network is sufficient for estimating representative aggregated travel distances for MFD models. The model developed by Batista et al. (2022) can sample any supply-related feature. However, it is not suited for demand-related features. Active learning models start from an empty (and not representative) dataset. The model then iteratively populates this dataset with new data points to create a richer and more representa-

tive dataset. Once the dataset becomes representative, adding new data points becomes redundant. Therefore, we need to know the domain of a dataset to infer its representativeness. In the case of the supply information, this domain is known (i.e. the transport network). However, the domain (i.e., trip characteristics) is unknown for the demand.

It is often challenging to retrieve information, for instance, about how many users travel with a certain transportation mode or within a specific origin-destination pair. Even when this information is available, the resulting sample is often large. For example, large metropolitan areas count millions of users, each of them with different characteristics, preferences, and mobility habits. This results in a large dataset that would require prohibitive computational times for existing Active Learning models to analyze. To solve this issue, we propose a new and enhanced framework designed for transport analysis and able to capture both supply- and demand-related features. First, we define an observation (or data point) as follows:

$$x_{ij} = \{x_{ij}^{Demand}, x_{ij}^{Supply}\} \tag{1}$$

where i and j are two nodes in the transport network, x_{ij}^{Demand} represents the characteristics (or features) of the demand that travels between them, while x_{ij}^{Supply} includes the characteristics of the supply. For instance, demand features could be the number of trips between i and j, or the socio-demographic characteristics associated with those trips. Supply features might instead include the travel distance, free-flow travel time, or type of infrastructure. Using this representation, the domain of x_{ij} is known (i.e. the transport network). Therefore, the algorithm proposed in Batista et al. (2022) could now be used to sample both x_{ij}^{Demand} and x_{ij}^{Supply}.

Unfortunately, x_{ij}^{Demand} is sparse, meaning that the majority of the supply-related nodes have (close to) zero demand (e.g., nodes representing infrastructural change on a motorway). Note that, supply-related nodes could also be defined as just simply origin and destination nodes. To deal with sparsity, we develop a framework that leverages the Branch and Bound (B&B) algorithm to sample the sparse vector x_{ij}^{Demand}. The model starts from a non-representative dataset, which includes some observations about the demand in the network. At each iteration, x_{ij}^{Demand} is assumed to be equal to its mean $\bar{x}^{Demand}$ for all observations in the dataset. If this difference is above a given threshold δ, the B&B partitions the network into sub-regions and samples new information for each of them. The size of the partition depends on the current number of observations and influences directly the number of iterations o the model. For each partition, the average demand $\bar{x}^{Demand}$ is computed. Then, the difference between available observations and $\bar{x}^{Demand}$ is evaluated, and the procedure is repeated until the error is below the established tolerance threshold δ for all sub-regions. As the domain is sparse, the model outputs large regions where the demand is zero while will intensify its search in those sub-regions where the demand is heterogeneous (as the variance of the residuals is larger).

Figure 1 shows an illustrative example, where the model samples origin-destination pairs of nodes from a synthetic demand matrix. The full (unknown) matrix is the starting point. At each iteration, the matrix is divided into two and the highest and lowest values are sampled. If the difference between these two is higher than a given threshold ($\delta = 3$, in this example), the model stops. Otherwise, the model keeps partitioning the matrix. Note that the values of the estimated matrix do not necessary match those of the original matrix. For instance, in Partition A, all unlabelled data are assumed to be equal to the mean of the available data-points, therefore $\bar{x}^{Demand} = \left(\frac{3-0}{2}\right) = 1.5$.

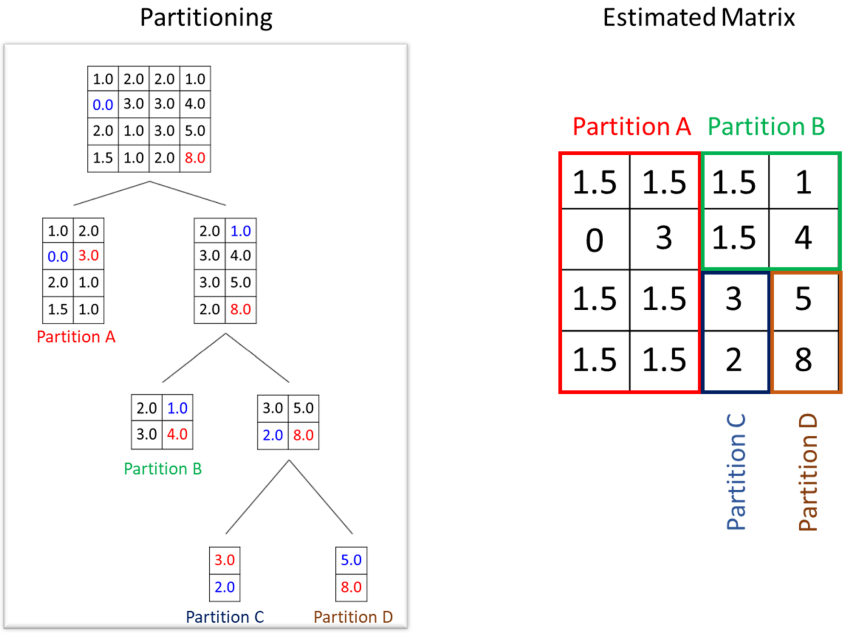

Figure 1: Schematic example that shows how the developed B&B heuristic works.

3. RESULTS AND DISCUSSION

The proposed heuristic model is used here to sample bike-sharing demand data. We utilize the open data provided by CitiBike, for the bike-sharing system of New York city. The data is publically available online at https://s3.amazonaws.com/tripdata/index.html. The dataset includes 746 stations and almost 2,000,000 trips. Figure 2 shows the spatial distribution of the demand over the bike stations in downtown New York city for December 2018.

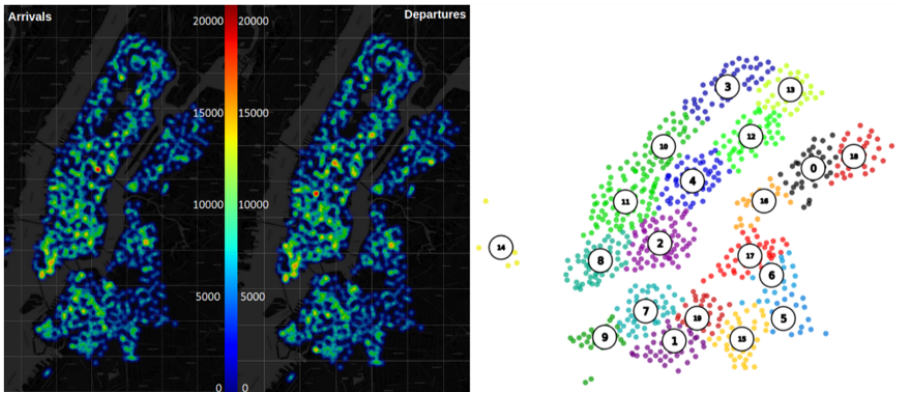

Figure 2: Spatial distribution of departures and arrivals of trips in the bike stations (left); Clusters of bike-sharing stations (right).

There were a total of 41,172 rides per day for this month. Bike-Sharing stations were partitioned into 20 zones using Gaussian Mixture models (GMM). Figure 2 depicts the resulting partitioning. We estimated 400 station-to-station demand matrices, i.e. one for each pair of origin-destination zones. Some of these matrices are sparse while others are not. The developed heuristics was then used to create an approximated station-to-station demand matrix for each pair of zones. Results were evaluated in terms of Precision (P), Accuracy (A), and Efficiency (E) as follows:

$$P = 1 - \frac{\log(\delta + 1)}{\log(x^{Max})} \tag{2}$$

$$A = 1 - \frac{\log(RMSE(\delta))}{\log(RMSE(\delta^{Max}))} \tag{3}$$

$$E = 1 - \frac{N_S}{N_T} \tag{4}$$

Where δ represents the threshold of the model – i.e. the maximum acceptable error for the B&B; δ^{Max} is the maximum value for the threshold $RMSE(\delta)$ is the Root Mean Squared Error between the real and the estimated OD matrix for a given threshold δ; N_s is the sample size; and, N_T is the total number of observations in the dataset. The term x^{Max} is the largest observation in the dataset, such as $x^{Max} \geq x_{ij}^{Demand}, \forall(i,j)$.

The precision depends only on the parameter δ, and it can be seen as an input of the B&B. For low values of the threshold δ, we expect a low variance of the residuals (i.e. $P = 1$ should represent the perfect match). The Accuracy is a similar indicator but is based on the actual error between the estimated and the observed demand values. This means that it can only be used for validation, as the real value x_{ij}^{Demand} is unknown. The efficiency indicates how many data points are necessary to achieve a certain precision. For instance, $E = 0.8$ means that only 20% of the data is required to achieve a specific level of precision. Figure 3 shows the results for two demand matrices, a sparse matrix, and a less sparse matrix.

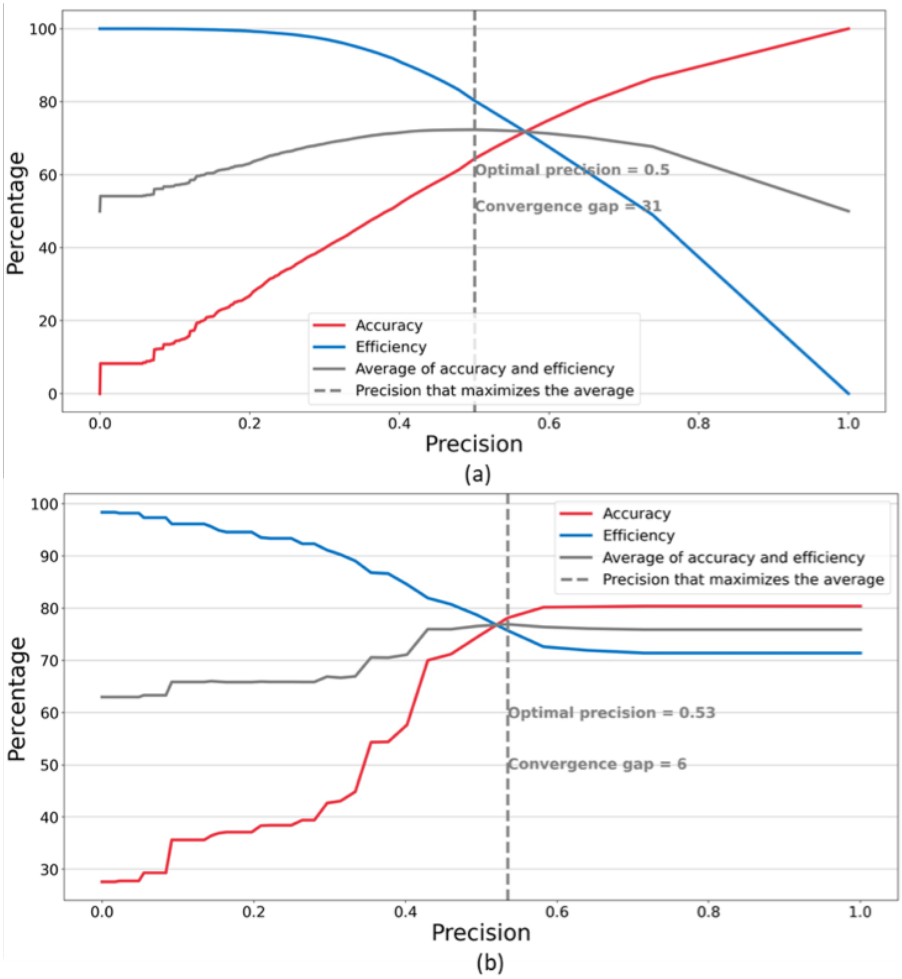

Figure 3: Accuracy, Efficency, and Precision for: (a) a non-sparse matrix; and (b) a sparse matrix.

Results suggest two observations. First, while the model works for sparse and non-sparse matrices, in the first case (Figure 3 (b)), both accuracy and efficiency are extremely high, meaning that good matrices can be obtained with few data points. For example, 30% of the database is sufficient to achieve 80% of accuracy. As the sparsity decreases (Figure 3 (a)), the model still performs well, but the trade-off between accuracy and efficiency becomes almost linear, meaning that a reduced precision translates unavoidably into a less representative dataset.

4. CONCLUSIONS

We argue in this study that there is a lack of methodologies that focus on collecting and building representative data sets for transport analysis. Therefore, this paper introduces a new and enhanced Active Learning framework designed to answer this question. The model combines traditional Active Learning Techniques (Gaussian Processes) and optimization techniques (Branch and Bound) to sample supply- and demand-related data from a transport network. The model is general and allows collecting different types of information such as observed travel times (using Google API or GPS data), venue popularity (Google Popular Times), travel distances (network feature), and OD matrices (using trip records). The model focuses on capturing heterogeneity at a spatial level. However, by combining Gaussian Processes and B&B, more complex heterogeneity can also be captured. For instance, Gaussian Processes can be used to sample information in a given geographical area, while the B&B model might indicate which areas require larger samples. Following this line, future work will focus on combining the two models and sampling more complex features (e.g., socio-demographic), as well as on assessing existing datasets' representativeness.

ACKNOWLEDGMENT

Sérgio Batista and Mónica Menéndez acknowledge support by the NYUAD Center for Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Award CG001.

REFERENCES

Batista, S. F. A., Cantelmo, G., Menéndez, M., & Antoniou, C. (2022). A gaussian sampling heuristic estimation model for developing synthetic trip sets. *Computer-Aided Civil and Infrastructure Engineering*, *37*(1), 93-109. doi: https://doi.org/10.1111/mice.12697

Bramich, D., Menéndez, M., & Ambühl, L. (2022). Fitting empirical fundamental diagrams of road traffic: A comprehensive review and comparison of models using an extensive data set. *IEEE Transactions on Intelligent Transportation Systems*.

Cantelmo, G., & Viti, F. (2019). A big data demand estimation framework for multimodal modelling of urban congested networks. In E. G. Nathanail & I. D. Karakikes (Eds.), *Data analytics: Paving the way to sustainable urban mobility* (pp. 139–146). Cham: Springer International Publishing. doi: 10.1007/978-3-030-02305-8_17

Chaniotakis, E., Antoniou, C., & Pereira, F. (2016). Mapping social media for transportation studies. *IEEE Intelligent Systems*, *31*(6), 64-70. doi: 10.1109/MIS.2016.98

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285-299. doi: 10.1016/j.trc.2016.04.005

Cipriani, E., Gemma, A., Mannini, L., Carrese, S., & Crisalli, U. (2021). Traffic demand

estimation using path information from bluetooth data. *Transportation Research Part C: Emerging Technologies*, *133*, 103443. doi: 10.1016/j.trc.2021.103443

Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., & Wang, F.-Y. (2018). Modeling real-time human mobility based on mobile phone and transportation data fusion. *Transportation Research Part C: Emerging Technologies*, *96*, 251-269. doi: 10.1016/j.trc.2018.09.016

Lahoorpoor, B., Faroqi, H., Sadeghi-Niaraki, A., & Choi, S.-M. (2019). Spatial cluster-based model for static rebalancing bike sharing problem. *Sustainability*, *11*(11). doi: 10.3390/su11113205

Loder, A., Ambühl, L., Menendez, M., & Axhausen, K. W. (2019). Understanding traffic capacity of urban networks. *Scientific reports*, *9*(1), 1–10.

Ma, T., Antoniou, C., & Toledo, T. (2020). Hybrid machine learning algorithm and statistical time series model for network-wide traffic forecast. *Transportation Research Part C: Emerging Technologies*, *111*, 352-372. doi: https://doi.org/10.1016/j.trc.2019.12.022

Mahajan, V., Kuehnel, N., Intzevidou, A., Cantelmo, G., Moeckel, R., & Antoniou, C. (2021). Data to the people: a review of public and proprietary data for transport models. *Transport Reviews*, *0*(0), 1-26. doi: 10.1080/01441647.2021.1977414

Pereira, F. C., Rodrigues, F., Polisciuc, E., & Ben-Akiva, M. (2015). <italic>why so many people</italic>? explaining nonhabitual transport overcrowding with internet data. *IEEE Transactions on Intelligent Transportation Systems*, *16*(3), 1370-1379. doi: 10.1109/TITS.2014.2368119

Singh, A., Halgamuge, M. N., & Moses, B. (2019). 5 - an analysis of demographic and behavior trends using social media: Facebook, twitter, and instagram. In N. Dey, S. Borah, R. Babo, & A. S. Ashour (Eds.), *Social network analytics* (p. 87-108). Academic Press. doi: 10.1016/B978-0-12-815458-8.00005-0