Dynamic OD Matrix Estimation using Data-Driven Modelling under Data-Scarcity: an application of Sparse Variational Gaussian Process

Giovanni Tataranno*

Faculty of Science, Technology and Medicine (FSTM)
University of Luxembourg, Esch-Sur-Alzette, Luxembourg, L-4364
Email: giovanni.tataranno@uni.lu

Federico Bigi

Faculty of Science, Technology and Medicine (FSTM)
University of Luxembourg, Esch-Sur-Alzette, Luxembourg, L-4364
Email: federico.bigi@uni.lu

Francesco Viti

Faculty of Science, Technology and Medicine (FSTM)
University of Luxembourg, Esch-Sur-Alzette, Luxembourg, L-4364
Email: francesco.viti@uni.lu

* Corresponding author

Short abstract for presentation at Transport Research Days 2025 - BIVEC/GIBET 2025

Introduction

Origin-destination (OD) matrix estimation is essential in transportation planning, as it provides a detailed understanding of travel patterns, supporting infrastructure development, traffic management, and urban planning decisions. As for Viti (2012), OD matrices have been traditionally estimated in different ways: direct sampling estimation (e.g. travel surveys and traffic counts) and model estimation; the former involves applying a system of models that computes the approximate number of journeys made with a certain mode, for a specific purpose during a certain period of time. OD estimation methods can typically generate either time-dependent (dynamic) or time-independent (static) matrices Peterson (2007), and can be estimated using either trip-based or activity-based models, Dong et al. (2006). While many methodologies have been explored in OD estimation, probabilistic data-driven models started emerging as a strong alternative, especially in data-scarce scenarios, as they aim to infer the underlying distributions, rather than focusing solely on raw data, Nakatsuji (2011). This is a common problem in OD estimation when using survey data, as these samples are often limited in size or in quality. Krishnakumari et al. (2020) presents their research in which they address OD matrix estimation under data scarcity by leveraging 3D supply patterns. While this approach simplifies the estimation process, it relies on various assumptions about route choice, which may limit flexibility and accuracy. To position ourselves within the current

research, the presented paper describes a methodology for dynamic OD estimation using trip-based modeling with activity components using data-driven modeling, specifically Sparse Variational Gaussian Processes (*SVGPs*), described in Section . Unlike previous methods, *SVGP* makes no assumptions about the dataset and does not involve fitting; instead, its predictions are based solely on the statistical relevance and noise/uncertainty of the data Carvalho (2014). This affirmation is confirmed by the Case Study, presented in Section , in which the *SVGP* is tested under data-scarcity conditions (up to 10% of the initial dataset as the training set), providing robust and precise predictions. Our model integrates elements from both trip and activity-based approaches by sequentially predicting trip attributes—such as start time and destination, alongside the trip's purpose, allowing us to predict both static and dynamic OD matrices. To the authors' knowledge, no previous trip-based dynamic OD matrix estimation research has included activity information within the model.

Methodology

A Gaussian Process (*GP*) (Gelman et al. (2021)) is a statistical model used to estimate probability distributions for both linear and non-linear data, allowing predictions to be made in unseen regions of a problem, complete with associated uncertainty. However, due to the high computational complexity of these models($O(n^3)$), in real applications, *GP*s are often replaced by Sparse Variational Gaussian Processes (*SVGP*), which significantly reduces computational complexity by approximating the full *GP*, Ghosh et al. (2006). The use of the *SVGP* for OD matrix estimation offers several advantages. Being a data-driven model, it allows different types of data to be used as input (e.g. traffic counts), ensuring flexible modeling. In addition, the *SVGP* allows for control over the computational complexity by reducing the number of *induction points*, a subset of points used to approximate the full *GP*, thus simplifying the computations without significantly compromising accuracy. Finally, the accuracy of these models does not depend on the amount of data available, but on its *statistical relevance* Titsias (2009). In this study, we use the *SVGP* to predict dynamic OD matrices, specifically our goal is to use *SVGP*s to predict, for each trip, the **start time**, **destination**, **activity type**, and the **arrival time**. The ability to predict individuals' activity types, along with their associated times and locations, provides key insights into how people use transportation systems. The strength of this methodology is that these insights are derived **without** relying on assumptions (e.g. amount of shortest paths used for each OD pair, distribution of flows on the network), offering a more unbiased approach. To maximize prediction accuracy while minimizing the number of *SVGP* models used, we opted for a chain architecture where each *SVGP* model's output serves as the input for the subsequent model, as shown on Figure 2. During the training phase, all inputs to each *SVGP* come solely from the training dataset, while in the prediction phase, the inputs are provided by the outputs of the previous *SVGP* models.

Figure 3 shows the full framework of our model to predict a full trip for each individual. Figure 1 shows the structure of the prediction process for a single trip, where we begin by training the first *SVGP* using socio-demographic data to predict the departure time. This prediction, combined with the socio-demographic inputs, is then fed to a second *SVGP* to predict the activity to be performed. The same iterative approach is applied to predicting

the destination and arrival time. Figure 2 shows then how we extend this methodology to the full activity chain: after the first trip has been predicted, subsequent trips are predicted sequentially, using information from previous trips and input about the total number of trips to be made. To reduce computational complexity and enhance the prediction, we applied Principal Component Analysis (*PCA*, Djukic et al. (2012)) to the input data, as it reduces the input space while filtering irrelevant components. As the *SVGP* is still computationally intensive, we optimized the process by allowing our method to make predictions simultaneously over a large population, making it scalable and efficient, rather than predicting individually for each person. The problem faced working with *SVGPs* in multi-output predictions is that these models produce only one continuous output at a time. To predict four variables per trip — departure and arrival times, destination zone, and activity type — we separated the tasks into discrete-choice predictions (zone and activity) and regression predictions (departure and arrival times). For the discrete-choice predictions (zone and activity), as only one output is provided by each model, we needed to train one *SVGP* for each choice option. The *SVGP* were then trained to output a continuous value between 0 and 1, which represents in our case the likelihood that a particular input corresponds to that specific choice. To achieve this, we applied a softmax-like transformation to the outputs of all *SVGP* to generate a Probability Density Function (*PDF*) to obtain the relative probabilities, constructing then a Cumulative Density Function (*CDF*) to sample and select the specific choice. For the remaining variables to predict (arrival and departure time), we trained one *SVGP* for each regression. This whole process is then repeated for each possible trip that is available in the dataset, as shown in Figure 2. By then chaining these predictions, we were able to obtain the final OD matrices.

Results

To demonstrate the effectiveness of the proposed methodology, we conducted experiments to reconstruct full Origin-Destination (OD) matrices from different levels of limited samples. As input, we used a synthetic travel survey performed including more than 80,000 trips. To assess the capability of the model under data scarcity, we created training sets sampling from 80%, 50%, 30%, and 10% of the full dataset. The synthetic dataset has been set to have about 95% of travellers performing 6 trips or fewer daily, with a maximum of 10 trips. For this specific study and for computational purposes, we opted to cap the number of predicted trips at 6 predictions, as it is still representative of the dataset.

		Full OD		Work		Home		School	
		NRMSE	MAE	NRMSE	MAE	NRMSE	MAE	NRMSE	MAE
Train set %	80%	0.0105	0.1916	0.0071	0.0906	0.0218	0.0329	0.0063	0.0240
	50%	0.0113	0.1226	0.0089	0.0579	0.0257	0.0205	0.0061	0.0149
	30%	0.0101	0.0752	0.0089	0.0359	0.0223	0.0124	0.0055	0.0090
	10%	0.0109	0.0264	0.0086	0.0120	0.0122	0.0043	0.0047	0.0033

Table 1: Prediction error results for the Full OD matrix estimation, as well for the OD matrices for Work, Home and School.

Table 1 compares the real OD matrices with the predicted output for each level of the training set. First, we compare the **Full OD** matrix, meaning without any segmentation per activity,

and then 3 OD matrices divided by activity, namely **Work**, **Home**, and **School**. We use *NRMSE* (Normalized Root Mean Square Error) and *MAE* (Mean Absolute Error) to evaluate the prediction accuracy of our model, as these metrics provide insights into both the relative error scale and absolute deviations, respectively. What we show with Table 1 is that the *SVGP* model not only provides strong prediction accuracy, but this remains consistent even with reduced data. This is due to its Bayesian nature, which infers a posterior distribution that depends more on the statistical relevance of the data than on its volume. Moreover, the granularity of the presented model, with each prediction using its own *SVGP*, allows for an efficient update using new data, reducing the need for full model retraining. Additionally, if any prediction shows lower accuracy, we can refine its training parameters individually without affecting other *SVGP*s in the model.

To further showcase the capability of this model under data scarcity, Figure 4 shows the absolute error when trying to predict the entire dataset using only 10% of it as the training set for the *SVGP*. While the prediction error is concentrated in a few areas (mainly due to the almost absence of data), 90% of the predicted trips fall within a margin of error considered acceptable, especially considering the training set size. As previously mentioned, this concentration of errors is not a major concern, as the structure of the model allows for individual updates to *SVGP* without the need for full retraining.

Conclusions

This research presented the application and effectiveness of the *SVGP* model for dynamic OD matrix estimation under data-scarcity conditions. The model successfully provides accurate predictions for all trip variables and activities without relying on any assumptions, while showcasing good prediction capability even under data-scarcity conditions. This achievement provides valuable foundations for exploring future theoretical models exploring the relationship between trips and activities. Further research will include the test on a real travel survey, the prediction of transport modes and future travel patterns. In addition, the flexible structure of the model accommodates different types of input data and allows for optimization of individual zones, enabling rapid updates of the *SVGP*s. This could potentially enable the model to be used for real-time forecasting applications in the future.

References

Carvalho, L. (2014). A bayesian statistical approach for inference on static origin-destination matrices in transportation studies. *Technometrics*, 56(2):225–237.

Djukic, T., Flötteröd, G., van Lint, H., and Hoogendoorn, S. (2012). Efficient real time OD matrix estimation based on Principal Component Analysis. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pages 115–121. ISSN: 2153-0017.

Dong, X., Ben-Akiva, M. E., Bowman, J. L., and Walker, J. L. (2006). Moving from trip-based to activity-based measures of accessibility. *Transportation Research Part A: Policy and Practice*, 40(2):163–180.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2021). Bayesian Data Analysis Third edition.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer Texts in Statistics. Springer, New York, NY.

Krishnakumari, P., van Lint, H., Djukic, T., and Cats, O. (2020). A data driven method for OD matrix estimation. *Transportation Research Part C: Emerging Technologies*, 113:38–56.

Nakatsuji, T. (2011). A Comprehensive Approach for Data Scarcity Problem in Real-Time Od Matrix Estimation. *Journal of Japan Society of Civil Engineers*.

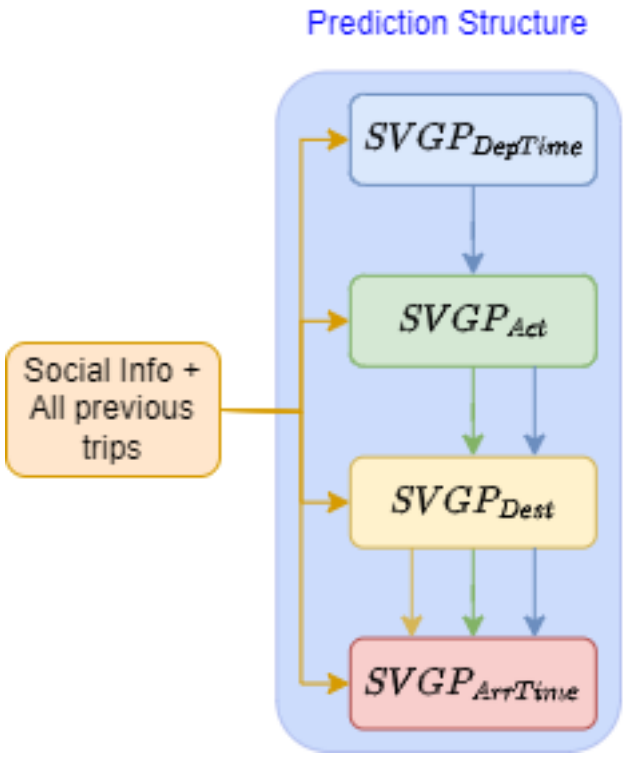
Peterson, A. (2007). *The origin-destination matrix estimation problem: analysis and computations*. Department of Science and Technology, Linköpings universitet, Norrköping.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. *Journal of Machine Learning Research*, 5:567–574.

Viti, F. (2012). State-of-art of O-D Matrix Estimation Problems based on traffic counts and its inverse Network Location Problem: perspectives for application and future developments. In *2012*.

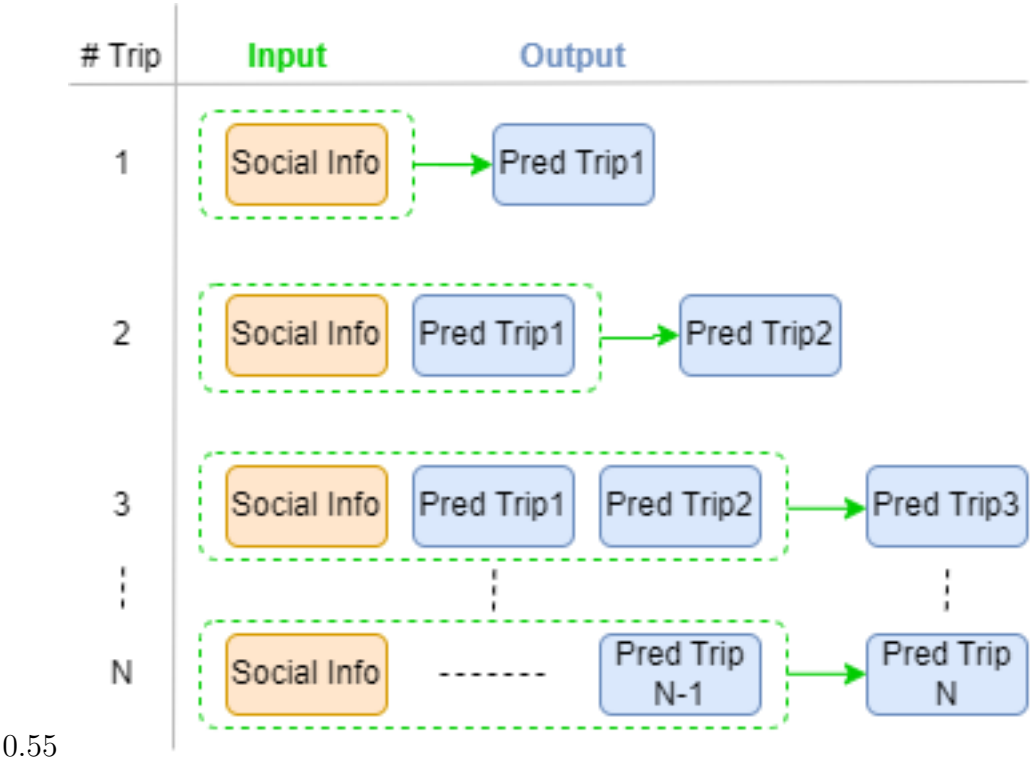
ACKNOWLEDGMENT

We would like to acknowledge the EU projects ACUMEN (N. 101103808) and ODIN (European Regional Development Fund – Programme ERDF "investissement pour la croissance et l'emploi" 2021-2027).



0.35

Figure 1: Single Trip Prediction



0.55

Figure 2: Full Trip Prediction

Figure 3: Proposed trip prediction structure.

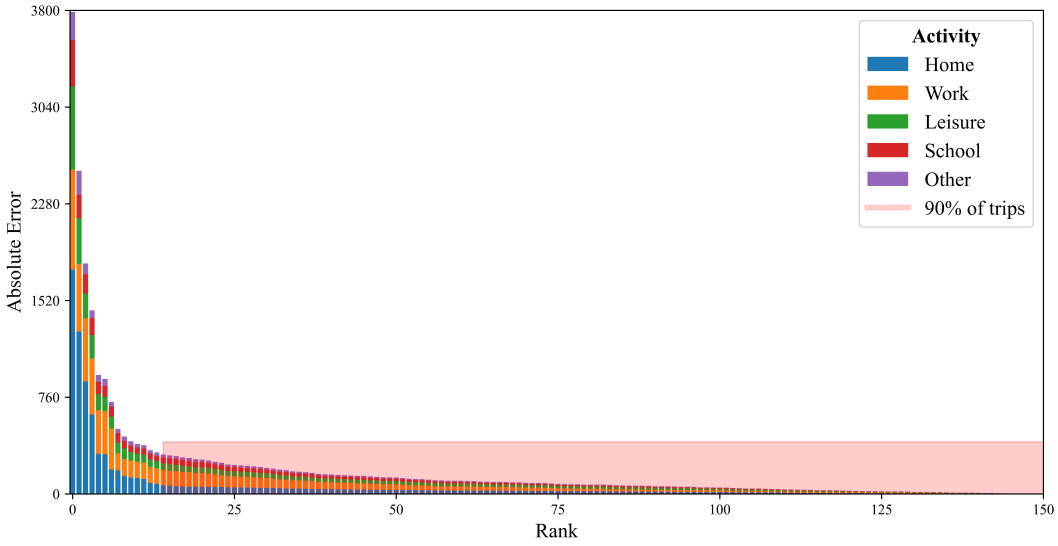


Figure 4: Absolute error ranking for trip activity predictions using 10% Training set