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

Giovanni Tataranno*[1], Federico Bigi[2], and Francesco Viti[3]

[1]PhD Student, Faculty of Science, Technology and Medicine (FSTM), University of Luxembourg, Esch-Sur-Alzette, Luxembourg, L-4364, Email: giovanni.tataranno@uni.lu
[2]Post Doctoral Researcher, University of Luxembourg, Luxembourg
[3]Associate Professor, University of Luxembourg, Luxembourg

Short summary

Origin-Destination (OD) Matrix estimation in the transportation domain is a major challenge, often limited by data availability to estimate reality objectively. This state of "data scarcity" limits the reliability and capability of generalization of models, especially when simplifying assumptions are used to compensate for this lack of information. To address this issue, we applied advanced statistical methods that do not rely on prior assumptions, which are currently one of the most effective solutions to mitigate the effects of data scarcity. For this, Gaussian Process (GP) models have been developed and tested on a large dataset simulating realistic conditions to support this methodology. The test involved training multiple GPs with a training set that contained only 10% of the total dataset, which is the amount of data typically available in an operational and realistic scenario. The model was then used to predict the entire dataset, predicting key variables such as Departure Time, Arrival time, Activity Type, and Destination. The results demonstrate the effectiveness of the proposed method under limited data conditions and confirm the model's ability to generalize to complex and realistic scenarios.
Keywords: data-driven modelling; data scarcity; demand modelling; gaussian process; OD estimation.

1 Introduction

Estimation of the Origin-Destination (OD) matrix is essential in transportation planning, as it provides a detailed understanding of travel patterns and information that supports decision-makers for infrastructure development, traffic management, and urban planning. Traditionally, OD matrices have been estimated either using direct sampling (e.g., travel surveys and traffic counts) or model-based approaches, which rely on theoretical frameworks and socio-demographic data to compute the approximate number of journeys by mode, trip purpose, or time of day Viti (2012). These techniques can produce static (time-independent) or dynamic (time-dependent) OD matrices, Peterson (2007), and are grounded in trip-based or activity-based perspectives of travel demand (Dong et al. (2006)). Despite the large body of research developed on OD estimation, probabilistic data-driven models are emerging as a promising alternative, particularly under data-scarcity scenarios, as these methods aim to infer the underlying distributions rather than relying exclusively on extensive (and sometimes costly) data collection (Nakatsuji (2011)). A common challenge faced in OD estimation is that survey data may be limited in size or quality, leading to potentially inaccurate or incomplete OD matrices. One recent example of addressing data scarcity has been proposed in Krishnakumari et al. (2020), where they leverage three-dimensional supply patterns to simplify OD estimation. However, this method assumes particular route-choice behaviors, restricting flexibility and accuracy. To position our paper within the current research, we propose a dynamic OD estimation method incorporating trip-based and activity components through Gaussian Processes (GPs). Unlike many model-based techniques that require complex fitting assumptions, GPs derive predictions solely from the statistical relevance and noise/uncertainty present in the data (Carvalho (2014)), making no assumptions other than the dataset's reliability. Our method integrates elements of trip-based and activity-based models by sequentially predicting trip attributes — such as start time, destination, and trip purpose. To the best of the authors' knowledge, previous dynamic OD modeling research has not incorporated activity information (i.e., the reason or purpose behind each trip) as directly as we do here using GP modeling with the proposed architecture. The remainder

of the paper is structured as follows: Section 2 quickly introduces travel demand, describing trip and activity-based models, highlighting also the use of discrete choice and Machine Learning (ML) models. Section 3 introduces the GP and how it was implemented for the prediction, specifically by developing its approximate version (Sparse Variational Gaussian Processes, SVGP). Section 4 showcases the result of the application of the GP for OD estimation, discussing the strengths and limitations of the presented methodology. Finally, Section 5 concludes the manuscripts and highlights further directions.

2 Literature Review

Travel demand modeling provides the conceptual and methodological foundation for OD matrix estimation. Broadly, modeling for this problem can be classified into two main types: Trip-Based and Activity-Based model (Chu et al., 2012). While both approaches aim to represent travel behavior in a way that can inform OD flows, they differ in how they treat and simulate individual and collective travel decisions.

- **Trip-Based Models**: In trip-based modeling — often aligned with the four-step framework (trip generation, trip distribution, mode choice, and route assignment) — OD estimation typically emerges in the second step (trip distribution). While widely used in practical applications, trip-based models sometimes lack the behavioral implementation of activity-based approaches, which results in good flow estimation but poor multi-stop trip precision.
- **Activity-Based Models (ABMs)**: Activity-based models focus on why people travel, linking each trip to an underlying activity (work, shopping, leisure, etc.). These models are inherently more detailed, representing full daily or weekly activity patterns, departure times, and destination choices. Activity-based modeling often employs Discrete Choice Models (DCMs) to capture complex decision-making processes regarding activity scheduling and travel patterns. For instance, Ben-Akiva & Bowman (1998) presents one of the earliest comprehensive DCM-based ABMs, while Västberg et al. (2020) formulates a dynamic discrete choice model (DDCM) to represent activity-travel planning as a Markov Decision Process.

In recent years, ML models have gained traction in activity-travel behavior, mainly due to the rise of big data, offering higher precision in prediction. More and more researchers are highlighting ML's potential in activity-travel behavior analysis and prediction, particularly in mode choice decisions prediction (Koushik et al., 2020; Pineda-Jaramillo, 2019). While mode choice is indeed fundamental for the correct functioning of all activity-based models, it can look a bit contradictory, as Miller (2023) points out that all travel demand models, including activity-based ones, face serious challenges in being able to predict activity location choices robustly. The only example found in the literature where this problem is directly tackled is in Hesam Hafezi et al. (2022), where they developed the Scheduler for Activities, Locations, and Travel (SALT), a comprehensive framework for forecasting and replicating individuals' travel behavior based on daily time-use patterns. The SALT model integrates behaviorally-based econometric, ML, and data-mining techniques to create a 24-hour schedule for individuals, proposing a ML model for activity location prediction. Some strong points for using DCMs for activity location are in their capability of embedding models for the prediction, together with the rest of the activity chain, while still providing a strong theoretical background to rely on which, supported by good interpretability of the model, something which instead is a notable issue with ML techniques, as identified in the literature, with few studies trying to address this limitation Koushik et al. (2020). This observation supports our thesis on the current limitations in activity location prediction models within activity-based modeling, underscoring the need for more advanced and reliable methodologies in this area.

3 Methodology

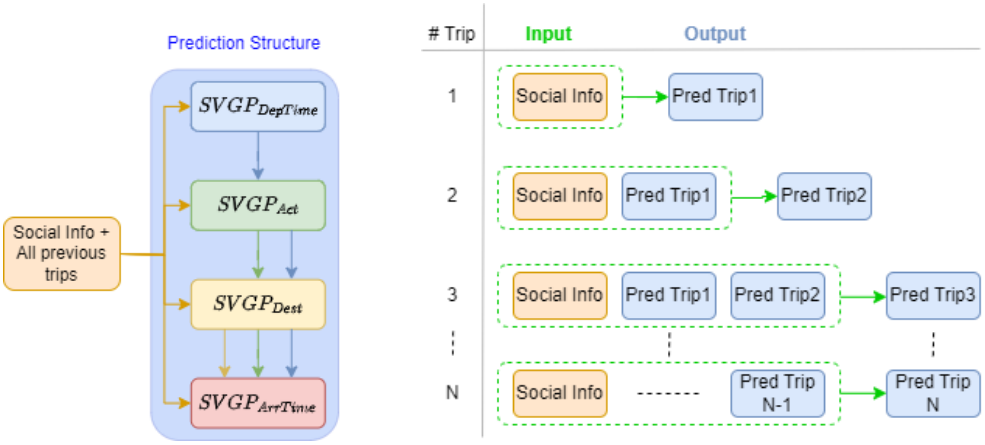
Gaussian Process

It is known and intuitive that the choice of the activity location, which impacts the OD estimation, is never *only* related to the activity that has to be performed and the hour of the day of the trip, but is instead a conditional distribution that depends for example on the relevant sociodemographic characteristics (Hanson & Hanson, 1981), the activities to perform prior and after the current (Scheffer et al., 2021), the choice of the mode of transport etc. To estimate conditional probabilities

while accommodating multiple potentially unrelated distributions, we evaluated the effectiveness of incorporating the GP method into our approach. In probability theory and statistics, a conditional distribution $P(Y \mid X)$ describes how the PDF (Probability Density Function) of an observed variable Y depends on an input X. This distribution is fundamental in OD estimation as we're trying to relate different inputs (e.g. sociodemographic and trip variables) between each other. In parametric approaches (e.g., linear regression), we assume a specific structure $f(x, \phi)$ governed by a finite set of parameters ϕ. While this can lead to smooth and nice functions, if the true form of f lies outside the assumed structure, then we need a strong fitting to get the predictions right, de facto tailoring our model to the dataset, therefore reducing its capability of generalizing. By contrast, nonparametric methods place a prior directly on functions without specifying a fixed form. Examples include Markov Chain Monte Carlo (MCMC), Generative Adversarial Networks (GANs), and Gaussian Processes (GPs) (Nelles, 2020). These methods are more flexible and can yield higher accuracy when the underlying relationship is complex. GP can define a distribution over functions and can model complex, non-linear relationships in data. As per Rasmussen (2004), the GPs can be defined as: *"a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions"*. As the GP prior is non-parametric, this approach puts a prior directly on function values. Nelles (2020) formally defines the GP model as using a GP as a prior on f. A GP prior on f means that a priori the joint distribution of a collection of function values $f = [f(x_1), \ldots, f(x_{m_0})]^\top$ associated with any collection of m_0 inputs $[x_1, \ldots, x_{m_0}]^\top$ is defined by a multivariate normal (Gaussian) distribution $p(f|X, \psi) = N(f|m, K)$ with mean m and covariance matrix K. The stochastic process in GPs defines a distribution over functions, which is what we are aiming to implement when it comes to predicting trips based on conditional distributions. Training a GP involves learning the hyperparameters by optimizing the negative log marginal likelihood, which adjusts the model to the data probability distribution. Unlike conventional fitting, which forces a model to minimize the residuals between predicted and observed values, this approach maximizes the likelihood of the data under the model's prior assumptions, selecting the most probable functions that align with the observed data. Nonetheless, all these model capabilities come with a high computational complexity ($O(n^3)$). That is why in real applications GPs are often replaced by its variant, 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 different advantages, the most important of which is the 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, as the SVGP accuracy does not depend on the amount of data available, but on its *statistical relevance* (Titsias, 2009).

GP Implementation

For this study, we use multiple SVGPs to predict the desired variables. Specifically, for each trip, several SVGP are employed to predict the **Start Time**, **Destination**, **Activity Type**, and the **Arrival Time**. 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 in Figure 1b. 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.



(a) Single Trip Prediction (b) Full Trip Prediction

Figure 1: Proposed trip prediction structure.

Figure 1 shows the full framework of our model to predict a full trip for each individual. Figure 1a 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 given as inputs, is then fed to a second SVGP to predict the Activity Type that has to be performed. The same iterative approach is applied to predicting the Destination and Arrival Time. Figure 1b shows 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. The problem faced when working with SVGPs for 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 (Destination Zone and Activity Type) and regression predictions (Departure and Arrival Times). For the discrete-choice predictions, as only one output is provided by each model, we needed to train one SVGP for each choice option. The SVGPs 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, we trained one SVGP for each. This whole process is then repeated for each possible trip that is available in the dataset, as shown in Figure 1b. By then chaining these predictions, we were able to obtain the final OD matrices.

4 Results and Discussion

Results

To test the effectiveness of the presented SVGP's architecture, we used it to predict an entire dataset starting from a 10% training set, randomly selected through uniform sampling. Figure 2 shows the distribution of the data in the full dataset (in blue) versus the training set (in red).

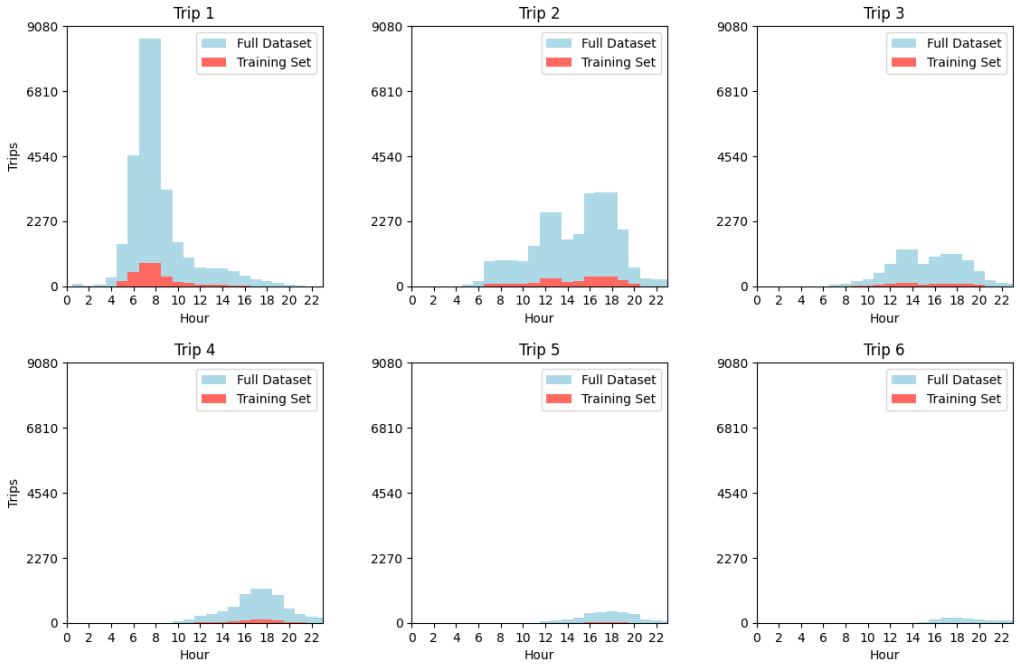


Figure 2: Full Dataset and Training Set comparison.

The comparison between the activities predicted by the model and those in the dataset is shown in Figure 3. Activities were analyzed in terms of frequency and distribution to highlight any discrepancies between the model predictions and the actual data.
Table 1 shows an overview of the errors for each of the variables predicted by the SVGP, using the following metrics to analyze distributions for synthetic populations, as per Bigi et al. (2024):

- Normalized Root Mean Squared Error (NRMSE): measures the normalized root mean squared error between the GP prediction and the data set.
- Jensen-Shannon Divergence (JSD): quantifies the similarity between the distributions of each variable.
- Hellinger Distance (HELL): measures the distance between the above distributions.

Trip	Metrics	Dest. Zone	Act. Type	Arr. Time	Dep. Time
1	NRMSE	0.04	0.00	0.14	0.18
	HELL	0.10	0.01	0.12	0.14
	JSD	0.01	0.00	0.01	0.02
2	NRMSE	0.02	0.02	0.17	0.17
	HELL	0.10	0.04	0.10	0.10
	JSD	0.01	0.00	0.01	0.01
3	NRMSE	0.08	0.14	0.16	0.18
	HELL	0.17	0.07	0.08	0.08
	JSD	0.03	0.00	0.01	0.01
4	NRMSE	0.05	0.03	0.17	0.17
	HELL	0.19	0.06	0.06	0.07
	JSD	0.03	0.00	0.00	0.00
5	NRMSE	0.13	0.28	0.22	0.23
	HELL	0.42	0.25	0.06	0.07
	JSD	0.14	0.06	0.00	0.00
6	NRMSE	0.13	0.09	0.27	0.33
	HELL	0.53	0.25	0.08	0.10
	JSD	0.22	0.05	0.01	0.01

Table 1: GP's predictions metrics

The results show a relative consistency for all the metrics in almost all the trips, with an exception for Trip 5 and Trip 6. For the latter, there is a significant discrepancy between the NRMSE, JSD, and HELL metrics. But why this occurs? Looking at the Destination Zone and Activity

Type, we observe that the NRMSE error is relatively small, but with a marked increase in the JSD and HELL. This suggests that although the overall model error for these variables is low, the SVGP's ability to accurately reproduce their distributions has decreased. For the Arrival Time and Departure Time instead, the NRMSE increases with lower JSD and HELL. This means that the model manages to represent these variables well, even in the presence of significant noise in the original data. In other words, the GP not only accurately represents the variables, but also captures the noise inherent in the data. A higher NRMSE is not necessarily a negative result. Although there is a tendency to minimize the NRMSE, in realistic scenarios the 'true' reality (here represented by the data set) is not fully known, and therefore the degree of error committed by the model cannot be accurately determined. Furthermore, the ability to model noise in the data is a non-trivial aspect that brings the model closer to the realistic and chaotic behavior observed in everyday life. The increase in NRMSE reflects the ability of the model to reproduce noise. Since no two instances of noise can be identical, a model that also captures intrinsic noise will inevitably show an increase in NRMSE. However, this does not indicate an error in the model, but rather its adherence to the complexity of the observed data. Thus, the SVGP predicts the variables well, but intrinsic noise is an unavoidable component of the prediction. Digging more inside the results, Figure 3 shows the comparison between the GPs prediction and the Full Dataset activities.

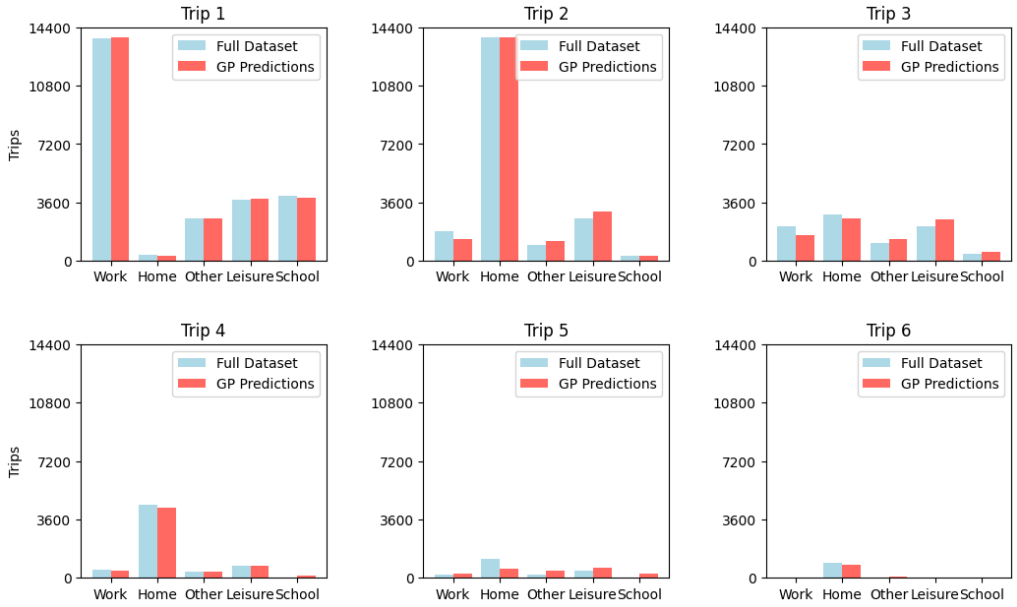


Figure 3: Comparison between GP's predictions and Full Dataset activities.

What can be observed is that there is a minimum discrepancy between the predictions and the real dataset, even for Trip 5 and 6 where few data area are available. This suggests that among the input data provided to the SVGP, there may be one or more variables on which each activity appears to be strongly dependent.

Discussion

This approach, based on a chain architecture as presented in Section 3, introduces a major challenge: error propagation. In particular, the concatenation is not limited to variables, but also includes individual trips. This can also be observed in Table 1, where the error between trips tends to increase for most of the metrics, with the SVGPs trying to distinguish between noise and relevant data, representing some error propagation filtered by the statistical relevance. In addition, as the training set contains less and less information about the trips further in the sequence, the expected error increases proportionally and tends to be more pronounced for the last trips.

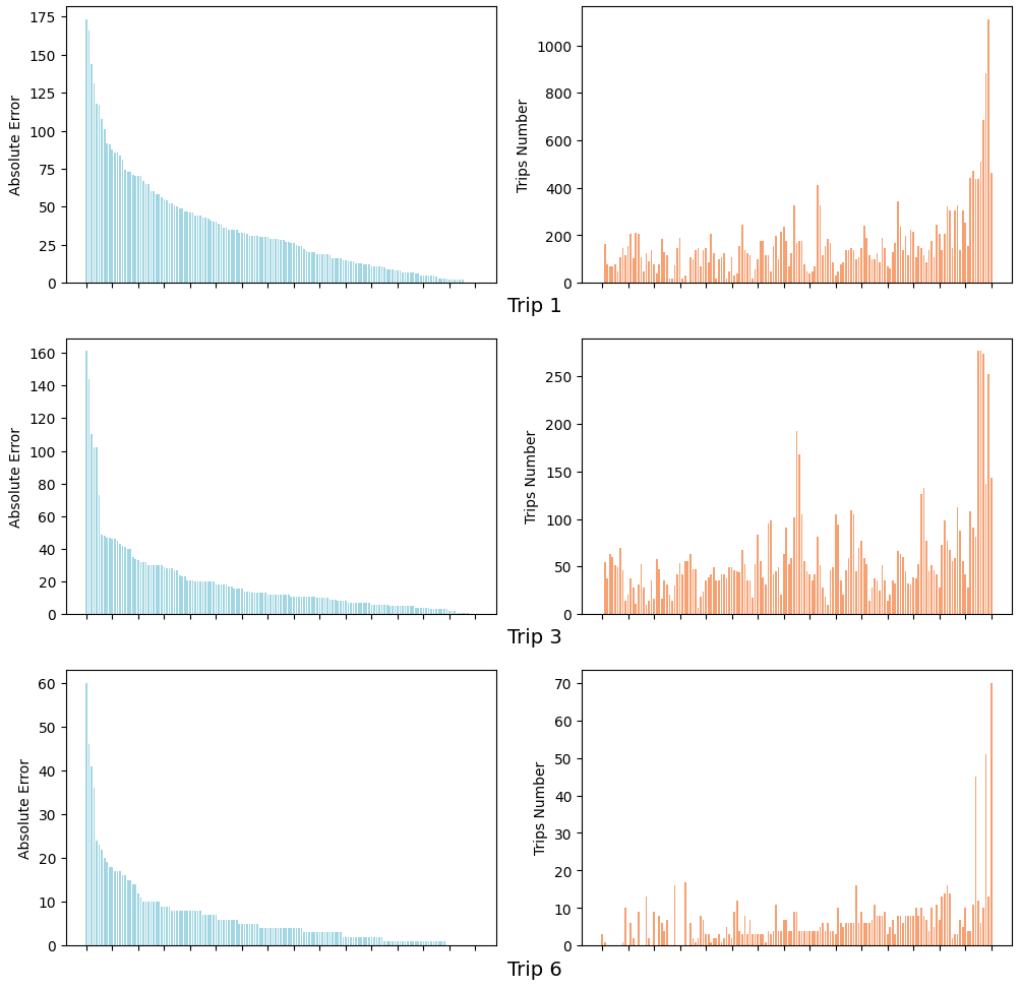


Figure 4: GP's absolute error for destination (blue) and overall number of trips (orange) for Trips 1, 3, and 6.

Figure 4 shows the absolute error (blue) made by the SVGP model in predicting a specific Destination Zone against the full available data (orange). When looking at Trip 1, 3, and 6, there is no visible propagation of the error. Instead, the observed trends are almost identical, with the error concentrated exclusively in areas where there is little or no data. This behavior occurs because the GP reconstructs a conditional probability incorporating an explicit quantification of the noise present in the Training Set, as well as the one expected in the input. This aspect is governed by a fundamental hyperparameter that allows the model to deal effectively with the noise, significantly reducing the effects of error propagation. Furthermore, this capability allows the GP to select areas where there are few or no associated trips in the training set, overcoming one of the main challenges of data scarcity. While this feature is a significant advantage in data scarcity contexts, the trade-off for such flexibility is the presence of a small error in areas where, in reality, there are no trips. This behavior highlights the limitations and opportunities of the model under conditions of incomplete data, and provides useful insights for further improvements and applications.

5 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. 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 SVGPs. This could potentially enable the model to be used for real-time forecasting applications in the

future. Further research will include the prediction of transport modes and future travel patterns.

Acknowledgements

This work is financially supported by the EU-FEDER project grant ODIN (n. 2023-1-8) and the EU Horizon Europe ACUMEN (n. 101103808)

References

Ben-Akiva, M. E., & Bowman, J. L. (1998). Activity Based Travel Demand Model Systems. *Equilibrium and Advanced Transportation Modelling*, 27–46. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4615-5757-9_2 doi: 10.1007/978-1-4615-5757-9{_}2

Bigi, F., Rashidi, T. H., & Viti, F. (2024). Synthetic Population: A Reliable Framework for Analysis for Agent-Based Modeling in Mobility. *Transportation Research Record*. doi: 10.1177/03611981241239656

Carvalho, L. (2014). A bayesian statistical approach for inference on static origin-destination matrices in transportation studies. *Technometrics*, *56*(2), 225–237. doi: 10.1080/00401706.2013.826144

Chu, Z., Cheng, L., & Chen, H. (2012). A review of activity-based travel demand modeling. *CICTP 2012: Multimodal Transportation Systems - Convenient, Safe, Cost-Effective, Efficient - Proceedings of the 12th COTA International Conference of Transportation Professionals*(December), 48–59. doi: 10.1061/9780784412442.006

Dong, X., Ben-Akiva, M. E., Bowman, J. L., & Walker, J. L. (2006, February). Moving from trip-based to activity-based measures of accessibility. *Transportation Research Part A: Policy and Practice*, *40*(2), 163–180. Retrieved 2024-10-11, from https://linkinghub.elsevier.com/retrieve/pii/S0965856405000820 doi: 10.1016/j.tra.2005.05.002

Ghosh, J. K., Delampady, M., & Samanta, T. (2006). *An Introduction to Bayesian Analysis*. New York, NY: Springer. Retrieved 2024-10-02, from http://link.springer.com/10.1007/978-0-387-35433-0 doi: 10.1007/978-0-387-35433-0

Hanson, S., & Hanson, P. (1981). The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. *Economic Geography*, *57*(4), 332–347. doi: 10.2307/144213

Hesam Hafezi, M., Sultana Daisy, N., Millward, H., & Liu, L. (2022). Framework for development of the Scheduler for Activities, Locations, and Travel (SALT) model. *Transportmetrica A: Transport Science*, *18*(2), 248–280. doi: 10.1080/23249935.2021.1921879

Koushik, A. N., Manoj, M., & Nezamuddin, N. (2020, 5). Machine learning applications in activity-travel behaviour research: a review. *Transport Reviews*, *40*(3), 288–311. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01441647.2019.1704307 doi: 10.1080/01441647.2019.1704307

Krishnakumari, P., van Lint, H., Djukic, T., & Cats, O. (2020, April). A data driven method for OD matrix estimation. *Transportation Research Part C: Emerging Technologies*, *113*, 38–56. Retrieved 2024-09-25, from https://www.sciencedirect.com/science/article/pii/S0968090X18317388 doi: 10.1016/j.trc.2019.05.014

Miller, E. (2023). The current state of activity-based travel demand modelling and some possible next steps. *Transport Reviews*, *43*(4), 565–570. Retrieved from https://www.tandfonline.com/action/journalInformation?journalCode=ttrv20 doi: 10.1080/01441647.2023.2198458

Nakatsuji, T. (2011, January). A Comprehensive Approach for Data Scarcity Problem in Real-Time Od Matrix Estimation. *Journal of Japan Society of Civil Engineers*. Retrieved 2024-10-21, from https://www.academia.edu/123614533/A_Comprehensive_Approach_for_Data_Scarcity_Problem_in_Real_Time_Od_Matrix_Estimation

Nelles, O. (2020). Gaussian Process Models (GPMs). *Nonlinear System Identification*, 639–708. doi: 10.1007/978-3-030-47439-3{_}16

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

Pineda-Jaramillo, J. D. (2019, 10). A review of machine learning (ML) algorithms used for modeling travel mode choice•. *DYNA (Colombia)*, *86*(211), 32–41. doi: 10.15446/DYNA.V86N211.79743

Rasmussen, C. E. (2004). Gaussian Processes in machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3176*, 63–71. doi: 10.1007/978-3-540-28650-9{_}4

Scheffer, A., Connors, R., & Viti, F. (2021, 1). Trip chaining impact on within-day mode choice dynamics: Evidences from a multi-day travel survey. *Transportation Research Procedia*, *52*, 684–691. doi: 10.1016/J.TRPRO.2021.01.082

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

Västberg, O. B., Karlström, A., Jonsson, D., & Sundberg, M. (2020). A dynamic discrete choice activity-based travel demand model. *Transportation Science*, *54*(1), 21–41. doi: 10.1287/trsc.2019.0898

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.*