TOWARDS A VALIDITY FRAME OF MULTI-MODAL SURROGATE MODELS FOR TRAFFIC SIMULATION

Raheleh Biglari[a], Claudio Gomes[b], and Joachim Denil[a,c]

[a]Faculty of Applied Engineering, Cosys-Lab, University of Antwerp, Belgium
[b]Department of Electrical and Computer Engineering, Aarhus University, Denmark
[c]Flanders Make@Uantwerpen, Belgium

ABSTRACT

Multi-modal systems, such as urban traffic networks, require considering specific modeling considerations for varying operational and environmental conditions. Surrogate modeling offers a practical solution to represent these modes through different surrogate models. However, switching between modes needs careful management of the models' experimental frames. In multi-model setups, overlap between these frames introduces unnecessary complexity and ambiguity in model selection. The concept of validity frame of multi-modal surrogate models is the model's experimental frame plus the process allowing to change this frame. This paper introduces an approach to define such a validity frame ensuring that the union of all experimental frames covers the full system operating domain without redundancy and focusing on clean separation between models and interpretable mode switching. The proposed ensemble of surrogate models, coupled with adaptive techniques, represents a significant step forward in the performance modeling of dynamic and computationally expensive models.

Keywords: Surrogate Modeling, Experimental Frame, Validity Frame, Traffic Networks.

1 INTRODUCTION

Complex systems such as traffic systems, climate simulations, or multi-physics industrial processes include different system and environment modes, each requiring specific modeling considerations. For example, the traffic system of a city needs to consider multiple modes, such as peak rush-hour congestion and off-peak flow patterns, to ensure a complete representation of real-world dynamics. To address this, surrogate modeling, particularly machine learning-based surrogate modeling [1] serves as a practical method to develop simplified models with reduced computational cost for different modes, but switching between modes requires careful management of the models' experimental frames. ***The experimental frame of a model*** *is the set of inputs for which the model can provide valid predictions*. One of the challenges is to ensure that the experimental frames of the models are non-overlapping. This is crucial to avoid the extra complexity in the models of each mode, and to make it clearer which model to use at any given time.

The validity frame of a model *is defined as its experimental frame, plus a process that allows us to change the experimental frame*. Such process must include minimizing the overlap between the experimental frame of the model with respect to the other experimental frames in a multi-model setup.

In this paper, we show one possible approach to define such a validity frame. To illustrate this, we create multiple experimental frames, each belonging to a surrogate model. Each surrogate model is specifically

Proc. of the 2025 Annual Modeling and Simulation Conference (ANNSIM'25)
May 26-29, 2025, Universidad Complutense de Madrid, Madrid, Spain
J.L. Risco-Martín, G. Rabadi, D. Cetinkaya, R. Cárdenas, S. Ferrero-Losada, and A. Bany Abdelnabi, eds.

designed to operate within a distinct validity region, enabling efficient and accurate predictions across different system and environment modes. In addition, we describe the process used to define the experimental frame of each model.

As part of our methodology, we also address the key challenge of building complete and non-redundant experimental frames. That is, the union of our models' experimental frames needs to form a complete system operating domain. Achieving this requires access to a complete dataset, which is derived by ensuring the proper distribution of data across the system's operational and environmental modes. This partitioning approach ensures complete coverage of model validity while reducing computational costs and improving scalability.

We choose the traffic system as a case study using SUMO (Simulation of Urban MObility) [2], a renowned open-source micro-traffic simulator, suitable for optimizing traffic light configurations. Additionally, available domain knowledge about the two main system's operating modes enables us to easily propose two surrogate models.

The presentation of our work will follow this outline: Section 2 discusses the background of this research, then in Section 3 an overview of related work, highlighting existing research and identifying gaps that this study aims to address. Section 4 details the proposed approach and presents the experiment with the experimental setup, datasets, and results. Section 5 discusses the main limitations of the approach. Finally, Section 6 concludes the paper by summarizing the key contributions, addressing the research objectives, and suggesting potential directions for future work.

2 BACKGROUND

2.1 Surrogate Modeling

Surrogate models have become popular in engineering design due to their capability to approximate computationally expensive engineering systems. A surrogate model is a model that is less computationally expensive and faster to run [3]. Selecting the appropriate surrogate model structure is critical in practice because of the multiple variants and variances in their hyper-parameter configurations. The process of designing complex engineered systems generally entails the exploration of the whole design space, which requires many runs of expensive simulations. In order to reduce the computational cost associated with these simulations, researchers have developed surrogate modeling techniques to approximate these simulations and significantly reduce computational costs.

The selection of a surrogate model is frequently approached in one of two ways [4]: 1) Manual comparison-based surrogate model selection. 2) Evolutionary algorithm-based surrogate model selection. Selecting the right surrogate model is important, as it directly affects the accuracy, efficiency, and reliability of the model. An inappropriate model can lead to inaccurate predictions, poor decisions, or increased computational costs. Factors such as the properties of interest, design space complexity, available resources, and the balance between accuracy and cost must guide model selection. An appropriate surrogate model minimises computational cost while keeping accuracy. Careful evaluation of the model's characteristics and compatibility with the properties of interest and goals are essential to ensure optimal performance.

2.2 Validity frame

Surrogate models are not necessarily applicable to the entire range of contexts that a system may encounter. It is necessary to identify a frame within which the model stays applicable. The concept of frames originates from the early 1980s, introduced as "experimental frames" by Zeigler [5]. An experimental frame represents a set of conditions under which a system is experimented. Building on this idea, validity frames explicitly encode the contexts in which a model provides valid results for specific properties relative to a real-world system [6, 7]. The concept of validity frames formalizes the contexts in which a model provides reliable results, and it is closely related to the methods for verification and validation in simulation modeling defined by Sargent [8].

While we extend the idea of validity frame to machine learning-based surrogate modeling, the validity frame of a model is defined as its experimental frame plus the process for adapting or redefining that frame. This process must include a mechanism to minimize overlap with the experimental frames of other models in a multi-model setup.

3 RELATED WORK

The increasing complexity of simulations and the need for efficient computational methods have driven significant advancements in techniques for surrogate modeling, adaptive abstraction, and runtime model integration. These approaches aim to balance precision, computational cost, and scalability while maintaining the validity of simulation results across diverse domains such as agent-based modeling, traffic systems, and building energy simulations. We briefly summarize relevant work in machine learning for surrogate modeling and adaptive abstraction techniques, emphasizing their relevance to our focus on ensuring the completeness and reliability of experimental frames in dynamic simulation environments. The problem of dynamically switching surrogate models can be seen as a hybrid system identification problem, and we refer the reader to [9] for an introduction to the topic.

3.1 Machine Learning for Surrogate Modelling

The work in [1] investigates the use of machine learning methods as surrogate models for agent-based models (ABMs), focusing on reducing computational costs in sensitivity analysis and parameter calibration. Surrogate models replace computationally expensive ABMs with statistical models that replicate their behavior, enabling faster evaluations. The study compares various ML methods, including neural networks, gradient-boosted trees, and Gaussian processes, and finds that artificial neural networks (ANNs) and gradient-boosted trees outperform traditional Gaussian process surrogates, especially in capturing non-linear or chaotic behaviors. It advocates for further development of ML-based surrogate models, particularly for complex, policy-relevant ABMs, and suggests that these techniques could democratize access to detailed model analyses in resource-constrained environments. Moreover, Ribeiro et al. introduce the local surrogate model which is a simplified model trained to approximate a complex model's behavior around a specific input or region [10].

For an extensive review of constructing neural network-based models for simulating dynamical systems, we highlight the surveys in [11]. These surveys provide a comprehensive discussion of data-driven modeling techniques and challenges, particularly those related to dynamical systems.

Similar to our work on quantifying and characterizing validity frame from a methodological point of view, we highlight the work [12, 13, 14] where validity frames are computed/quantified for different models. With a focus on neural networks, the work of [13] highlights the importance of distinguishing interpolation (reliable predictions within the training data range) from extrapolation (unreliable predictions outside the data range). To tackle this, the study proposes a methodology to calibrate novelty detection algorithms specifically for extrapolation detection in ML models. This approach focuses on ensuring reliable model predictions within the validity region while highlighting the risks of extrapolation in building energy systems.

3.2 Adaptive Abstraction and Approximation

The work of Bosmans et al. [15] explores a strategy for reducing computational costs in large-scale simulations, particularly those modeling complex systems like traffic networks or IoT systems. It introduces an opportunistic model approximation technique leveraging information theory to identify and abstract areas of the simulation that contribute minimally to global behavior. These low-information regions are dynamically transformed into less detailed representations, effectively reducing computational demand while preserving the validity of the emergent behaviors at the global scale. The technique involves entropy-based transformations, where entropy measures the informational complexity of simulation areas. Regions with lower entropy—indicating less influence on overall system behavior—are approximated to save computational resources.

The paper in [16] discusses adaptive abstraction techniques in agent-based simulations, emphasizing their potential to enhance computational efficiency without significantly compromising result accuracy. Adaptive abstraction dynamically switches between levels of detail in simulations, depending on context, to strike a balance between execution speed and model precision. Using a traffic simulation case study, the authors explore two major patterns—PUPPETEER and ZOOM—for aggregating and disaggregating agent data. While PUPPETEER preserves detailed individual states, ZOOM relies on statistical aggregation, sacrificing some detail for computational efficiency. Results demonstrate that adaptive abstraction can effectively reduce computational costs while maintaining valid emergent behaviors, such as traffic jams, though domain-specific design decisions significantly influence outcomes.

Bosmans et al. [17] serves as motivation to the importance of our own work. It presents a framework for adaptive multi-level traffic simulation, combining micro-level (individual vehicles) and meso-level (aggregated groups) models to balance computational cost and simulation accuracy dynamically. By leveraging the MAPE-K feedback loop and experimental frames, the framework assesses the validity of meso-level models in real-time and switches to micro-level models when higher fidelity is required. Using a custom-built simulation environment, the authors demonstrate the approach in urban traffic scenarios, showcasing significant improvements in computational efficiency while maintaining simulation validity. Results indicate that the dynamic approach outperforms static hybrid simulations in balancing performance and error rates.

With a focus on run-time model swapping, the work in [18] explores a mechanism for dynamically integrating and swapping models during simulation time without disrupting the simulation. It introduces a framework that facilitates the replacement of individual models or structural changes in coupled-models, enabling updates based on evolving system requirements. While the proposed method can be used for adaptively varying the abstraction level at runtime, as done in [16] and [17], our work is more focused on identifying the triggering conditions for model swapping.

Our work builds on prior research in surrogate modeling and adaptive abstraction by shifting the focus toward constructing non-overlapping experimental frames and formalizing the concept of a validity frame. Unlike existing methods that often rely on runtime model switching or overlapping data regions, we propose a data-driven approach to partition the validity region upfront. This enables cleaner model boundaries, reduces redundancy, and supports more interpretable and efficient surrogate modeling in multi-modal systems, as demonstrated in our traffic simulation case study.

4 APPROACH

This section introduces an approach to define the validity frame of the surrogate models using the asset of a machine learning-based surrogate modeling through a traffic simulation use case. It also details the experimental setup, data generation process, and model development, addressing key challenges such as ensuring dataset diversity, defining surrogate models' experimental frames, and minimizing redundancy in training data.

4.1 Problem Description

Traffic systems are complex systems, and their simulations necessitate computationally expensive models, constraining their scalability for practical implementations. We aim to predict the average total waiting time of all cars in the traffic system by simulating a large part of Wilrijk, a district in Antwerp, Belgium, as seen in Figure 1, with deep neural network models trained on SUMO-generated data.

Figure 1: Covered traffic system in SUMO.

4.2 Experimental Setup

We model the traffic system for our intended map, by using SUMO. SUMO is a micro-level traffic simulator, meaning each vehicle and its dynamics are modeled individually. It has several main advantages. First, it allows us to generate a large-scale map of a certain area. The road network is imported directly from OpenStreetMap (OSM) [19], eliminating the need to build a network from scratch. Next, using the Traffic Control Interface (TraCI), we can connect to an ongoing simulation. This allows us to collect information about the simulation and to update the ongoing simulation. Traffic scenarios are generated using SUMO's randomTrips.py, which generates a set of random trips for a given network. In the absence of real-world data; therefore, we generate data for our traffic system and run the simulation. SUMO allows us to keep a clear record of the simulation results. We can view the total average waiting time of all vehicles in the system.

4.3 Surrogate Models

Drawing from domain knowledge, we focus on two distinct traffic modes: rush-hour mode, characterized by congestion and high traffic density, and off-peak mode, with lower traffic density. For each mode, we develop a dedicated surrogate model: the low-density model, trained on low-density data, and the high-density model, trained on high-density data. While the low-density model performs better during off-peak periods, the high-density model is expected to be more accurate when the number of cars surpasses a certain threshold, making it more suitable for rush-hour conditions.

The input vector for both networks will contain the combination of the traffic light configurations and the number of cars per road. The output will be a single value representing the total waiting time for the next

500 seconds derived from running the simulation. By dividing the data into separate density categories, each model can specialize in a specific region of the problem.

4.4 Data Generation

Since the training data is not readily available for our experiment, we generate training data for the surrogate models by simulating several traffic scenarios with differing traffic densities and traffic light configurations. To generate low-density scenarios, we use SUMO's randomTrips.py and --*insertion-density* argument = 15. This argument is the number of vehicles per hour per kilometer of road that the user expects. And to generate high-density scenarios, we use SUMO's randomTrips.py and --*insertion-density* argument = 25. These values are selected based on calculations of the number of cars at high-demand intersections such as how shown in Figure 2. Through testing different settings, we observed that a value of 25 resulted in significantly denser traffic, whereas a value of 15 produced lighter traffic flow. Although there may be alternative methods for identifying high-demand regions that could further enhance this research. This is, however, outside the scope of this paper.

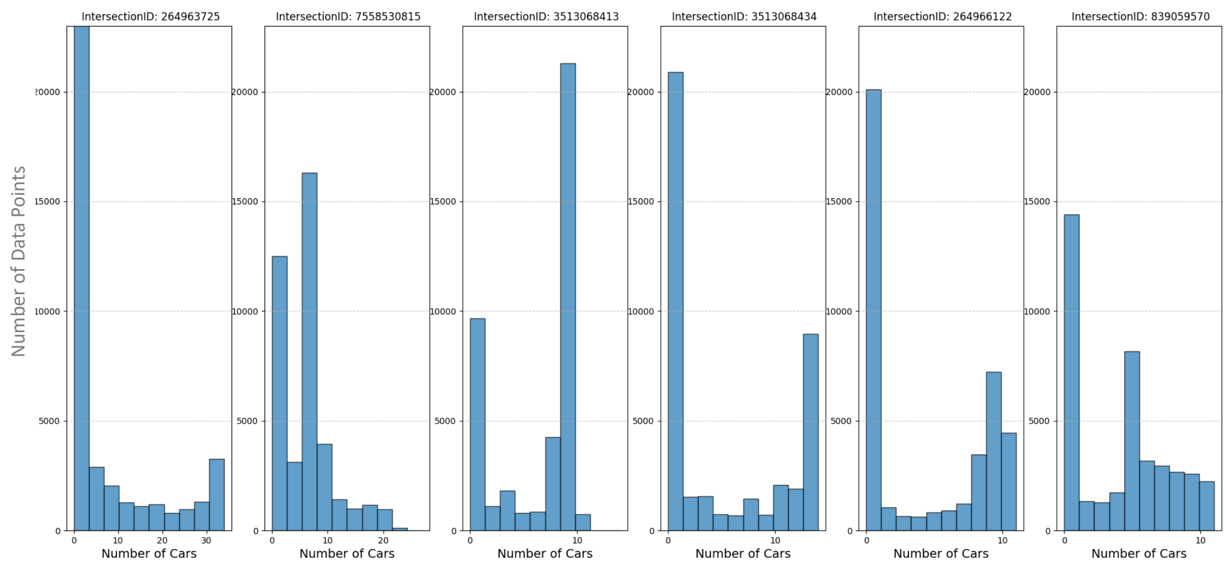

Figure 2: Distribution of vehicle counts across high-traffic intersections over 600 simulation runs demonstrating balanced and diverse data coverage.

For each scenario, we saved 15 states. These states serve as the input values in our dataset. Each of these states is then simulated for 500 time steps. The waiting times after the simulation are the output values in the dataset. Separate datasets are prepared for the low-density and high-density models, ensuring each model is trained on data relevant to its intended use case.

4.5 A Complete and Balanced Experimental Frame

For a model to be accurate, it needs to operate within a clearly defined experimental frame. These boundaries outline the range of conditions and scenarios it is designed to handle accurately.

An essential element of establishing an experimental frame is ensuring it includes the complete range of scenarios the model is expected to face in its intended use. This does not imply that the model must include every potential scenario, but rather, it should reflect the range of actual conditions it is likely to manage. In machine learning-based surrogate modeling, if the model's training data or calibration does not cover crucial scenarios, such as unusual or extreme events, it may fail when these situations occur. For example, a traffic simulation model considered exclusively normal traffic hours may have difficulty predicting traffic during a concert in the city center or a football match. A complete experimental frame needs:

Understanding the system: Clearly identifying the system's goals, properties of interest, conditions, and modes that describe the system under study.

Balanced Data: A complete experimental frame needs balanced data, which does not mean all possible situations but rather representative and diversified data. Therefore, there is a need for uniformity in data, which means the dataset needs to be uniformly distributed over situations, scenarios, and features depending on the specific context and application of the system under study.

In this experiment, the dataset must accurately represent varied traffic configurations and include the most pertinent and commonly observed traffic light configurations. Moreover, careful consideration is devoted to preserving consistency in data distribution to prevent biases. This approach guarantees the model stays robust and reliable across various expected operational scenarios. A diverse dataset helps accurately reflect the system's behavior, avoiding the overrepresentation of specific subsets and promoting a balanced understanding of traffic dynamics.

To ensure that we have balanced data in this use case, we compare high-demand intersections with heavy traffic. We visually inspect the distribution with a histogram shown in Figure 2, which verifies data distribution aligns with the required balance and diversity. We could focus on all possible traffic light configurations, but in practice it's not computationally possible since there are 186 traffic lights in this map, each having at least 3 phases, and for each phase, there are many duration possibilities. Therefore, we chose to focus on high-demand intersections.

Furthermore, to build a complete experimental frame, the combined experimental frames must cover the complete operational domain of the system under study. This ensures the surrogate models collectively keep the accuracy and robustness necessary for reliable system representation across all operational scenarios.

4.6 Minimally overlapping experimental Frames

In the context of surrogate models, it is essential to define distinct experimental frames for each model. To ensure clear model selection and eliminate ambiguity, these experimental frames must be minimally overlapping. This guarantees that each model matches perfectly to a unique subset of the problem space, consequently avoiding prediction conflicts.

Achieving minimally overlapping experimental frames requires creating corresponding minimally overlapping datasets for training the models since we are using surrogate models. This means the data used to train one model should not overlap with the data used for another. In this experiment, we divided the dataset into two categories, shown in Figure 3.

Additionally, avoiding redundancy helps ensure that the surrogate models are trained on the most informative data, allowing them to perform accurately and efficiently within their respective experimental frames. To achieve low redundancy in the dataset, there are two main approaches: 1) Randomly removing data points from the dataset. 2) Removing data points with high similarity to others. In this research, we use the second approach, which prioritizes the removal of similar data points to ensure diversity in the dataset. To determine similarity, we define a notion of distance between data points. We use the most familiar distance metric, n-dimensional Euclidean distance. The number of roads in our network plus the number of phases for all traffic lights in our map is n and the calculation of distance takes $O(n)$.

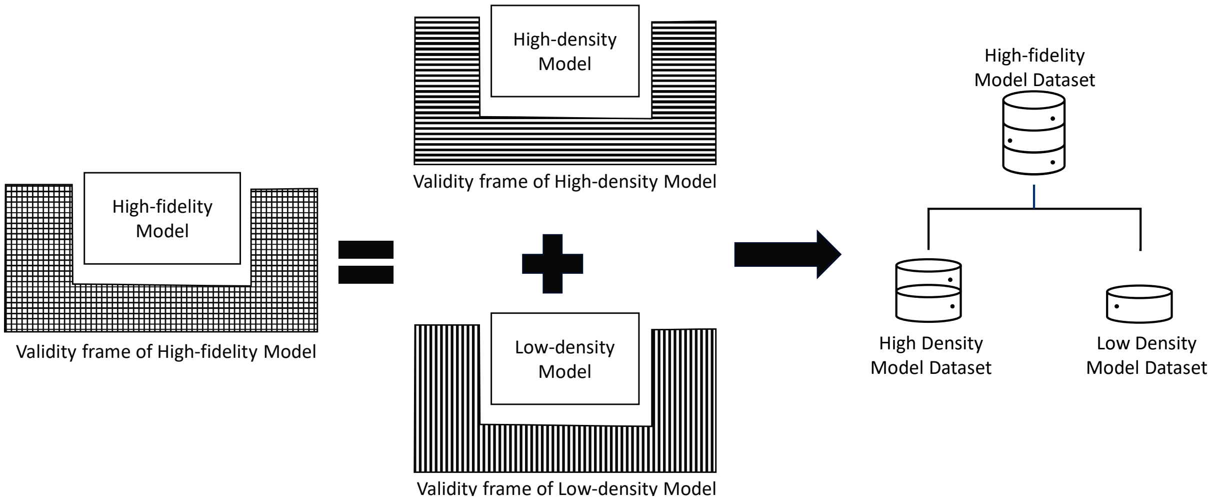

Figure 3: Minimally overlapping experimental frame and minimally overlapping datasets.

In this experiment, each data point represents a combination of the traffic light configurations and the number of cars per road. The Euclidean distance between two points is calculated based on these attributes. If the distance between two points is small, they are considered similar, and one of them is removed to reduce redundancy. In this example, we eliminate 8,046 out of 8,999 samples identified as redundant, based on observations from the heatmap graph in Figure 4, where a distance of 50 appears to indicate data points that are too close to each other. A similar approach is applied to the low-density dataset, where we remove 7,710 out of 8,999 data points. It is important to note that the distance metric is domain-specific and varies depending on the particular use case. The calculation of these distances between P data points is $O(P^2)$.

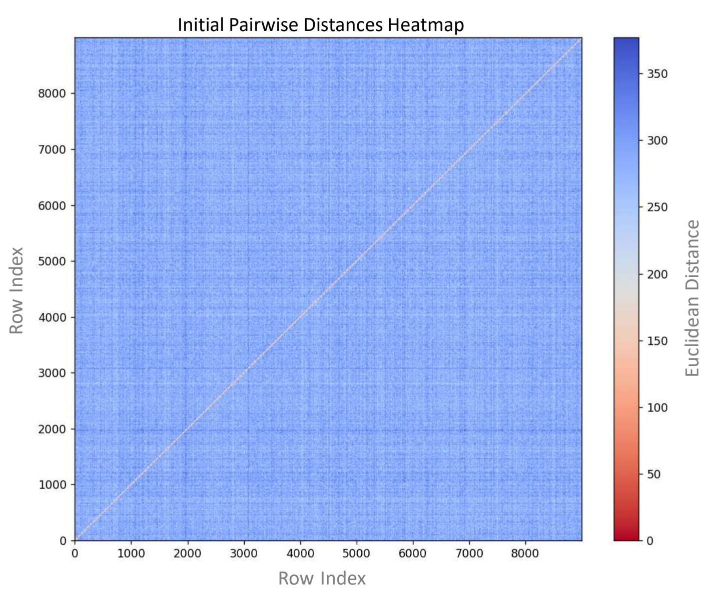

Figure 4: Pairwise distance heatmap for the high-density dataset.

After removing redundant data points, we retrain the model depicted in Figure 5 and Figure 6, showing the retrained model compared to the previous model for the high-density model and the low-density model, respectively. The figures show the model predictions before and after removing the redundancy.

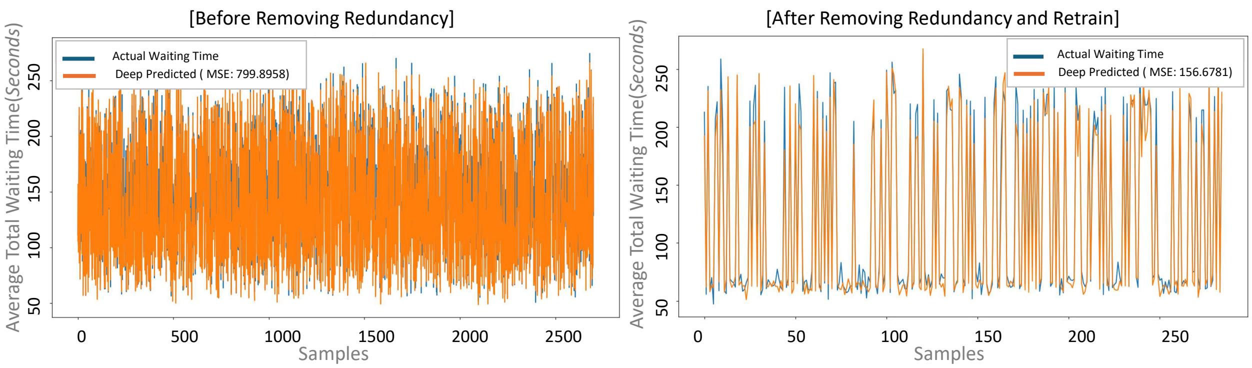

Figure 5: Removing redundancy and retraining high-density model (in the right figure, about eight times fewer samples are used in the training).

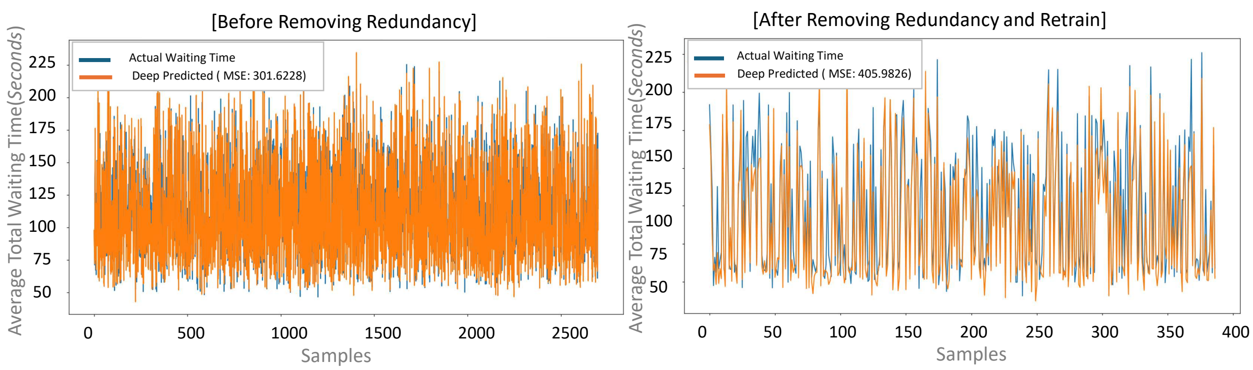

Figure 6: Removing redundancy and retraining low-density model (in the right figure, about seven times fewer samples are used in the training).

4.7 Predictive Accuracy for Interpolation and Extrapolation

We perform a complete evaluation to validate the model's predictive capabilities. We check the interpolation (predictions within the range of training data), results shown in Figure 7 for the high-density model and in Figure 8 for the low-density model. These results demonstrate that–even after partitioning the dataset and removing redundant data–the model keeps making accurate predictions. However, when we check extrapolation (predictions outside the range of training data), we do not get correct results. This outcome aligns with what we focus on: that predictions outside the experimental frame of the model are unreliable. Extrapolation is scenarios with --*insertion-density* argument = 50 which creates a very dense traffic situation and falls outside the validity region of the high-fidelity model.

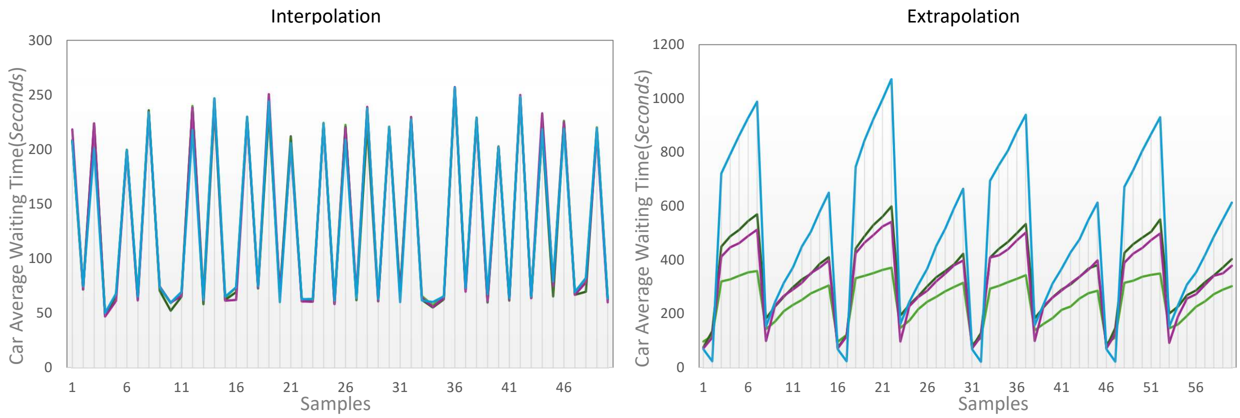

Figure 7: Comparison of interpolation and extrapolation predictions for the high-density model.

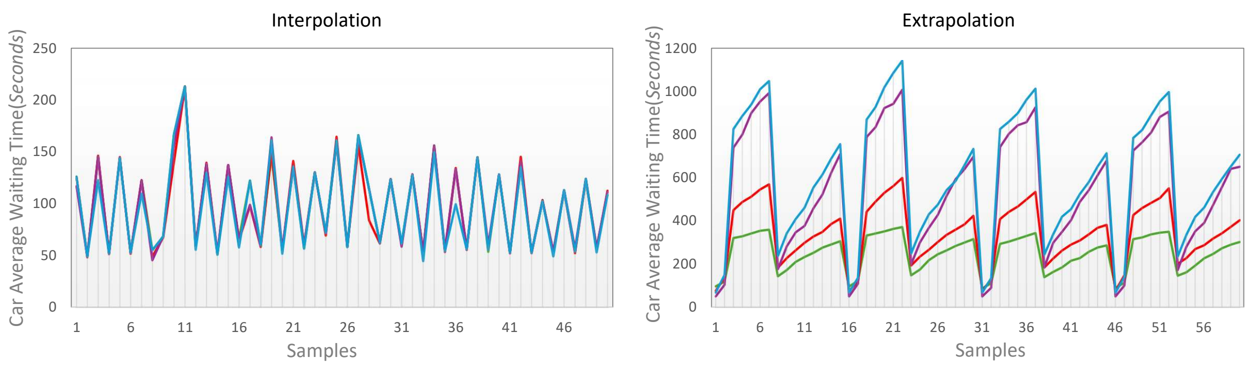

Figure 8: Comparison of interpolation and extrapolation for the low-density model.

5 DISCUSSION AND LIMITATIONS

Choosing Between Linear Regression and Deep Neural Networks: When designing a predictive model, the choice between using Linear Regression (LR) or Deep Neural Network (DNN) depends on the complexity and characteristics of the problem. In this challenge, we discuss selecting the appropriate approach based on the structure of the dataset and the predictive requirements of the traffic system model.

For this experiment, we developed both an LR model and a DNN model to evaluate their performance. Figure 9 illustrates the results of each model. The data structure and requirements align more closely with DNN's capabilities, making it the most appropriate option for this experiment. Since the training duration of the DNN model is longer than that of the LR model, we run the training on an NVIDIA GeForce MX450 GPU instead of a CPU using CUDA which is a parallel computing platform and programming model developed by NVIDIA.

Balanced Data: The distribution of randomly generated datasets often depends on the specific process used for data generation, which can lead to imbalanced data and consequently limit the model's abilities to achieve optimal generalization. Therefore, we changed the dataset by generating different traffic light configurations to keep the duration phase of 90 seconds for each traffic light and the yellow phase between 3 and 6 seconds according to ADOT Traffic Engineering Guidelines and Processes, section 600 [20].

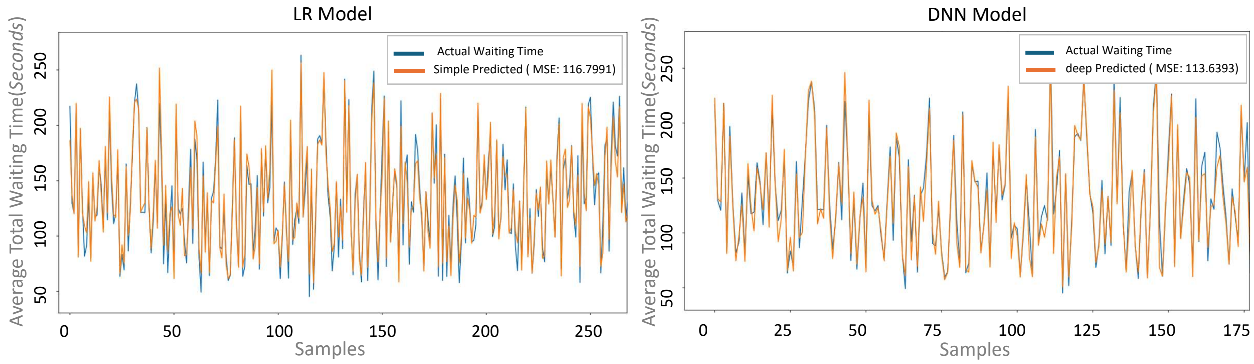

Figure 9: Zoomed-in view of linear regression and deep neural network models.

Choice of Metric: In this usecase, we chose Euclidean distance as the metric space. Euclidean distance measures the straight-line distance between two points in space. However, this metric is not universally suitable for all use cases. Its effectiveness depends on the domain and data characteristics. For example, in high-dimensional spaces or non-linear data, alternative metrics (e.g., cosine similarity or Manhattan distance) might be more appropriate. The choice of metric should align with the specific requirements and structure of the data in the application.

Computational Considerations in Pairwise Distance Calculation: Calculating pairwise distances between every data point in the dataset can be computationally expensive and may pose scalability challenges. To mitigate this issue, an alternative approach is to randomly sample data points and compute their distances rather than performing a full pairwise comparison. This strategy ensures that redundancy and overlap in the dataset can still be effectively identified while reducing computational overhead.

Boundary and overlapping Regions: The focus is on minimizing the overlap between experimental frames as much as possible; however, some overlap remains. Moreover, there are some boundaries. These areas introduce ambiguity in model selection and highlight the need for further investigation to improve decision-making in these regions.

6 CONCLUSION AND FUTURE WORK

This paper presented an approach to surrogate modeling for multi-modal systems focusing on traffic simulations. We defined the concept of the validity frame as an experimental frame plus a process that allows us to change the experimental frame. In this paper, we introduced a method to define this validity frame by partitioning the system operating domain into distinct experimental frames and training separate surrogate models on minimally overlapping datasets, we enable more efficient simulation while maintaining predictive accuracy within each model's intended scope.

In this research, we showed how to construct surrogate models for low and high-density traffic scenarios using data generated in SUMO, and we demonstrated how balancing the dataset and reducing redundancy improves model performance. Our approach ensures that the union of the experimental frames forms complete and non-redundant experimental frames.

While this method improves scalability and clarity in surrogate modeling, overlap at the boundaries between modes poses the challenge of uncertainty in choosing the appropriate model. Future work will focus on refining model-switching mechanisms, exploring adaptive and runtime approaches for validity frame management, and expanding this method to more complex or real-time traffic systems.

ACKNOWLEDGMENTS

Raheleh Biglari is funded by the Bijzonder Onderzoeksfonds (BOF) fund at the University of Antwerp.

Claudio Gomes' work is partially supported by the RoboSAPIENS project funded by the European Commission's Horizon Europe programme under grant agreement number 101133807.

REFERENCES

[1] C. Angione, E. Silverman, and E. Yaneske, "Using machine learning as a surrogate model for agent-based simulations," *PLOS ONE*, vol. 17, no. 2, p. e0263150, Feb. 2022.

[2] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, "Microscopic traffic simulation using sumo," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2575–2582.

[3] L. Wright and S. Davidson, "How to tell the difference between a model and a digital twin," *Advanced Modeling and Simulation in Engineering Sciences*, vol. 7, no. 1, pp. 1–13, 2020.

[4] L. Jia, R. Alizadeh, J. Hao, G. Wang, J. K. Allen, and F. Mistree, "A rule-based method for automated surrogate model selection," *Advanced Engineering Informatics*, vol. 45, p. 101123, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034620300926

[5] B. P. Zeigler, *Multifacetted modelling and discrete event simulation*. Academic Press Professional, Inc., 1984.

[6] B. Van Acker, P. De Meulenaere, H. Vangheluwe, and J. Denil, "Validity frame–enabled model-based engineering processes," *Simulation*, vol. 100, no. 2, pp. 185–226, 2024.

[7] S. Van Mierlo, B. J. Oakes, B. Van Acker, R. Eslampanah, J. Denil, and H. Vangheluwe, "Exploring validity frames in practice," in *International Conference on Systems Modelling and Management*. Springer, 2020, pp. 131–148.

[8] R. G. Sargent, "An introductory tutorial on verification and validation of simulation models," in *2015 winter simulation conference (WSC)*. IEEE, 2015, pp. 1729–1740.

[9] F. Lauer and G. Bloch, *Hybrid System Identification: Theory and Algorithms for Learning Switching Models*. Springer International Publishing, 2019.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[11] C. Legaard, T. Schranz, G. Schweiger, J. Drgoňa, B. Falay, C. Gomes, A. Iosifidis, M. Abkar, and P. Larsen, "Constructing neural network based models for simulating dynamical systems," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–34, Feb. 2023.

[12] B. Sanguino, "Parameter estimation and analysis of wind impact on ship maneuvering model," Master's thesis, NTNU, 2024.

[13] M. Rätz, P. Henkel, P. Stoffel, R. Streblow, and D. Müller, "Identifying the validity domain of machine learning models in building energy systems," *Energy and AI*, vol. 15, p. 100324, Jan. 2024.

[14] R. Biglari and J. Denil, "Model validity and tolerance quantification for real-time adaptive approximation," in *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, ser. MODELS '22. ACM, Oct. 2022, pp. 668–676.

[15] S. Bosmans, S. Mercelis, P. Hellinckx, and J. Denil, "Reducing computational cost of large-scale simulations using opportunistic model approximation," in *2019 Spring Simulation Conference (SpringSim)*. IEEE, Apr. 2019, pp. 1–12.

[16] R. Franceschini, S. V. Mierlo, and H. Vangheluwe, "Towards adaptive abstraction in agent based simulation," in *2019 Winter Simulation Conference (WSC)*. IEEE, Dec. 2019, pp. 2725–2736.

[17] S. Bosmans, T. Bogaerts, W. Casteels, S. Mercelis, J. Denil, and P. Hellinckx, "Adaptivity in multi-level traffic simulation using experimental frames," *Simulation Modelling Practice and Theory*, vol. 114, p. 102395, Jan. 2022.

[18] H. Ejersbo, K. Lausdahl, M. Frasheri, and L. Esterle, "Dynamic runtime integration of new models in digital twins," in *2023 IEEE/ACM 18th Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*. IEEE, May 2023, pp. 44–55.

[19] OpenStreetMap contributors, "Openstreetmap," 2025, accessed: 2025-03-27. [Online]. Available: https://www.openstreetmap.org/

[20] Arizona Department of Transportation, "Traffic group photolog summary report: Tgp0621," https://azdot.gov/sites/default/files/2019/05/tgp0621-2018-01.pdf, 2018, accessed: 2025-04-10.

AUTHOR BIOGRAPHIES

RAHELEH BIGLARI is a doctoral student at the University of Antwerp, Faculty of Applied Engineering in the Department of Electronics and Information and Communication Technologies, Cosys-Lab. Her research interests lie in performance modeling, model-based system engineering, and digital twin engineering. Her email address is raheleh.biglari@uantwerp.be.

CLAUDIO GOMES is an assistant professor at the Department of Electrical and Computer Engineering at Aarhus University, Denmark. His research interests include co-simulation, digital twin engineering, and machine learning for digital twins. His email address is claudio.gomes@ece.au.dk.

JOACHIM DENIL is an associate professor at the University of Antwerp, Faculty of Applied Engineering in the Department of Electronics and Information and Communication Technologies, Cosys-Lab, and is associated with Flanders Make. His research interests are enabling methods, techniques, and tools to design, verify, and evolve Cyber-Physical Systems. Specifically, he is interested in the performance modeling and simulation, model-based systems engineering, and verification and validation of models and systems. His email address is joachim.denil@uantwerp.be.