Collective Dynamics

The Evaluation of Data Fitting Approaches for Speed/Flow Density Relationships

Arthur Rohaert[1] · **Jonathan Wahlqvist**[1] · **Hana Najmanová**[2] · **Nikolai Bode**[3] · **Enrico Ronchi**[1]

[1] Division of Fire Safety Engineering, Lund University, Sweden
E-mail: enrico.ronchi@brand.lth.se
[2] Czech Technical University in Prague, Czech Republic
[3] University of Bristol, United Kingdom

Received: 9 Octobre 2023 / Last revision received: 7 June 2024 / Accepted: 10 June 2024
DOI: 10.17815/CD.2024.177

Abstract This paper presents guidance on data-fitting approaches in the context of pedestrian and evacuation dynamics research. In particular, it examines parametric and non-parametric regression techniques for analysing speed/flow density relationships. Parametric models assume predefined functional forms, while non-parametric models provide flexibility to capture complex relationships. This paper evaluates a range of traditional statistical approaches and machine-learning techniques. It emphasises the importance of weighting unbalanced datasets to enhance model accuracy. Practical applications are illustrated using traffic and pedestrian evacuation data.

This paper is intended to stimulate discussion on best practices for developing, calibrating, and testing macroscopic and microscopic evacuation models. It does not prescribe a one-size-fits-all solution for evacuation data fitting approaches, but it provides an overview of existing methods and analyses their advantages and limitations.

Keywords data fitting · speed/flow density relationship · pedestrian dynamics · traffic dynamics

1 Introduction

How the speed/flow of pedestrians and traffic depends on the density plays an important role in transportation and safety research, as this relationship characterises transport systems, and informs the design and management of transport facilities. For example, the flow's maximum value indicates a system's capacity, a concept used at spatial scales ranging from exit doors to cities [1]. Increasing traffic density beyond this point will render the system less efficient [2]. These relationships also represent a common tool for various engineering applications, including capacity analysis of evacuation routes or pedestrian flow prediction in various egress components. Moreover, macroscopic evacuation models rely on the relationships between speed/flow and density to simulate evacuation outcomes. In microscopic evacuation models, the relationships emerge from the agent-based behavioural rules [3–6]. Because of the importance of these relationships, many studies have reported empirical speed, flow, and density data and attempted to describe the trends in terms of mathematical functions that facilitate prediction, and comparison to other contexts. This process requires the fitting of mathematical functions to the data, which is the subject of this paper.

This paper explains how to fit both parametric regressions and non-parametric regressions to data. Practical applications are illustrated using pedestrian and traffic evacuation data collected during real-world scenarios. We have published all the data in an open repository [7–9], as well as the scripts to establish the regressions [10].

2 Parametric regression

Parametric regression involves modelling the relationship between the dependent and independent variables using a predetermined equation with a fixed number of parameters. This means that the shape and form of the regression curve are determined by the chosen speed/flow density relationship (model) and the estimated parameter values. Well-known examples include the parabolic model (Greenshields [11] in traffic dynamics, equivalent to Older [12], Fruin [13], and many others in pedestrian dynamics), the bi-linear model of Daganzo [14] (traffic dynamics) and the exponential model (first discussed in traffic dynamics by Newell [15] and Franklin [16] but also know as the model of del Castillo and Benítez [17] and equivalent to Weidmann [18] in pedestrian dynamics). These macroscopic models can be fitted to data by minimising the sum of squares of the speed prediction errors.

The advantage of these parametric models is that they can incorporate prior knowledge that suggests the trend of the data. Although most of the early models (such as Daganzo's model [14]) are based on empirical data, they were later derived from rules used in agent-based models. In addition, many models use parameters with physical meaning, making it easier to measure and interpret these parameters in real-world applications.

Empirical speed-density data are typically unbalanced, simply because higher pedestrian and vehicle densities are less common than lower densities. This leads to a poor fit for high densities. To counteract this effect, models can be fitted by weighting the sum of least squares, prioritizing points with fewer or more distant neighbours, as suggested by Qu et al. [19]. We demonstrated the need to weight the residuals by fitting the model of Underwood [20] to data that was collected during the 2019 Kincade wildfire evacuation (see Fig. 1) [8, 21]. Note that higher traffic density conditions are rare in this dataset: 98.6 % of the 69.116 data points have a density below 20 veh/km/lane. The importance of weighting can be demonstrated analogously with pedestrian data.

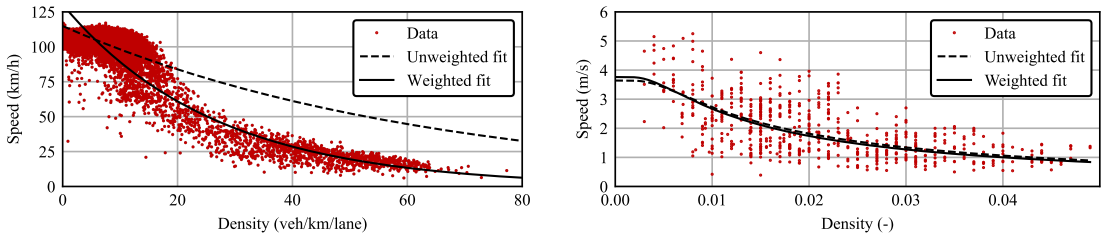

Figure 1 Left: The speed-density data on US 101 during the 2019 Kincade Fire, USA [22] and the fitted exponential model of Underwood [20], with and without applying the weights proposed by Qu et al. [19]. Note that the simple two-parameter model of Underwood does not accurately represent real traffic dynamics for low traffic densities. Right: The speed-density data from evacuation exercises with preschool children [23]. Note that the dimensionless density is determined using the methodology established by Predtechinski and Milinski [5]. The fitted model is the three-parameter model of Weidmann [18]. The need to weight data decreases when the dataset is better balanced, as seen in this evacuation experiment, and when the model more accurately captures the trend.

3 Non-parametric regression

As shown in Fig. 1, predefined speed-density relationship may represent the true trend inaccurately, leading to biased predictions. Non-parametric regression, on the other hand, offers more flexibility. It allows us to compare different datasets and can provide better estimates of certain key parameters, such as the free-flow speed or the capacity. The values of these parameters can be difficult to obtain from parametric regression. For example, the free-flow speed (speed values at 0 density) in Fig. 1 clearly depends on the ability of the model to fit the data well in the low-density region.

3.1 Traditional non-parametric regression

The trend of a dataset can be represented by a moving average. In its simplest form, the number of neighbouring data points is considered fixed. However, the average value can also be obtained by looking at all the neighbours within a fixed window, often referred to as kernel smoothing. In Fig. 2, a Gaussian kernel smoothing has been applied.

The same smoothing technique can be used to calculate empirical percentiles and visualise the dataset's variation, including heteroscedasticity and asymmetry.

3.2 Machine learning regression

Machine learning algorithms have become increasingly important in recent years. In this section, we will explore some popular non-parametric regressions: Kernel Ridge Regression (KRR), Support Vector Regression (SVR), and Gaussian Process Regression (GPR). KRR and SVR offer a deterministic single predicted speed for a given density, while GPR, a probabilistic regression, provides a speed probability distribution. Each of the mentioned algorithms employs the kernel trick, which transforms input features into a higher-

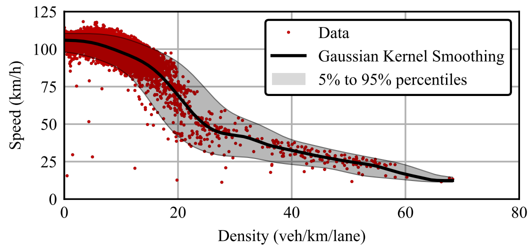

Figure 2 The speed-density data on US 101 during the 2020 Glass Fire, USA [21]. The weighted moving average and the empirical percentiles employ a Gaussian Kernel Smoothing. Here, the percentiles are calculated by linear interpolation of the cumulative distribution function [29]. The length scale of the kernel is optimised by employing cross-validation.

dimensional space to capture complex relationships. The radial basis function effectively considers an infinite series of polynomials (Mercer's theorem), which contributes to its flexibility in modelling complex functions [24]. We will also discuss hyper-parameter optimization, which is an important process required to improve the performance of the algorithms.

Kernel Ridge Regression (KRR) performs a ridge regression on the data after it has been transformed into a higher dimensional space (after the kernel trick has been applied). Ridge regression then minimises the sum of squared errors in the transformed space. The two key hyperparameters are alpha (the regularisation) and gamma (the range of influence of the kernel) [25].

Support Vector Regression (SVR) optimises the position of a hyperplane by minimising the prediction error while ensuring a margin constraint. SVR is robust to outliers because it only depends on support vectors (data points closest to the hyperplane). The hyperparameters are the regularisation parameter C and the tolerance margin epsilon [26,27].

Gaussian Process Regression (GPR) probabilistically models the speed-density relationship as a Gaussian process, estimating both a mean function and a covariance function for the expected speed. It captures the distribution of potential functions fitting the training data by utilizing a prior function distribution and kernel function to shape expectations. The posterior distribution over functions is then calculated via Bayes' rule to make speed predictions [28].

3.3 Hyperparameter optimization

In the context of regression, an over-fitted model can result in a very low training error (i.e., the model predicts the training data very accurately), but a high test error (i.e., the model performs poorly on new data [30]). On the other hand, an under-fitted model can result in a high training error and a high test error. One approach to finding an adequate fit (see Fig. 3) is to optimise the hyperparameters by means of cross-validation [31].

Cross-validation is a widely used technique, used to evaluate the performance of a model and to select the best hyperparameters. The idea is to partition the data into multiple

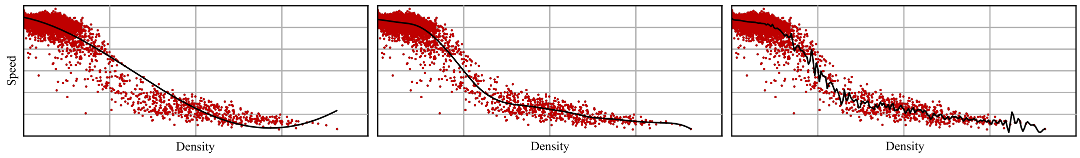

Figure 3 An example of an under-fitted (left), well-fitted (centre) and over-fitted (right) kernel ridge regressions. The under-fitted regression fails to capture the trend of the data accurately, while the over-fitted model has incorporated noise of the dataset in the regression.

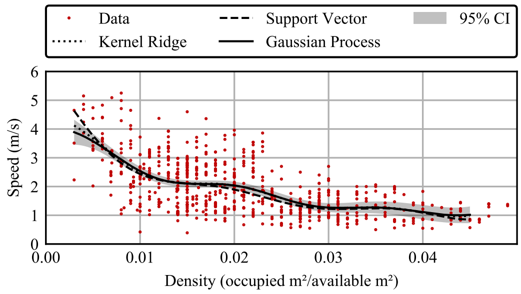

Figure 4 The speed-density data from evacuation exercises with preschool children [23] and three popular machine-learning regression techniques (Kernel Ridge, Support Vector, and Gaussian Process). All methods deliver similar regressions. Gaussian Process Regression is computationally the most expensive, but also provides confidence intervals.

subsets, or folds, and to train the model on some sets and test it on the remaining sets. This process is repeated for all combinations of training and testing sets, and the performance metrics are averaged across all folds [31]. Here, a five-fold cross-validation is performed. The model is trained on 4 folds and tested on the remaining fold. This process is repeated 5 times, with each fold serving as the testing set once.

An example is given in Fig. 4, in which the three earlier mentioned algorithms have been applied to an evacuation exercise with preschool children [23]. The optical parameters were the result of a grid search (an exhaustive technique that tries every combination of hyperparameters specified), during which the performance was measured by a five-fold cross-validation.

4 Conclusion

This paper serves as a guide for researchers in the field of pedestrian and evacuation dynamics, presenting practical approaches for developing, calibrating, and testing both macroscopic and microscopic models. Rather than proposing a universal solution, it outlines key methods and provides practical guidelines for their application to different datasets.

What follows is a list of practical recommendations

- **Consider both parametric and non-parametric regression**: parametric models offer interpretability and can represent physical process that occur in the flow, while non-parametric models offer flexibility and are great to compare data without the biased or limitation of a parametric model.

- **Address unbalanced datasets**: when working with unbalanced datasets, such as those commonly found in pedestrian and traffic dynamics research, weighting techniques should be applied to improve model accuracy.
- **Study outliers and artefacts**: attempt to understand them (e.g., measurement equipment malfunction, unique person, etc.)
- **When choosing a parametric model**, prefer models with few parameters that fit well to the data.
- **Optimize Hyperparameters**: this is crucial for improving the performance of machine learning regression models. Employ techniques like cross-validation to evaluate model performance and select the best hyperparameters for your dataset.
- **Question the outcome** of machine learning algorithms and traditional fits. Check if the regressions have a physical meaning (e.g., monotonic functions).

Acknowledgements Arthur Rohaert and Enrico Ronchi would like to express their gratitude to the National Research Council of Canada for the funding provided through the ”RACE Wildfires: Resilience and Adaptation to Climatic Extreme Wildfires” project, grant agreement # ST-R2-0202, which supported this research.

Author Contributions Arthur Rohaert: Conceptualization, Methodology, Software, Formal Analysis, Resources, Writing / Jonathan Wahlqvist: Supervision / Hana Najmanová: Conceptualization, Resources, Reviewing / Nikolai Bode: Conceptualization, Reviewing / Enrico Ronchi: Conceptualization, Reviewing, Supervision.

References

[1] Geroliminis, N., Sun, J.: Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transportation Research Part B: Methodological **45**, 605–617 (2011). doi:10.1016/j.trb.2010.11.004

[2] Cepolina, E.M.: Phased evacuation: An optimisation model which takes into account the capacity drop phenomenon in pedestrian flows. Fire Safety Journal **44**, 532–544 (2009). doi:10.1016/j.firesaf.2008.11.002

[3] Lighthill, M.J., Whitham, G.B.: On kinematic waves II. a theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **229**, 317–345 (1955). doi:10.1098/rspa.1955.0089

[4] Richards, P.I.: Shock waves on the highway. Operations Research **4**, 42–51 (1956). doi:10.1287/opre.4.1.42

[5] Predtechenskii, V.M., Milinskiĭ, A.I.: Planning for foot traffic flow in buildings. The National Bureau of Standards, U.S. Dept. of Commerce, and the National Science Foundation, Washington, DC (1978)

[6] Gwynne, S.M.V., Rosenbaum, E.R.: Employing the hydraulic model in assessing emergency movement. In: SFPE Handbook of Fire Protection Engineering, pp. 2115–2151. Springer New York (2016). doi:10.1007/978-1-4939-2565-0_59

[7] Najmanová, H., Ronchi, E.: Data for: Experimental data about the evacuation of preschool children from nursery schools, Part II: Movement characteristics and behaviour (2023). doi:10.5281/ZENODO.7377045

[8] Rohaert, A., Kuligowski, E.D., Ardinge, A., Wahlqvist, J., Gwynne, S.M., Kimball, A., Noureddine Bénichou, Ronchi, E.: Dataset of traffic dynamics during the 2019 Kincade Wildfire Evacuation (2022). doi:10.5281/ZENODO.7410114

[9] Rohaert, A., Araghi, N.J., Kuligowski, E.D., Ronchi, E.: Dataset of traffic dynamics during the 2020 Glass Wildfire Evacuation (2023). doi:10.5281/ZENODO.7483487

[10] Rohaert, A.: Data fitting script for speed/flow density relationships (2024). doi:10.5281/zenodo.10990917

[11] Greenshields, B., Bibbins, J., Channing, W., Miller, H.: A study of traffic capacity. Highway Research Board proceedings (1935)

[12] Older, S.J.: Movement of pedestrians on footways in shopping streets. Traffic Engineering & Control **10**, 160 (1968)

[13] Fruin, J.J.: Designing for pedestrians a level of service concept. Polytechnic University (1970)

[14] Daganzo, C.F.: The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B: Methodological **28**, 269–287 (1994). doi:10.1016/0191-2615(94)90002-7

[15] Newell, G.F.: Nonlinear Effects in the Dynamics of Car Following. Operations Research **9**(2), 209–229 (1961). doi:10.1287/opre.9.2.209

[16] Franklin, R.: The structure of a traffic shock wave. Civil Engineering Pulb. Wks. Rev **56**, 1186–1188 (1961)

[17] Del Castillo, J.M., Benítez, F.G.: On the functional form of the speed-density relationship - Part I: General theory. Transportation Research Part B: Methodological **29**, 373–389 (1995). doi:10.1016/0191-2615(95)00008-2

[18] Weidmann, U.: Transporttechnik der Fussgänger: Transporttechnische Eigenschaften des Fussgängerverkehrs. Tech. rep., ETH Zurich (1993). doi:10.3929/ETHZ-B-000242008

[19] Qu, X., Wang, S., Zhang, J.: On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models. Transportation Research Part B: Methodological **73**, 91–102 (2015). doi:10.1016/j.trb.2015.01.001

[20] Underwood, R.T.: Speed, volume and density relationships. Quality and Theory of Traffic Flow. A Symposium, Yale University Bureau of Highway Traffic pp. 141–188 (1961). Publisher: Bureau of Highway Traffic, Yale University

[21] Rohaert, A., Janfeshanaraghi, N., Kuligowski, E., Ronchi, E.: The analysis of traffic data of wildfire evacuation: the case study of the 2020 Glass Fire. Fire Safety Journal **141** (2023). doi:10.1016/j.firesaf.2023.103909

[22] Rohaert, A., Kuligowski, E.D., Ardinge, A., Wahlqvist, J., Gwynne, S.M., Kimball, A., Bénichou, N., Ronchi, E.: Traffic dynamics during the 2019 Kincade wildfire evacuation. Transportation Research Part D: Transport and Environment **116** (2023). doi:10.1016/j.trd.2023.103610

[23] Najmanová, H., Ronchi, E.: Experimental data about the evacuation of preschool children from nursery schools, Part II: Movement characteristics and behaviour. Fire Safety Journal **139** (2023). doi:10.1016/j.firesaf.2023.103797

[24] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[25] Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **12**, 55–67 (1970). doi:10.1080/00401706.1970.10488634

[26] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995). doi:10.1007/BF00994018

[27] Muller, P., Parmigiani, G.: Optimal Design via Curve Fitting of Monte Carlo Experiments. Journal of the American Statistical Association **90**, 1322–1322 (1995). doi:10.2307/2291522

[28] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2005). doi:10.7551/mitpress/3206.001.0001

[29] Parzen, E.: Nonparametric Statistical Data Modeling. Journal of the American Statistical Association **74**, 105–121 (1979). doi:10.1080/01621459.1979.10481621

[30] Bode, N.W., Ronchi, E.: Statistical Model Fitting and Model Selection in Pedestrian Dynamics Research. Collective Dynamics **4** (2019). doi:10.17815/CD.2019.20

[31] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY (2013)