Treating Noise and Anomalies in Vehicle Trajectories from an Experiment with a Swarm of Drones

Vishal Mahajan, Emmanouil Barmpounakis, Md Rakibul Alam, Nikolas Geroliminis and Constantinos Antoniou

***Abstract*—Unmanned aerial systems, known as "drones," are relatively new in collecting traffic data. Data from drone videography can have potential applications for traffic research. Drones can record the vehicles from their aerial point-of-view and provide their naturalistic driving behavior. Processing raw data from drones to remove noise and anomalies is crucial to ensure that the data are fit for subsequent applications, e.g., the development of traffic flow or crash risk models. This study uses a part of the pNEUMA dataset, a large dataset with almost half a million trajectories captured by a swarm of drones over Athens, Greece. This novel dataset offers an opportunity to analyze the data attributes and treat the noise and outliers in the data. We use a combination of smoothing filters and Extreme Gradient Boosting with adaptive regularization to process the speed and acceleration profiles of the vehicle trajectories in the dataset. Our approach can help prospective data users treat this or similar trajectory datasets alternatively to applying manual thresholds and assist in accelerating research in microscopic traffic analysis.**

***Index Terms*—drone data, trajectory data, anomaly detection, machine learning**

I. Introduction

RESEARCHERS and practitioners need real-world traffic data to study traffic behavior and implement traffic management strategies. Traffic data collection is classified into point, edge, and area-wide measurements [1]. Traditional traffic data collection, such as count data collection from loop detectors, is a form of point measurement. Point-to-point measurement using Floating Car Data (FCD) and Bluetooth scanners have also gained prominence in the last few years. These data collection methods have limited observability, i.e., for a few selected road segments or from the selected fleet of vehicles. In recent years, advances such as fast microprocessors, efficient storage, and wireless communication technologies have allowed the use of drones for many civil applications [2]. Aerial footage from Unmanned Aerial Systems (UAS) or more commonly known as "drones", is one of the newest methods for collecting traffic data and has notable advantages such as observation of naturalistic driving behavior and detailed driving trajectory [3], [4].

In 2018, a large-scale drone data collection (pNEUMA) was conducted in Athens, Greece, which covered a large extent in the context of urban driving [5]. The pNEUMA dataset collected by [5] is part of an open science initiative that researchers from different disciplines can use to develop and test their models. The authors describe the challenges of collecting video data for an extended area. An essential set pertains before and during the drones' flight, which must be planned and accurately determined because of weather, battery backup, video quality, regulatory approvals, and technical expertise. However, another set of essential challenges pertains to post-processing the video recording after the drone flight. The researchers use state-of-the-art computer vision algorithms to detect and track vehicles and extract vehicle trajectories from the raw videos. Since the errors in the vehicle's position are in the order of 20 cm or less [5], trajectory extraction follows quite accurately the vehicle's position. Nevertheless, vehicles are not point objects but cover significant space; minor errors in the position can accumulate when speed or acceleration variables are calculated. Furthermore, the urban driving environment is more complex than highways due to the increased heterogeneity of vehicle classes, traffic signals, congestion, intersections, parking vehicles, and occlusion due to high-rise buildings and trees [6]. All these factors could introduce noise and anomalies or outliers in the extracted trajectories from the

Article DOI: 10.1109/TITS.2023.3268712 This research was partially funded by Swiss National Science Foundation (Grant no. 200021_188590) "pNEUMA: On the new era of urban traffic models with massive empirical data from aerial footage", the German Research Foundation DFG (TRAMPA Project, Grant 415208373) and the DAAD Project number 57474280 "Verkehr-SuTra: Technologies for Sustainable Transportation", within the Programme "A New Passage to India - Deutsch-Indische Hochschulkooperationen ab 2019" (Corresponding author: vishal.mahajan@tum.de). Original data source: pNEUMA https://open-traffic.epfl.ch/. The codes and processed data from this study can be accessed at https://github.com/vishalmhjn/pneuma_treatment

Vishal Mahajan, Md. Rakibul Alam and Constantinos Antoniou are with Chair of Transportation Systems Engineering, TUM School of Engineering and Design, Technical University of Munich, Germany.

Emmanouil Barmpounakis and Nikolas Geroliminis are with the Urban Transport Systems Laboratory (LUTS) of the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

drone videos, and thus, the data may require additional treatment. There is an opportunity to address this issue by analyzing the trajectories and treating the noise and anomalies. These data contain many naturalistic trajectories, and thus filtering the anomalies and smoothing the noise could accelerate subsequent research attempts. For researchers to fully take advantage of such a detailed and large dataset, it is first necessary to find appropriate techniques to detect these cases and filter them efficiently.

We identify the problematic cases based on the acceleration from the original dataset in this paper. First, we treat the noise and anomalies in the vehicles' speed and acceleration time series obtained from vehicles' trajectories using noise filters and an anomaly detection model. Then, a sensitivity analysis is conducted to choose the best parameter setting to classify representative/ usual and anomalous data. Our study's main contribution is applying existing noise filters and machine learning based anomaly detection model to data collected from drones by using the challenging pNEUMA dataset. In this pursuit, we use Extreme Gradient Boosting (XGBoost) with adaptive regularization to create an anomaly mask for each trajectory. The remainder of the paper is structured as follows: Section 2 concisely reviews the literature on this topic, Section 3 presents the methodology of the study, Section 4 provides the data description and illustrates some of the problematic cases, followed by Section 5 on data analysis, Section 6 shows the results of this research, and lastly, we conclude the study and present its limitations and implications.

II. Literature Review

Sensor measurements often come with errors. Noise and anomalies (also referred to as outliers) are two of the common types of errors [7]. These errors deviate the measured signal from its desirable value, and therefore the data require processing before use. The desired value is the best possible representation of the *true underlying signal* that can be measured [8]. The desired value may differ from the absolute true value of a signal that may or may not be possible to measure. Processing time-series or sequence data is classified into three main tasks: filtering, smoothing, and prediction. Kalman [9] formalized the distinction between filtering and smoothing. Suppose, the observed time sequence $y(t_0), \ldots, y(t_n)$ of length $n+1$ from which we need to estimate the unobservable or desired value of the true signal at $t = t_i$, where t_i is the time of interest. If $t_i < t_n$, it is data smoothing, whereas if $t_i = t_n$, it is data filtering and if $t_i > t_n$, then it is a prediction task [9]. In the following paragraphs, we briefly introduce the topics of noise and anomaly, followed by a discussion of a few studies on vehicle trajectory datasets from the traffic research domain.

A. Noise

Noise is the unwanted component of the signal which is not relevant to a specific task. Removal or treatment of noise is a general prerequisite for data usage. The source of noise in the data can be the measuring device or sensor and its surroundings during the data collection, e.g., in drone videography, noise can be introduced due to the vibrations of the camera apparatus. This could be characterized by the presence of a periodic high-frequency signal superimposed on the desired value. Data processing methods can also introduce noise in the data, e.g., extraction of trajectories from even a stabilized video could be noisy depending on the algorithms and tools used. For time-series data, smoothing refers to a broad array of methods to remove the noise from the data. This is commonly done by allowing only the low frequency of the signal to pass while attenuating the high-frequency component [10]. The moving average filter is one of the most common low-pass filters, where the current estimate is the rolling average of neighboring values.

Savitzky-Golay (SG) filter is another example of a low-pass filter. SG filter (SGF) is an efficient method of data smoothing using local least-square polynomial approximation [11]. Polynomial fitting on a sub-sequence of length 2M+1 (where M is a positive integer) and then evaluation of the polynomial's output at the central point is equivalent to the convolution of the sub-sequence with a fixed set of integers (impulse response) [12]. The output samples (y) obtained from the discrete convolution of fixed weights (h) with the input sample (x) is shown in (1) [11]:

$$y[n] = \sum_{m=-M}^{M} h[m] \cdot x[n-m] = \sum_{m=n-M}^{n+M} h[n-m] \cdot x[m] \tag{1}$$

For a uniformly spaced sequence, the weights are computed only once based on the length of the sub-sequence (window size 2M+1) and the polynomial degree. This is beneficial for the sensor data because the sensor data are generally generated at a fixed frequency and equally spaced. Common weights for every convolution operation on the sub-sequence make it highly efficient in speed and memory. The output of the SG filter is suitable for subsequent application of the anomaly detection algorithm due to the high signal-noise ratio. For a polynomial of degree 0 or 1, SG filter is equivalent to a moving average filter [12].

Gaussian Filter (GF) is another popular filter, especially in image processing. As the name suggests, the input sample is convoluted with a Gaussian kernel (2) to get the smoothed estimates:

$$N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{2}$$

Wherein σ is the standard deviation. The kernel is truncated symmetrically beyond a specified number of σ. The

output of the GF is weighted more towards the central values of the input samples due to the characteristic shape of the Gaussian kernel. This makes the GF a gentler smoothing filter than a moving average filter [13]. We want to point out that the methodical steps in noise filters, such as numerical approximation or truncation, could introduce or exacerbate errors in processed data [14]. It is a matter of trade-off between introducing these errors and removing the data's noise. Filters are justified if their output is closer to the desired value than the input data.

B. Anomalies

The term "anomaly" refers to a behavior different from the usual or representative behavior of the system [15]. When the primary objective is to recover the representative signal, anomalies are detected, removed, or replaced. In contrast to the noise, anomalies are not always unwanted. Anomaly detection in time-series data is an active research topic in various fields, such as finance, network security, and health [15]. Time-series data can consist of either a single-point anomaly or an anomalous sub-sequence. In high-frequency sensor data, sequence anomalies can be prominent as a single disturbance in the signal can span over multiple points. For instance, unrealistic high transient values or peaks can characterize anomalies in the data. Chandola *et al.* [15] categorized the anomaly detection techniques under classification, clustering, nearest-neighbor, statistical, information-theoretic, and spectral-based methods. Anomaly detection can be seen as a supervised learning task, but this is practically constrained due to the often unavailability of ground-truth labels. This is why unsupervised techniques hold significant potential for anomaly detection. These methods aim to find the best separation between the usual and anomalous data points/ sequences based on the specified parameters (distance, density, and probability). For instance, Eskin [16] used a machine learning model to learn the probability distribution over the data and then applied a statistical test to detect the anomalies.

C. Noise and anomalies in vehicle trajectory data

It is essential to clarify what is considered usual or representative behavior of the system viz-a-viz "abnormal" or "unusual" within the scope of our study. This study deals with the naturalistic trajectory data, which is of great interest to the researchers as it provides in-situ driving behavior. In a recent study [17], the authors review the trajectory smoothing/ filtering techniques to process trajectory data from diverse data sources such as ground-based camera videos, drone videos, and instrumented vehicles. Since our study focuses on anomaly detection, we list the range of acceleration and deceleration noted in some previous studies (Fig. 1). We point out that the drone videography data are still in the early phases, and thus there are limited studies on processing such data. We consider a few prominent studies from other data sources dealing with vehicle trajectories to compensate for this. We also identify the context and data sources in these studies. The values are from different contexts (driving environment, vehicle types, data collection, and processing methods) and have varying ranges. Therefore, the context of a study should be considered before the acceptability of the acceleration range.

Bokare and Maurya [18] analyzed the acceleration and deceleration behavior of different vehicles using Global Positioning System (GPS) data. They noted that acceleration rates (for all vehicles except trucks) increase from minimum to maximum at initial speeds and then decrease with speed. Sangster *et al.* [19] used naturalistic data ("100-Car study") from different sources such as GPS, On-board Diagnostics (OBD) and accelerometer box for studying the car-following behavior. Their study found that lags in GPS data can result in oscillations in calculated speed. They identified the outliers in the speed and acceleration time series by checking the observed data against the anticipated physical limitations and replaced outliers with the interpolated data. As a result, they transformed the maximum instantaneous acceleration (negative sign for deceleration) ranging [-303.6 m/s^2, 303.0 m/s^2] in the raw data to [-8.9 m/s^2, 9.0 m/s^2] in the smoothed data. Punzo *et al.* [20] used jerk values (derivative of acceleration) to identify the infeasible accelerations in Next Generation Simulation (NGSIM) dataset, which pertains to highway context. Further on, Montanino and Punzo [21] reconstructed trajectories from the NGSIM dataset using steps such as outlier and noise removal and local reconstruction. They adopted a threshold of 8 m/s^2 and 5 m/s^2 for deceleration and acceleration, respectively, for outlier detection, and after reconstruction, longitudinal accelerations range in [-5 m/s^2, 3 m/s^2]. They mention that vehicle mistracking is a likely source of error when extracting the trajectories from the video recordings. Coifman and Lee [22] addressed the vehicle mistracking by manually re-extracting the trajectories with better quality from the NGSIM videos while reporting that the errors in the data cannot be corrected through cleaning or interpolation. Analysis by Kruber *et al.* [23] on a newer dataset [Highway Drone Dataset (highD dataset)] on German freeways found longitudinal accelerations in the range [-6.3 m/s^2, 5.6 m/s^2]. Xu *et al.* [24] collected longitudinal acceleration data using motion sensors on a two-lane mountain highway. They applied data filtering and peak detection algorithms to remove noise and determine the maximum accelerations. According to their findings, acceleration values ranged between [-7.1 m/s^2, 2.8 m/s^2] and [-1.2 m/s^2, 1.4 m/s^2] for small and heavy vehicles, respectively. In the urban driving context, Kanagaraj *et al.* [25] extracted trajectories from a video recording on a section of the road. They smoothed the data using locally weighted regression and obtained longitudinal accelerations in [-4.6 m/s^2, 4.7 m/s^2]. OpenACC is a recently released dataset related to car-

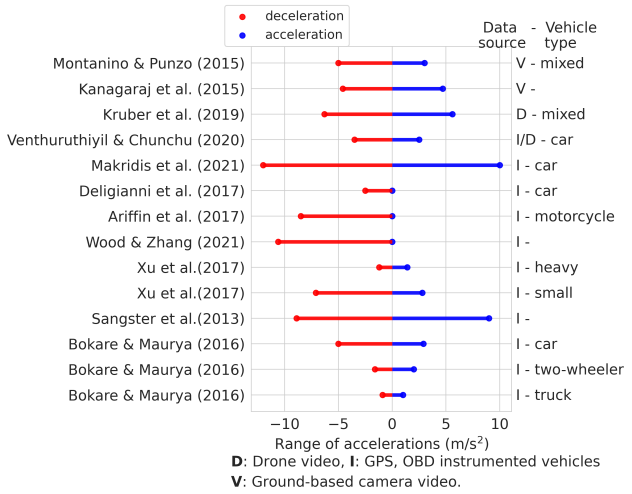


Fig. 1. Acceleration and deceleration ranges in the selected studies [17]–[19], [21], [23]–[26], [28]–[30] including those using naturalistic trajectory data.

following experiments. In this study, Makridis *et al.* [26] used U-Blox M8 devices that were equipped with motion sensors and Global Navigation Satellite System (GNSS) receivers. They post-processed speed and acceleration values using the piece-wise cubic polynomial to compensate for the noise levels in the raw data. The range of accelerations is about [-12 m/s^2, 10 m/s^2] considering human and Adaptive Cruise Control (ACC) drivers. Fard *et al.* [14] used wavelet transform and wavelet-based filter to process the outliers and noise in the NGSIM data. Their method detects outliers based on the local properties of the data and thus is an improvement over globally defined thresholds. Venthuruthiyil and Chunchu [27] reconstructed an error-prone trajectory from video data using locally weighted polynomial regression. In their recent work [17], they processed drone video data by retrieving missing data and then smoothing the data. Before smoothing, they removed the outliers using a median filter. A median filter is a statistical filter wherein the window size and threshold are specified to detect outliers. Afterward, they processed the data using the combination of Recursively Ensembled Low-pass filter (RELP) and adaptive tri-cubic kernel.

D. Summary and research gaps

From the above discussion, we find that studies for highway driving are more prevalent than urban driving. The range of practical or possible acceleration/ deceleration can depend on many factors, such as desired speed, driving context (intersection, highway, ramp), vehicle class and type, driving style, surrounding vehicles, and data sources. One of the challenges of drone videography is that the errors in the data are not consistent as it could be a result of extrinsic (wind burst, object occlusion) and intrinsic (image processing, object tracking [22]) causes, e.g., the reasons for outliers in the drone data could be i) a sudden wind burst that can move the drone, ii) tracked vehicles with reduced visibility (minor roads, occlusion due to buildings or trees), iii) vehicles being tracked in the edges or not well-calibrated areas of the video. Although computer vision algorithms have advanced massively during recent years, it has been recognized in previous studies that trajectory data from drone videography have a heavy-tailed data distribution due to outliers and needs special treatment [5], [17].

Compared to noise treatment, outlier detection is a relatively challenging task to identify systematic issues during the data collection process, thus requiring specialized treatment. Accelerations with unrealistic peaks characterize outliers. They are generally removed in the trajectory datasets using a pre-defined threshold [31] or a statistical filter (such as the median filter used in [17]) on the speed or acceleration series. Such thresholds are manually defined by domain experts with care so that possible driving observations are not classified as outliers (false positives) [21]. Simple heuristics such as global thresholds are also trajectory invariant and cannot account for complex scenarios. An all-embracing and insufficiently flexible filter would not be suitable, as there is always the caveat of over-smoothing or false positives. It overlooks crucial information, especially when it comes to lane-changing maneuvers or aggressive driving behavior like harsh acceleration or harsh deceleration [32]–[34]. Outlier detection is challenging, given the heterogeneity of traffic and driving behaviors.

We want to highlight the drawbacks of simple and popular outlier detection methods, namely z-score or modified z-score algorithms. Using domain expertise to label the anomalies, one needs to specify the window size (over which statistical measure such as mean or median is estimated) and threshold distance (such as the number of standard deviations). There is a trade-off between false positives and false negatives depending on the window size and threshold distance, which emphasizes fine-tuning these parameters. Using mean or median statistics can be biased in urban traffic when the vehicle is stationary at the intersection and thus needs tuning for each vehicle.

Research by Fard *et al.* [14] is based on local detection of outliers using wavelet transforms. The use of wavelet transform has its own challenges, such as the selection of the mother wavelet [14]. Large and diverse datasets, such as the pNEUMA dataset from drones, demand complex heuristics for anomaly detection, and their manual specification is impractical. The above aspects emphasize selecting a scalable methodology for a large dataset, which exploits data-driven or machine learning models and replaces complex heuristics.

While 2D tracking has been used in the AV motion planning-related literature [35] and could reduce the errors in the pNEUMA dataset; we aim to apply fast processing

to remove errors, improving the quality of the data for traffic-oriented purposes. While high-frequency noise can be addressed using available techniques, detecting anomalies is tricky since it is an unsupervised task. As a result, to fully take advantage of such a detailed and massive dataset, it is necessary to find appropriate techniques to detect the outliers (unrealistic transient peaks) and filter them efficiently. Therefore, outlier detection based on the local properties of the data could encompass variations among vehicle class, driving behavior, and anomalies would be helpful and a step toward extending the state-of-the-art. Therefore, we see an opportunity to propose an anomaly detection method that uses complete data (instead of a moving window) to detect the relevant anomalies, i.e., implausible accelerations. We present our methodology in the next section keeping in view the large dataset, high variability in the driving attributes and context, and minimum fine-tuning and processing speed.

III. Methodology

Before removing errors, we analyze the occurrences of excess values of accelerations. Spatially and temporally near vehicles can highly correlate due to traffic flow or car-following behavior. Spatially apart but temporally near vehicles can also show correlated errors due to global events such as wind disturbance to the drone. However, such errors can also occur on account of image processing or data processing. Spatially near but temporally far vehicles can show similar anomalies if passing through the same street obstructed from drone view at different times [6]. However, if the trajectories of vehicles are spatially and temporally apart, we expect little error correlation among them.

Our methodology for treating noise and anomalies is shown in Fig. 2. We treat vehicles one by one so that the noise and anomalies are identified flexibly depending on the vehicle's attributes.

Let us denote the raw speed data by s_t^i for the i^{th} vehicle at time t. We use the SG filter to remove the noise in the speed time series from the raw data to obtain output v_t^i. The output of the SG filter will be biased because the filter is applied to data containing anomalies. However, this is only an intermediate step, and we will address this specific problem subsequently. Smoothed output (from the previous step) is a better choice for evaluating acceleration from the speed time-series (3) than the raw data, as the gradient of noisy data could fluctuate and give even more unrealistic values of the accelerations.

$$a_t^i = \begin{cases} \frac{v_{t+\delta t}^i - v_t^i}{\delta t}, & \text{if } t = 1 \\ \frac{v_{t+\delta t}^i - v_{t-\delta t}^i}{2\delta t}, & \text{if } 1 < t < n \\ \frac{v_t^i - v_{t-\delta t}^i}{\delta t}, & \text{if } t = n \end{cases} \tag{3}$$

Where v is the smoothed speed output of the SG filter, δt= 0.04 seconds as per the frame frequency of 25 Hz.

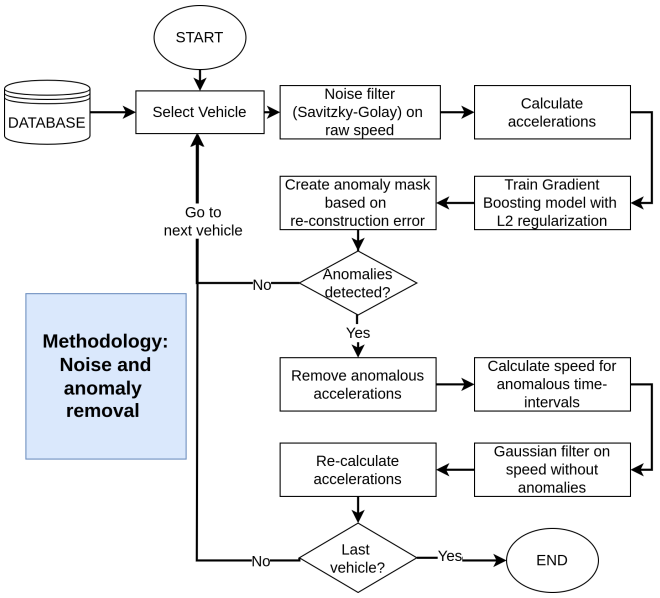


Fig. 2. Methodology flow chart

We aim to detect and process the unreasonable values of the acceleration for the anomaly detection task. This is akin to peak detection in a time series, where the peaks represent anomalous behavior. Our work also makes a similar assumption as [16], where the proportion of representative/ usual data is significantly larger than the anomalous data. This assumption is verifiable by plotting the density plots of data distribution and checking what portion of the data lies within the usual range. Next, we fit a regularized machine learning model to reconstruct the acceleration time series. The use of reconstruction error to classify anomalies is demonstrated in previous studies [36], [37] e.g., Sakurada and Yairi [37] used an autoencoder (a deep learning model) with a regularized objective function for this task.

Instead of a deep learning model, we select XGBoost model [38] for this purpose. XGBoost is based on the concept of Gradient Boosting Machines (GBM) but with certain algorithmic and software enhancements. We select this model because boosting models are generally considered "off-the-shelf classifiers" [39], and thus need lesser feature preprocessing and parameter tuning than other machine learning models such as neural networks. Two basic tunable parameters for a gradient boosting model are the number of iterations (or the number of estimators) and the size of each of the constituent trees (number of leaves in the tree) [39]. Boosting trees are computationally feasible on even large datasets since small trees are used as weak learners (depth of a tree varying between 4 to 8). In Boosting, observations with high residuals generally receive ever-increasing influence with each iteration [39]. Increasing the number of iterations and size of the tree will result in over-fitting. Thus, it is important to stop training the model before it starts to overfit the data. Another way

to prevent over-fitting is by using a regularization (similar to Ridge regression) to shrink the contribution of each tree. XGBoost [38] uses the following regularized loss or objective function:

$$\mathcal{L}^{(k)} = \sum_{t=1}^{T} l\left(a_t, \hat{a}_t^{(k-1)} + f_k\left(\mathbf{x}_i\right)\right) + \Omega\left(f_k\right) \tag{4}$$

Where l is a differential convex loss function that measures the difference between the target a_t and prediction $\hat{a}_t^{(k-1)}$ acceleration, each f_k corresponds to an independent k^{th} tree with structure q and leaf weights w, T is the number of frames (input samples) in the series, and

$$\Omega(f) = \gamma K + \frac{1}{2}\lambda \|w\|^2, \tag{5}$$

K is the number of leaves in the tree, γ is the parameter that penalizes large trees, and λ is the regularization parameter that penalizes the high values of w. The use of regularization controls the over-fitting so that the models are not sensitive to outliers. We refer the reader to the paper by [38] for more details on the XGBoost model.

Ideally, the model should mirror (or fit) the non-anomalous segments except the anomalous segments because we want to preserve all the information in the raw data except the anomalies. To achieve this, we provide the input features consisting of three features: a) smoothed speed series, b) lateral accelerations, and c) acceleration from (3). The target variable for the model is again the same acceleration as the input feature since the aim of the model is to reconstruct the acceleration time series. Therefore, the input features will tend to be correlated. We adopt L2 regularization (Tikhonov's regularization) to prevent over-fitting. The value of the regularization parameter (λ) is adapted for each trajectory, as given by:

$$\lambda^i = b \cdot max(|a^i|)^n, \tag{6}$$

b is a constant, $max(a)$ is the maximum acceleration value observed for a specific trajectory, and n is a positive real number. The rationale for using an adaptive λ is that the vehicle trajectory data could be diverse from different drivers, vehicles, and contexts. Thus, it makes more sense to define an outlier within the context of each trajectory. Therefore, we hypothesize that a single value of λ does not provide this flexibility. λ is directly proportional to the absolute maximum acceleration in the input data since we want the regularization to be highly effective for the unreasonably high acceleration values. High values of b and n cause high penalization of the anomalies, limiting the acceleration values range. In this paper, we use different sets of parameter combinations (b and n) to conduct the sensitivity of the regularization for anomaly detection. It is also possible that different combinations of b and n lead to similar results for specific maximum acceleration values.

We select a sufficiently large fixed value of the number of iterations (say M) and then constrain the model with the L2 regularization term to prevent the model from fitting the outliers. This is motivated by the fact that the regularization term can prevent the model from fitting the extreme points by imposing a high cost. The output of the boosting model is the reconstructed acceleration profile. The reconstructed series should almost replicate the input series for a representative (part of) time series. The reconstructed series will act as a mask to filter the anomalies for the problematic time series. Therefore, the reconstructed series is called an "anomaly mask". We define a tolerance level (τ) and check if the difference between the input series and reconstructed series exceeds that level to label the anomalous sections (0: representative, 1: anomaly) (7). A smaller value of τ will make the model more conservative, i.e., more data will be labeled as anomalous and vice-versa.

$$label_t^i = \begin{cases} 1, & \text{if } |a_t^i - \hat{a^i}_t| \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Where $label_t$ is the anomaly label for the t^{th} frame instance of a trajectory, and $\hat{a}_t$ is the reconstructed output. In our approach, the regularized XGBoost model and τ replace the statistical measures (mean or median) and distance metric (number of standard deviations) to do better than the existing methods without manually adapting the parameters for each trajectory.

Further, if the simultaneous anomalies in the trajectory are detected within a gap of f frames, we assign the complete sub-sequence as anomalous for subsequent correction. This completes the anomaly detection or labeling task. If no anomalies are detected in the previous step, processing for the current trajectory ends, and the next vehicle is selected. Thus, only nominal smoothing via the SG filter is applied to remove the noise in the absence of an anomaly.

After the anomalies are labeled, we need to recalculate the speed ignoring the anomalous accelerations, to obtain the refined or reconstructed speeds. We use the constant acceleration model for speed estimation using (8). The constant acceleration model is reasonable, given the high frame frequency in the trajectory data. Thus acceleration is only considered constant during the one-time step, e.g., 0.04 seconds for data recorded at 25 Hz. This step ensures that the speed and acceleration data are internally consistent, inspired by [21].

$$\hat{v^i}_t = \hat{v^i}_{t-1} + \hat{a^i}_t \cdot \Delta t \tag{8}$$

We replace the speed values for the anomalous points or segments with the reconstructed speed values based on the following:

$$v_t^{i,o} = \begin{cases} \hat{v^i}_t, & \text{if } label_t = 1 \\ s^i{}_t, & \text{otherwise} \end{cases} \tag{9}$$

The above speed time series is treated with the low-pass (Gaussian) filter to recover "unbiased" smoothed speeds. Therefore, this speed series is final since the smoothing of the raw data is done after removing anomalies. Finally, we compute the acceleration values (3) from these speed values. We repeat this process for all the vehicles in the sample. Due to low errors in the position of the tracked vehicle in the pNEUMA dataset, we do not reconstruct or adjust the positions using the processed speed vector. Due to this, we are prone to losing the internal consistency between position and speed [22], and this is a subject of a future study. Finally, we analyze the distribution of acceleration in the detected outliers and the rest of the data for all the vehicles in the sample.

Our method only relies on the feasible range of acceleration for validating the results. It is relevant to point out that we do not conduct jerk value analysis, as done in previous studies by [20], [21]. This is partially mitigated in the above step of Gaussian filter application on the speed series after removing anomalies. This eliminates the sharp edges in the acceleration profile.

IV. Data Collection

In October 2018, the pNEUMA experiment was conducted in Athens, Greece, aiming to record traffic streams over an urban setting using a swarm of ten drones. The pNEUMA experiment aimed to revolutionize how drones as an emerging technology could reshape our understanding of traffic congestion mechanisms. Specifically, the scope was to understand better how congestion forms and propagates in congested multi-modal urban environments through massive data from aerial footage, emphasizing disturbances generated by interactions among different types of vehicles. For the specific experiment, the morning peak (8:00 a.m. to 10:30 a.m.) was recorded for each working day of the week. For improved safety and cooperation, the swarm would take off from the two take-offs/ landing areas (H1 and H2 in Fig. 3) at the start of the experiment, and each drone would go to its area of responsibility. When all drones were at their hovering point, the recording of the traffic stream would start simultaneously, and when the battery ran low, they would return to the landing point. Considering that drones could hover for up to 25 minutes, including take-off, routing, and landing times, it was decided that each session would take place every 30 minutes for better coordination and standardization of the experiment. This setup allows 15 to 20 minutes of continuous monitoring of traffic. During the temporal blind spots, trajectories were not recorded and were not related between sessions.

The analyzed study area includes low, medium, and high-volume arterials, around 100 busy intersections (signalized or not), more than 60 bus stops, and close to half a million trajectories. This massive dataset contains trajectories of

Fig. 3. The study area of pNEUMA experiment. Source: [5]

every vehicle present in the study area, calibrated in the WGS-84 system every 0.04 seconds, as this is the maximum frequency allowed by the video's frame rate. The average ground sampling distance is calculated to be 16.5 cm/px. Except for the features that can be produced using the position information (speed, acceleration, and distance traveled), each vehicle type is available (car, taxi, motorcycle, bus, heavy and medium vehicle). We refer the reader to [5] for more details on the design of the experiment and to [6] for the recently released drone imagery. Since this dataset is also part of an open science initiative shared with the research community, these data are downloadable from https://open-traffic.epfl.ch.

V. Analysis

For this paper, the dataset corresponding to all drones' recordings from the last day of the experiment (10:00 a.m. - 10:30 a.m.) is selected. These data contain trajectories of about 10,500 vehicles with a vehicle's position, speed, and acceleration at 25 Hz.

First, we visualize the acceleration samples of all vehicles (Fig. 4). For motorcycles, median acceleration (0.62 m/s^2) is slightly higher than the other vehicles. However, the range of acceleration for motorcycles is the largest. To check the extreme values, we also visualize the maximum acceleration and deceleration according to the vehicle type (Fig. 4) and find noticeable differences. The median maximum acceleration or deceleration of motorcycles, taxis, and buses is greater than those of cars and medium vehicles. In contrast, heavy vehicles show the lowest median values. This can be partially explained by the different acceleration capabilities (motorcycles vs. heavy vehicles) and driving behavior (taxis vs. private cars). Specifically, motorcycles, due to their limited width and advanced maneuverability compared to other vehicles, make their tracking more challenging [40].

In Fig. 5, we illustrate the maximum accelerations of each vehicle at the location they occur. Excessive accelerations (red areas of the heatmap) appear mainly in the southeast

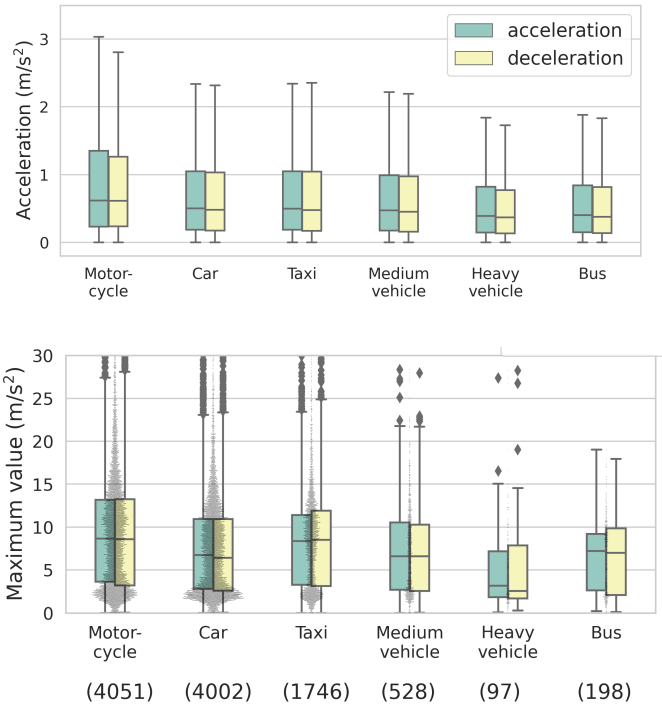


Fig. 4. Distribution considering the (top) all accelerations of all vehicles, and (bottom) maximum acceleration and deceleration for all vehicles before treatment. The number of unique vehicles for each category is mentioned in parentheses.

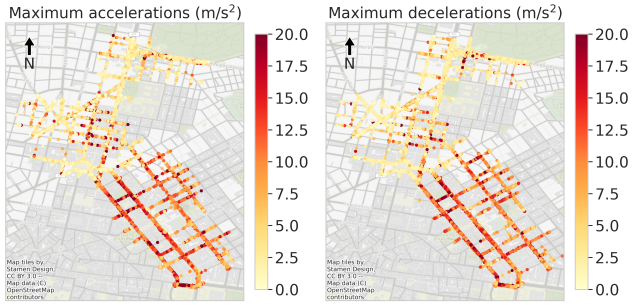
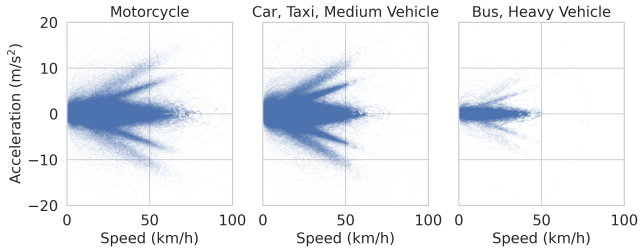


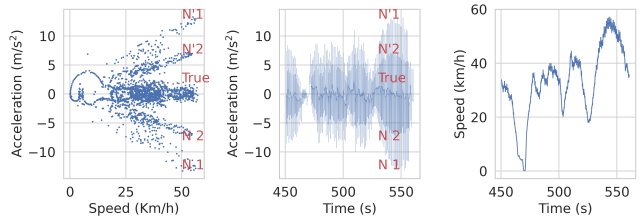
Fig. 5. Heatmap showing location-wise maximum acceleration and maximum deceleration for each vehicle.

(captured by drone numbers 1, 2, 3, and 4), whereas such instances are uncommon in the north. These can be explained by the limitations of drone videography, such as intersections due to bad lighting (shade, low contrast), roads on the edge of the recorded video (due to video distortion), other tracking issues, and data post-processing. While during the pNEUMA experiment, the videos were stabilized, the current work did not test for evidence of residual camera motion. Also, we did not see any evidence suggesting the presence of such non-vehicular motion during the experiment, processing of the videos, or data analysis efforts.

In the pretreated data, the speed-acceleration plot (Fig. 6) shows four "flanks" (two each for acceleration and deceleration). We discover these flanks are due to the noise in the speed (and acceleration) vector. The slope of the flanks has a unit of frequency (s^{-1}) and is about constant, which points to the presence of two components of noise with fixed frequencies in the dataset. Further investigation shows that this periodic noise is not present in the individual drone recordings but occurred while merging the datasets, e.g., drone 2 and drone 3. Thus, it is not related to the nature of the experiment or the CV algorithm but to the specific dataset.



(a) Pretreated data showing presence of four flanks (for a sample of 1000 vehicles)



(b) Example with two noise components (N1 and N2)

Fig. 6. Acceleration-speed plots

We calculate the number of vehicles for which the magnitude of acceleration and deceleration exceeds a cut-off limit (5 m/s^2, 10 m/s^2, 15 m/s^2, 20 m/s^2) over time. In Fig. 7, we expect and see that excess acceleration occurrence reduces with the increase in the cut-off limit. At a cut-off of 5 m/s^2, we see occurrences corresponding to the previously mentioned two components of high-frequency noise. Further, these occurrences are also sinusoidal over about 90 seconds. At a cut-off of 10 m/s^2, only one of the two high-frequency noises (about 1 Hz) is noticeable, meaning that the amplitude of the other noise is lesser than 10 m/s^2. Here too, the sinusoidal nature of occurrences is even more noticeable at a period of about 90 seconds. This systematic periodicity points to green waves on the arterial roads covered by drones 2 and 3, during which new traffic enters and leaves these arterial roads. At the same time, these sinusoidal occurrences exclude the effect of wind blows, which tend to be random. No high-frequency and sinusoidal occurrences are noticeable at a cut-off of 15 m/s^2 and 20 m/s^2, which means that very high

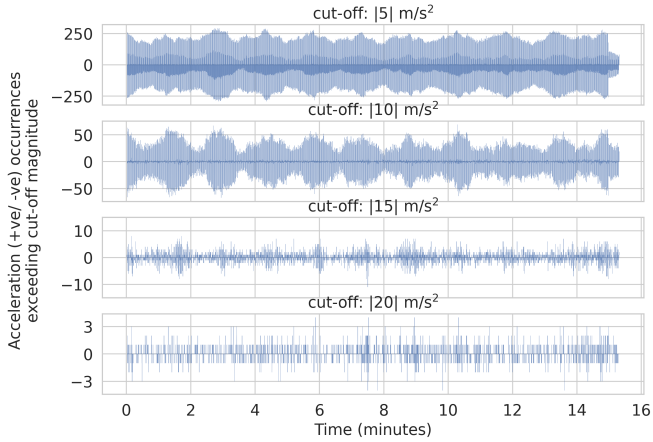


Fig. 7. Occurrences (positive for longitudinal acceleration and negative for longitudinal deceleration) exceeding the cut-off limit.

acceleration values, such as more than 20 m/s^2, are scattered.

Fig. 8. Heatmap showing location-wise speeds and accelerations (exceeding 6 m/s^2) at frames 45 seconds apart to highlight the correlation between them and the periodicity of excess accelerations at about 90 seconds. Similar findings were found for deceleration.

Further spatial analysis reveals that noise and sinusoidal behavior of anomalous instances is prevalent on drone recordings 1, 2, 3, and 4. This is shown in Fig. 8. The correlation between vehicle influx at high speeds (green wave) and excessive acceleration is noticeable in the bottom part of the map. Accurate vehicle tracking can be challenging for the object tracker during this sudden acceleration and deceleration behavior, thus resulting in such periodicity of anomalous instances. For terminology, we conclude that periodic excessive acceleration values are attributed to noise, whereas random (transient) peaks are anomalies.

In Fig. 9, the challenge of detecting the actual noise becomes even trickier due to the inconsistency of its occurrence. For example, vehicle trajectory id 758 (Fig. 9) does not need treatment; the same does not apply to vehicle id 1490, 493, or 146. The longitudinal acceleration is noisy for the whole trajectory and shows unrealistic values (around t= 40 s for vehicle 146). Additionally, for vehicle id 1780, it is seen that from time t_1= 0 s to t_2= 150 s, the noise is negligible. In contrast, specific treatment is necessary for the trajectory beyond t_2= 150 s due to unrealistic longitudinal acceleration values. In Fig. 9, we labeled the data according to the drone and thus found that noise is primarily contributed by drones 1, 2, 3, and 4. When we visualize the longitudinal acceleration for these vehicles over a few seconds (Fig. 10), we find that noise is synchronized for vehicles 146, 493, and 1490, which complements previous findings (Fig. 7 and Fig. 8) about temporal synchronization of noise and specific locations/ drones contributing to the noise. Thus, periodic noise and unrealistic transient values should be eliminated to recover the desired data.

VI. Results

The acceleration calculated from the second derivative of the position is noisy. It has unreasonably high values, possibly due to accumulated errors, as seen by the incidence of values up to $\pm$100 m/s^2 in Figure 11(a). The histograms in Fig. 11 are plotted on the log scale to visualize heavy-tailed data easily. Such large values are also expected, given the spatial resolution and high sampling frequency (25 Hz).

On the other hand, the first derivative of the speed results in less extreme values, shown in Fig. 11(b). This could be because the application of some smoothing filter preprocesses the speed attribute in the pNEUMA data. Thus the processed speed series is a better candidate for calculating acceleration, which is also noted by [18]. Indeed we find that the first derivative of the position with a moving average filter of a 1-second window (25 frames) has a somewhat similar distribution as the speed attribute given in the data. Therefore, we rely on the speed attribute for our model and take its first derivative to calculate the acceleration further.

The acceleration values range up to $\pm$75 m/s^2, emphasizing further data processing. We find that there is still high-frequency noise in the data, and it needs treatment. The SG filter is best suited for this task, also evident in Fig. 11(b). The SG filter of the polynomial order one is denoted by SG^1 (same as the moving average filter) and performs better than the polynomial of order two (SG^2). The acceleration distribution is also improved by a substantial reduction in its heavy-tailedness on the aggregate scale. In Fig. 11(b), the SG^1 filter removes the high-frequency noise and recovers the noise-free signal. Moreover, it also reduces the extreme values since each point is a weighted average of its neighborhood. Despite this, the anomalies remain, and thus the need for anomaly detection.

We use the XGBoost implementation [38] in Python for our study. We set the number of iterations or estimators for the XGBoost model as 300 and use different types of the regularization parameter, i.e., fixed or adaptive (6). Fig. 12 shows

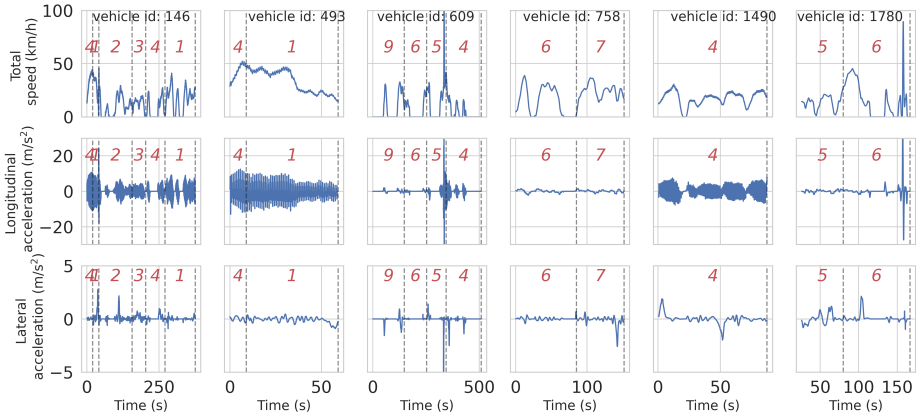


Fig. 9. Speed, longitudinal acceleration, and lateral acceleration plots of six vehicles showing different characteristics. The number within vertical dashed lines indicates the id of the drone which recorded the corresponding trajectory segment.

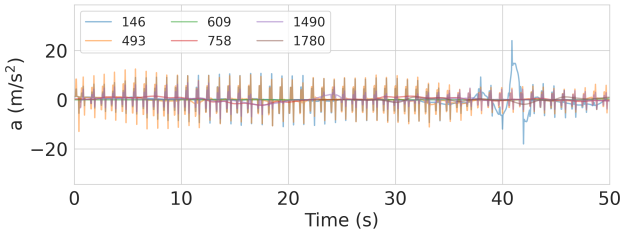
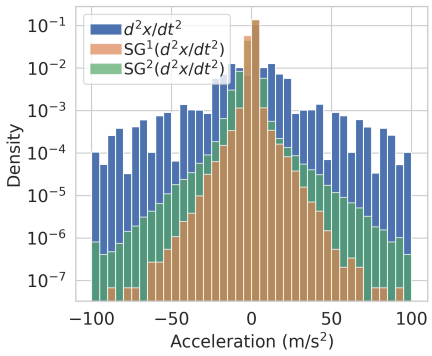


Fig. 10. Temporal synchronization of noise in the longitudinal acceleration of selected vehicles

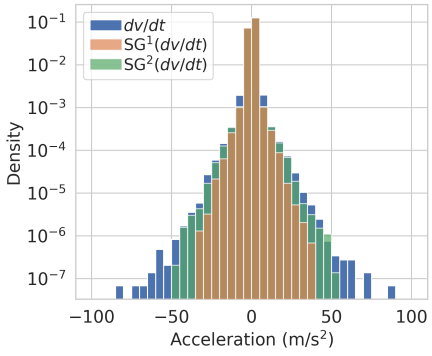
the effect of regularization. We find that L2-regularization prevents the reconstructed profile from achieving extreme values. A high λ squeezes the range of the acceleration values, as the histogram for $\lambda = 2000$ is narrower than that for $\lambda = 500$. After these preliminary trials, we use the adaptive λ (with $b = 20$ and $n = 2$), which performs better than the fixed λ in constraining the range of reconstructed accelerations.

We analyze the characteristics of the anomaly decision boundary in Fig. 12. The decision boundary is correlated with the maximum acceleration simply due to the formulation of the L2 regularization in (6). The Lowess fit (Locally Weighted Scatterplot Smoothing) shows a linear trend between (post-processed) maximum acceleration and anomaly decision boundary at small values. At high values, the decision boundary is asymptomatic at around 4 m/s^2, but the range of values goes up to 7 m/s^2. Thus the anomaly detection threshold is not fixed and depends on the trained model. We see this as an advantage of our approach compared to the fixed threshold, which does not adjust to vehicle-specific kinematics.

The sensitivity analysis is done on the regularization

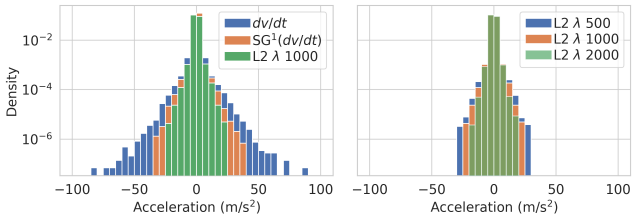


(a) Using position coordinates

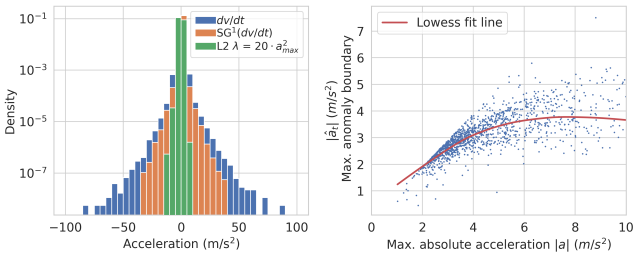


(b) Using processed speed series

Fig. 11. Savitzky-Golay filter to remove noise from acceleration series. the y-axis is plotted on the log scale for better visualization of distribution tails



(a) (left) Fixed λ and (right) effect of increasing λ



(b) (left) Adaptive λ with $b = 20, n = 2$ and (right) anomaly decision boundary

Fig. 12. Effect of regularization parameter λ.

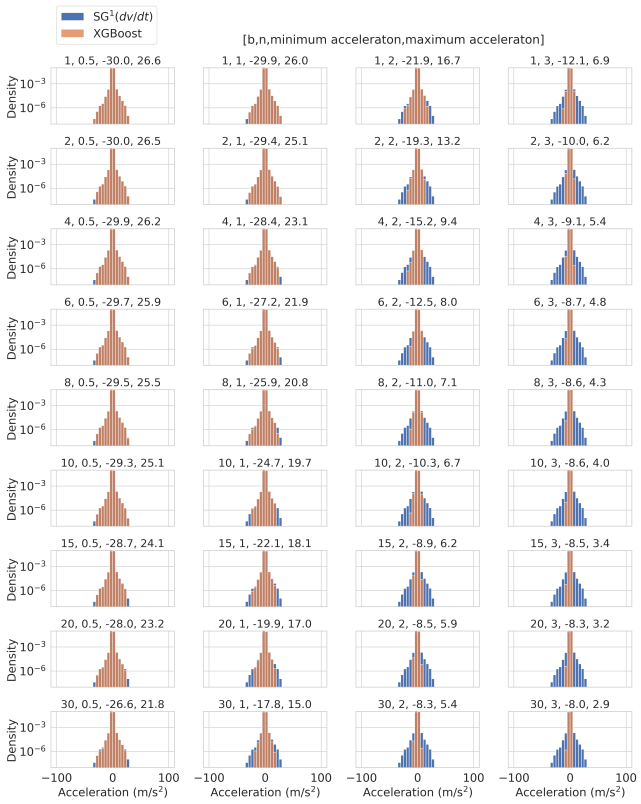


Fig. 13. Sensitivity analysis of the parameters b and n

parameters used in the boosting model (b and n). We adopt $b \in \{1, 2, 4, 6, 8, 10, 15, 20, 30\}$ and $n \in \{0.5, 1, 2, 3\}$ for a grid-based evaluation based on the combination of these parameters. We use the values of other parameters: τ and f as 0.1 and 10, respectively. The results are shown in Figure 13. For low values ($b = 1$, $n = 0.5$), there is hardly any reduction in the anomalies in the reconstructed trajectories. In contrast, the reconstructed trajectories are heavily biased due to false positives for the high values ($b = 30$, $n = 3$). It is also observed that the proposed method is more sensitive to n and less sensitive to b. In between ($15 \leq b \leq 20$, $n = 2$), the parameter values are optimal for our task. These findings also show that different combinations of b and n can produce similar results for specific maximum acceleration values.

After removing anomalies, we use the Gaussian filter for smoothing the data, i.e., the filter is applied to the original data without anomalies. We find that window sizes between 12 and 25 frames provide an acceptable range of processed accelerations. Thus, we do not recommend a single best value but a range of values for anomaly detection and smoothing, which can provide a practical solution. This is obvious because our criteria for acceptance are based on the range of final accelerations. However, it is an essential conclusion since multiple optimal parameter combinations help approximately recover the desired signal from the raw data.

We demonstrate the complete methodology in Fig. 14, showing the de-noising, anomaly detection and removal, retrieval of consistent speeds, and accelerations. Here window sizes for the SG and Gaussian filters are set as 25 frames and 25 frames, respectively. In the third and fourth subfigure in Fig. 14 for each vehicle, applying a low-pass filter without removing anomalies will result in biased profiles due to extreme values. We also demonstrate cases in Fig. 15 where the trajectories have either noise, anomalies, or none of both. In the case of vehicle id 493 and 505, reconstruction of speed and acceleration is skipped since no anomaly is detected. This also shows our method treats data so as not to cause significant over-smoothing when the data are without noise or anomalies. Thus the final output in both cases is the result of the smoothing only. The post-processed maximum acceleration and deceleration for all types of vehicles (Fig. 16) are within the reasonable range because their values for most of the sample vehicles' trajectories are less than 4 m/s^2. Still, for a few samples, values go up to 9 m/s^2. Fig. 16 should be compared with Fig. 4 to see the effect of noise and anomaly treatment.

In acceleration-speed plots (Fig. 17), we show the stepwise treatment process on 1000 vehicle trajectories with extreme acceleration values. The top row in Fig. 17 shows unrealistically high acceleration values due to noise and anomalies. The output after interim noise treatment using the SG filter (window size: 25 frames) is shown in the panel's second row (from the top). This is followed by the

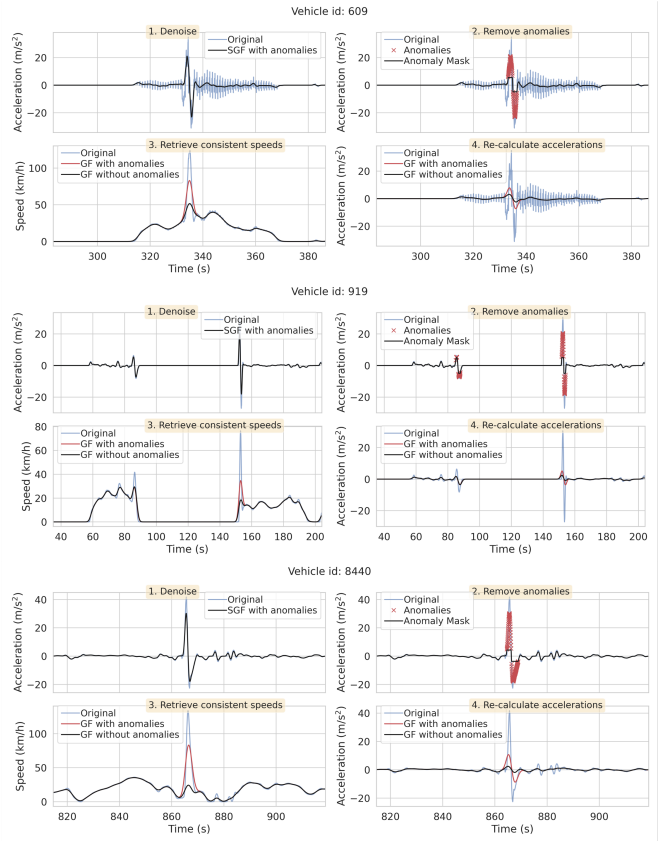


Fig. 14. Individual steps in the treatment of noise and anomalies for three example vehicles ids

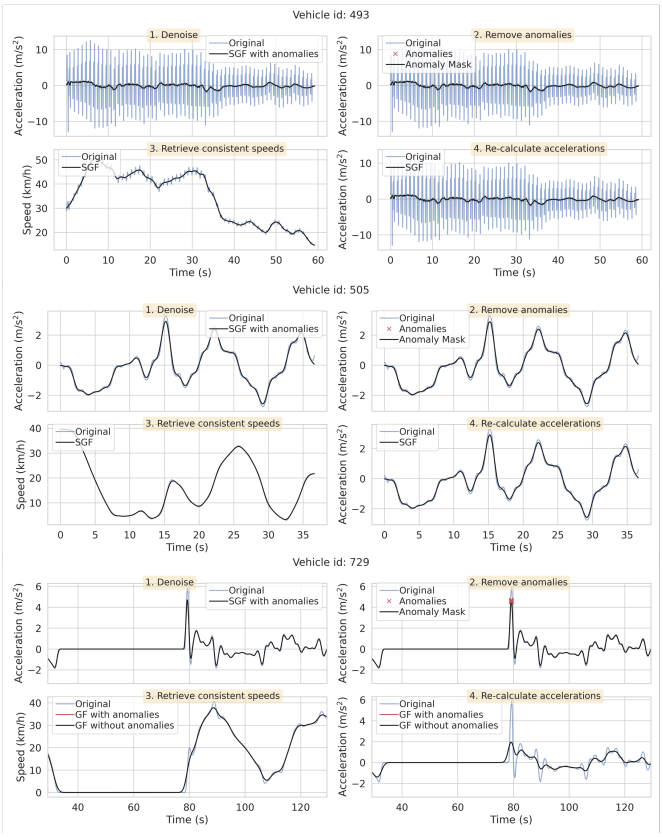


Fig. 15. Treatment examples when trajectory has (top) only noise, (middle) neither noise nor anomalies and (bottom) only anomalies.

anomaly treatment in the third row. The output after the final smoothing (after the removal of anomalies) by Gaussian filter (window size: 25 frames) is shown in the bottom row of Fig. 17. The final data in the speed-acceleration plot shows that the range of accelerations is confined within the reasonable range. Over-smoothing can distort the time-space diagram by drastically changing the speed or distance traveled compared to the pretreated values. To verify this, we relied on time-space diagrams of a sample of the vehicles. We did not find significant differences between the distance traveled with pre-treatment and that with post-treatment speeds.

It is also relevant to provide processing time statistics, which can depend on many factors, such as the hardware specifications, parallelization of the algorithms, and the number of samples in each trajectory. We used an HP desktop Machine with eight physical cores (i7-11700F @ 2.50GHz) and 16 GB RAM. We run our method sequentially, i.e., all vehicles are treated one by one. We use two cores for the XGBoost model via the parameter *n_jobs*. The computation mainly involves calculating numerical gradients, manipulating data frames, array operations, XGboost model training, and applying low-pass filters. We record the run-time statistics per vehicle trajectory and find that the run-time

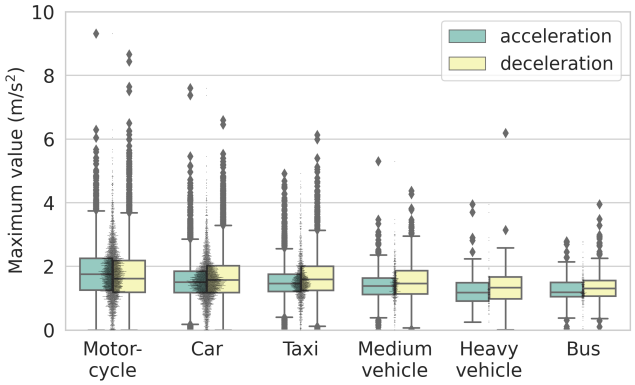


Fig. 16. Distribution of the maximum acceleration and deceleration for vehicles in the dataset after treating anomalies and noise. The same sample is used here as in Fig. 4

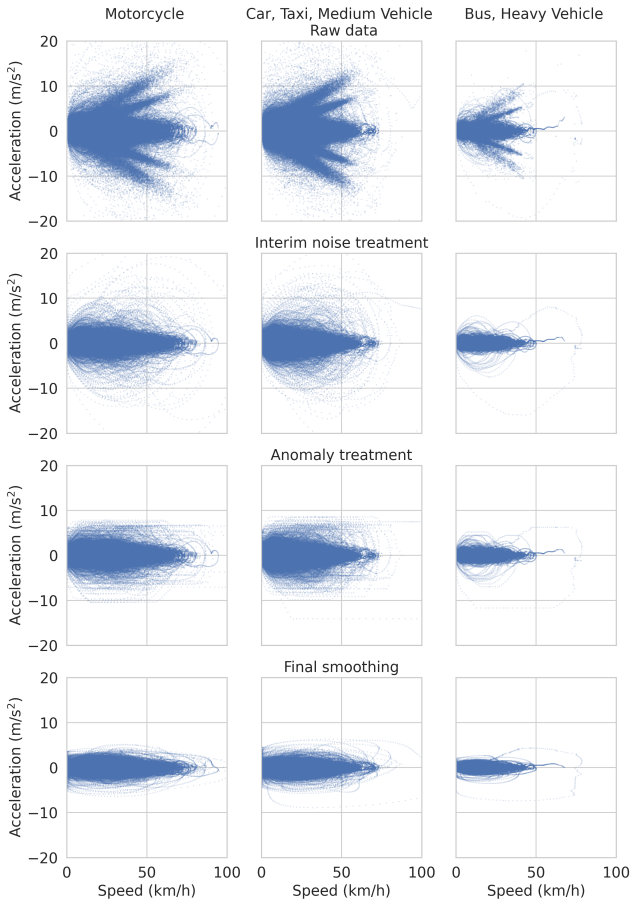


Fig. 17. Step-wise treatment output and errors

mean, median and standard deviation are 0.77 s, 0.59 s, and 1.02 s, respectively.

VII. Conclusion

Emerging traffic data collection methods have their benefits and challenges, which is true for drone data. Before using these data, noise filtering or anomaly detection are essential steps to improve the data quality by minimizing implausible values in the data. When dealing with vehicle trajectories, a balance should be maintained between filtering the data and retaining naturalistic driving behavior. In this paper, we demonstrate the application of noise smoothing and anomaly detection on the data from the aerial footage of the pNEUMA dataset. We use SG filter, XGBoost with adaptive regularization, and GF to remove the noise and anomalies. We show that our approach can accurately detect anomalies in the form of unrealistic transient peaks in the data. Adaptive regularization adjusts itself based on the maximum acceleration value and simplifies anomaly detection to fewer tunable parameters. Using an off-the-shelf model such as XGBoost reduces the number and effort of tuning the model.

Our approach can also be adapted to other trajectory or sequence datasets corrupted by noise and anomalies. However, for successful transfer, anomalies in new data should be similar to those in the pNEUMA data, i.e., a few unrealistic transient acceleration peaks. The treated data are much more suitable for microscopic traffic analysis, such as road safety analysis using surrogate measures or traffic flow modeling (car-following or lane-changing), and can thus help accelerate future research. In our study, we identified a range of parameters for anomaly detection and smoothing, which provide acceptable results. This indicates that any subsequent analysis (traffic flow, emissions, crash safety) using trajectory data will also be sensitive to these parameters. Thus, the researchers should estimate confidence intervals to quantify their results' uncertainty.

The research is not without its limitations. As discussed, while some of the errors are attributed to limitations of drone videography, the absence of ground truth labels prevents us from verifying the driving behavior or the exact causes behind the detected anomalies in the acceleration. Therefore, treated data might still contain errors beyond the definition of anomalies used in this study. For instance, we only addressed the anomalies of the unrealistic-peak character, but the errors of other characters may also be present in the dataset. In the absence of ground truth labels, validation of results is challenging, and the implications of such errors depend on the requirements and sensitivity of the subsequent analysis. Finally, machine learning model training can be ineffective in extremely short trajectories due to a lack of data. Although XGBoost is less data-hungry when compared to deep neural networks, XGBoost's performance to detect anomalies can still be affected when data is scarce. Future works could be done to adjust the trajectory positions as per the treated speed to ensure internal consistency [20], [21] (between position and speed), and platoon consistency (with leader vehicle and follower vehicle). Another crucial future work is decomposing the processed speed and acceleration vectors into longitudinal and lateral components relative to the street's orientation for analyzing lateral driving maneuvers.

VIII. Acknowledgments

We sincerely thank the three anonymous reviewers for their valuable comments and helping us enrich the manuscript.

References

[1] C. Antoniou, R. Balakrishna, and H. N. Koutsopoulos, "A synthesis of emerging data collection technologies and their impact on traffic management applications," *European Transport Research Review*, vol. 3, no. 3, pp. 139–148, Nov 2011.

[2] H. González-Jorge, J. Martínez-Sánchez, M. Bueno, and P. Arias, "Unmanned aerial systems for civil applications: A review," *Drones*, vol. 1, no. 1, p. 2, Jul 2017. [Online]. Available: http://dx.doi.org/10.3390/drones1010002

[3] E. N. Barmpounakis, E. I. Vlahogianni, and J. C. Golias, "Unmanned aerial aircraft systems for transportation engineering: Current practice and future challenges," *International Journal of Transportation Science and Technology*, vol. 5, no. 3, pp. 111 – 122, 2016, unmanned Aerial Vehicles and Remote Sensing.

[4] H. Q. Pham, M. Camey, K. D. Pham, K. V. Pham, and L. R. Rilett, "Review of unmanned aerial vehicles (uavs) operation and data collection for driving behavior analysis," in *CIGOS 2019, Innovation for Sustainable Infrastructure*, C. Ha-Minh, D. V. Dao, F. Benboudjema, S. Derrible, D. V. K. Huynh, and A. M. Tang, Eds. Singapore: Springer Singapore, 2020, pp. 1111–1116.

[5] E. Barmpounakis and N. Geroliminis, "On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 50 – 71, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X19310320

[6] S. Kim, G. Anagnostopoulos, E. Barmpounakis, and N. Geroliminis, "Visual extensions and anomaly detection in the pneuma experiment with a swarm of drones," *Transportation Research Part C: Emerging Technologies*, Jan 2023.

[7] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: a systematic review," *Journal of Big Data*, vol. 7, no. 1, p. 11, Feb 2020.

[8] T. O'Haver, "Signals and noise," in *A Pragmatic Introduction to Signal Processing*, T. O'Haver, Ed. Maryland, USA: O'Haver, Tom, 2022, p. 23.

[9] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 03 1960.

[10] J. L. Holloway, "Smoothing and filtering of time series and space fields," ser. Advances in Geophysics, H. Landsberg and J. Van Mieghem, Eds. Elsevier, 1958, vol. 4, pp. 351–389.

[11] R. W. Schafer, "What is a savitzky-golay filter? [lecture notes]," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[12] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[13] R. Fisher, S. Perkins, A. Walker, and E. Wolfart. (1997) 71 - hypermedia image processing reference (hipr). http://homepages.inf.ed.ac.uk/rbf/HIPR2/morops.htm (zuletzt aufgerufen: 18.10.2012).

[14] M. Rafati Fard, A. Shariat Mohaymany, and M. Shahri, "A new methodology for vehicle trajectory reconstruction based on wavelet analysis," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 150–167, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X16302261

[15] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, Jul. 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882

[16] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, p. 255–262.

[17] S. P. Venthuruthiyil and M. Chunchu, "Vehicle path reconstruction using recursively ensembled low-pass filter (relp) and adaptive tri-cubic kernel smoother," *Transportation Research Part C: Emerging Technologies*, vol. 120, p. 102847, 2020.

[18] P. Bokare and A. Maurya, "Acceleration-deceleration behaviour of various vehicle types," *Transportation Research Procedia*, vol. 25, pp. 4733–4749, 2017, world Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016.

[19] J. Sangster, H. Rakha, and J. Du, "Application of naturalistic driving data to modeling of driver car-following behavior," *Transportation Research Record*, vol. 2390, no. 1, pp. 20–33, 2013.

[20] V. Punzo, M. T. Borzacchiello, and B. Ciuffo, "On the assessment of vehicle trajectory data accuracy and application to the next generation simulation (ngsim) program data," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1243–1262, 2011.

[21] M. Montanino and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," *Transportation Research Part B: Methodological*, vol. 80, pp. 82 – 106, 2015.

[22] B. Coifman and L. Li, "A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset," *Transportation Research Part B: Methodological*, vol. 105, pp. 362–377, 2017.

[23] F. Kruber, J. Wurst, S. Chakraborty, and M. Botsch, "Highway traffic data: macroscopic, microscopic and criticality analysis for capturing relevant traffic scenarios and traffic modeling based on the highd data set," 2019.

[24] J. Xu, W. Lin, X. Wang, and Y.-M. Shao, "Acceleration and deceleration calibration of operating speed prediction models for two-lane mountain highways," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 7, p. 04017024, 2017.

[25] V. Kanagaraj, G. Asaithambi, T. Toledo, and T.-C. Lee, "Trajectory data and flow characteristics of mixed traffic," *Transportation Research Record*, vol. 2491, no. 1, pp. 1–11, 2015.

[26] M. Makridis, K. Mattas, A. Anesiadou, and B. Ciuffo, "Openacc. an open database of car-following experiments to study the properties of commercial acc systems," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103047, 2021.

[27] S. P. Venthuruthiyil and M. Chunchu, "Trajectory reconstruction using locally weighted regression: a new methodology to identify the optimum window size and polynomial order," *Transportmetrica A Transport Science*, vol. 14, no. 10, pp. 881–900, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2324993522001683

[28] J. S. Wood and S. Zhang, "Evaluating relationships between perception-reaction times, emergency deceleration rates, and crash outcomes using naturalistic driving data," *Transportation Research Record*, vol. 2675, no. 1, p. 0361198120966602, 2021.

[29] A. Ariffin, A. Hamzah, M. Solah, N. Paiman, Z. M. Jawi, and M. M. Isa, "Comparative analysis of motorcycle braking performance in emergency situation," *Journal of the Society of Automotive Engineers Malaysia*, vol. 1, no. 2, 2017. [Online]. Available: http://jsaem.saemalaysia.org.my/index.php/jsaem/article/view/53

[30] S. P. Deligianni, M. Quddus, A. Morris, A. Anvuur, and S. Reed, "Analyzing and modeling drivers' deceleration behavior from normal driving," *Transportation Research Record*, vol. 2663, no. 1, pp. 134–141, 2017.

[31] M. Montanino and V. Punzo, "Making ngsim data usable for studies on traffic flow theory: Multistep method for vehicle trajectory reconstruction," *Transportation Research Record*, vol. 2390, no. 1, pp. 99–111, 2013. [Online]. Available: https://doi.org/10.3141/2390-11

[32] E. Barmpounakis, G. M. Sauvin, and N. Geroliminis, "Lane detection and lane-changing identification with high-resolution data from a swarm of drones," *Transportation Research Record*, vol. 2674, no. 7, pp. 1–15, 2020.

[33] E. I. Vlahogianni and E. N. Barmpounakis, "Driving analytics using smartphones: Algorithms, comparisons and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 196 – 206, 2017.

[34] V. Mahajan, C. Katrakazas, and C. Antoniou, "Prediction of lane-changing maneuvers with automatic labeling and deep learning," *Transportation Research Record*, vol. 2674, no. 7, pp. 336–347, 2020.

[35] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, pp. 1826–1848, 2020.

[36] N. Japkowicz, C. Myers, and M. Gluck, "A novelty detection approach to classification," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 518–523.

[37] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, ser. MLSDA'14. New York, NY, USA: Association for Computing Machinery, 2014, p. 4–11. [Online]. Available: https://doi.org/10.1145/2689746.2689747

[38] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[40] E. N. Barmpounakis, E. I. Vlahogianni, and J. C. Golias, "Intelligent transportation systems and powered two wheelers traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 908–916, 2016.

Vishal Mahajan is a doctoral candidate at the Chair of Transportation Systems Engineering at the Technical University of Munich (TUM). He holds a Master of Science degree in Transportation Systems from the Technical University of Munich, Germany. His goal is to check and establish the efficacy of diverse publicly available data sources in traffic research and transport modeling. He is interested in using the publicly available data sources (including Open Data) to develop and calibrate large-scale traffic simulation models for real-time traffic management. To achieve his research goals, Vishal applies statistical and machine learning techniques for predictive modeling.

Emmanouil (Manos) Barmpounakis is a post-doctoral researcher in the Urban Transport Systems Laboratory (LUTS) of the École Polytechnique Fédérale de Lausanne (EPFL). He has a Ph.D. and diploma in Civil Engineering from the National Technical University of Athens (NTUA). His research interests focus primarily on urban transportation systems, traffic flow monitoring, microscopic traffic modeling, unmanned aerial systems and data science. He is a member of the editorial board of the International Journal of Transportation Science and Technology and a reviewer in 6 International Journals. He is a member of the IEEE ITSC Technical Committee on Smart Cities and Smart Mobility. He is part of the https://open-traffic.epfl.ch/ open-science initiative that includes a dataset of naturalistic urban trajectories collected by a one-of-a-kind experiment with a swarm of drones.

Md Rakibul Alam is a doctoral candidate (since 2020) at the Chair of Transportation Systems Engineering at the Technical University of Munich (TUM), Germany. He had completed his M.Sc. in Informatics from the same university. He had earned his B.Sc. in Computer Science from the Independent University, Bangladesh. He is currently a research associate at the Chair of Transportation Systems Engineering at TUM, and his research interests include data engineering and architecture in intelligent transportation systems.

Nikolas Geroliminis is a Full Professor at EPFL and the head of the Urban Transport Systems Laboratory (LUTS). He holds a diploma in Civil Engineering from NTUA, Greece and a MSc and Ph.D. in civil engineering from University of California, Berkeley. His research interests focus primarily on urban transportation systems, traffic flow theory and control, public transportation and on-demand transport, car sharing, Optimization, MFDs and Large Scale Networks. He is a recipient of the ERC Starting Grant METAFERW: Modeling and controlling traffic congestion and propagation in large-scale urban multimodal networks. He currently serves as Editor-In-Chief of Transportation Research part C: Emerging Technologies journal.

Constantinos Antoniou is Full Professor in the Chair of Transportation Systems Engineering at the Technical University of Munich (TUM), Germany. He holds a Diploma in Civil Engineering from NTUA (1995), a MS in Transportation (1997) and a PhD in Transportation Systems (2004), both from MIT. His research focuses on data analytics, modeling and simulation of transportation systems, Intelligent Transport Systems (ITS), calibration and optimization applications, road safety and sustainable transport systems.