Road Traffic Density Estimation Based on Heterogeneous Data Fusion

Philipp Zißner*, Paulo H. L. Rettore*, Bruno P. Santos†, Roberto Rigolin F. Lopes*, and Peter Sevenich*
*Dept. of Communication Systems, Fraunhofer FKIE, Bonn, Germany
Email: {philipp.zissner, paulo.lopes.rettore, roberto.lopes, peter.sevenich}@fkie.fraunhofer.de
†Dept. of Computer and Systems, Federal University of Ouro Preto, Joao Monlevade, Brazil
Email: bruno.ps@ufop.edu.br

***Abstract*—This investigation starts with the hypothesis that fusing heterogeneous data sources can increase the data coverage and improve the accuracy of traffic-related applications in Intelligent Transportation Systems (ITS). Therefore, we designed (i) a Data Fusion on Intelligent Transportation Systems (DataFITS) framework that allows collecting data from numerous sources and fusing them according to spatial and temporal criteria; (ii) a traffic estimation method that groups road segments into regions, identify correlations between them, and measure the traffic distribution to estimate traffic. As a result, DataFITS increased by 130% the number of road segments coverage and enhanced, by fusion process, around 35% of road overlapping data sources. We evaluate the traffic estimation of the 15 most correlated regions, where the fused data together with correlated areas resulted in the best traffic estimation accuracy by reaching up to 40% in some cases and 9% on average.**

***Keywords* - ITS, Smart Cities, Traffic estimation, Data Fusion**

I. Introduction

Intelligent Transportation Systems (ITSs) are a key concept to improve current transportation systems in Smart Cities by using data and communication technologies. ITSs aims to enhance traffic efficiency and reduce accidents while minimizing the time wasted in traffic jams, emissions, and fuel consumption. In this context, traffic can influence cities' dynamics by causing delays in the mobility of people and goods. Thus, understanding traffic behavior and transportation systems may help to better manage it by re-routing vehicles and adjusting traffic flow, which can improve overall urban mobility.

To better understand and model traffic behavior a significant amount of data is required. These data can be collected through various sources such as built-in vehicle sensors, traffic monitoring systems, news, social media, smartphones, and many others [1]. Traffic-related data may assist the ITSs in predicting road events such as traffic and accidents more accurately. Although there are plenty of data sources available nowadays, most of them provide data with low quality (e.g., Spatio-temporal gaps) or limited free data access.

In this sense, the data fusion concept, which attempts to improve these aspects by fusing many diverse data sources, can be a solution to the data quality and availability problem. While this appears to be a promising approach, the process is challenging due to data issues such as different data structures, errors in the acquisition and acquired data (e.g., wrong measurement, missing values), outliers, conflict, incompleteness, and vagueness [2].

In this paper, we propose a traffic density estimation framework based on heterogeneous data fusion. First, we designed the Data Fusion on Intelligent Transportation Systems (DataFITS), a framework that aggregates data from heterogeneous data sources and fuses them temporally and spatially. As a result, DataFITS produces more enriched and varied data that describes the transportation system. Second, we use the fused data to feed our proposed traffic density estimation model. The method groups road segments into regions and then it draws correlation between them to increase the sample of data. Such data will be used latter to estimate traffic based in the computed historical traffic density distribution data. Lastly, we evaluate the accuracy of our traffic estimation using various metrics (Coefficient of determination (R2), Dynamic-Time-Warping (DTW) [3], Granger Causality (GC) [4], Mean absolute error (MAE) and Root-Mean-Square Error (RMSE)). As a result, we show that by using fused heterogeneous data, it is possible to improve the traffic estimation accuracy up to 40% in some cases and 9% on average.

The main contributions of this paper are listed below:

- The design and evaluation of a heterogeneous data fusion framework.
- A traffic correlation over different regions in a city based on time windows and weekdays.
- A traffic density estimation based on simple statics using heterogeneous fused data.
- Discussion over quantitative results from experiments using real data, testing the benefits of data fusion to enhance a traffic estimation application.

In the remainder, this paper is organized as follows: First, Section II briefly reviews the recent literature, discussing the usage of heterogeneous data fusion and different traffic estimation models. In Section III, we explain the design of DataFITS and the traffic density estimation. The evaluation of the framework is described in Section IV, creating various data samples to test the accuracy of the respective traffic estimation. Finally, we conclude this paper in Section V, also reviewing future directions.

II. Related Work

This section examines the recent literature on data fusion techniques to support ITS, with a focus on investigations that describe traffic prediction models.

978-1-6654-9792-3/22/$31.00 ©2022 IEEE

In [5] the authors provide a platform to gather, process, and export heterogeneous data from smart city sensors and create various data visualizations. We share a similar motivation, further providing a fusion approach and a methodology for traffic prediction, instead of solely designing the data platform.

In general, data fusion is challenging due to the data semantic heterogeneity and the different spatio-temporal aspects [1], [2]. The present paper was motivated by our previous investigations [6]–[9] also using data fusion to provide more reliable and accurate applications for transportation systems within a smart city. Moreover, inspired by the ideas in [10], we extended Traffic Data Enrichment Sensor (TraDES) by designing a general data fusion framework DataFITS in this investigation. TraDES is a low-cost traffic sensor for ITS that combines vehicle and traffic data to train a machine-learning algorithm to learn from automotive characteristics such as speed, CO_2 emissions, fuel consumption, and so on, the road traffic levels, expanding the spatio-temporal data coverage.

Furthermore, there has been a lot of research conducted in recent years on traffic prediction and forecasting. Abadi, et al. [11] developed a model to predict traffic up to 30 minutes ahead in time using an autoregressive model with real-time and historical data. The model produces some reasonable predictions of short-term traffic, yet the quality of results is contingent on the amount of data available to train the model in a variety of circumstances (e.g., normal and incident traffic situations). Here, DataFITS gets around the problem using data fusion on different sources to enhance the input dataset for the traffic estimation process. The study of traffic prediction using these methods is a well-known area, and we may come across many comparable approaches [12], [13].

The use of correlated areas to increase data quantity is a rather unique approach, as most literature solely utilizes data from connected roads that are linked with one another. Wang et al. [14] investigated how combining traffic data from different urban areas might be used to improve fine-grained prediction of traffic. Their method grouped several correlating road links into urban regions and predicted both fine- and coarse-grained traffic. However, they are not considering increasing the amount of information by using data from multiple correlated regions or heterogeneous fused data.

The usage of heterogeneous data fusion to provide a traffic prediction model is also discussed in the literature, mostly combining traffic data from stationary sensors and probe vehicles [15], [16]. In [16], the authors provide a prediction model based on heterogeneous data fusion, by combining flow data acquired from cameras and travel time through GPS observations. We extend the current literature, by providing a framework that is capable of fusing various data types, not limited to traffic features, and use this data to create a traffic prediction model. Our methodology is not dependent on existing datasets and is going to allow users to add further data sources, once it is published as an open-source project.

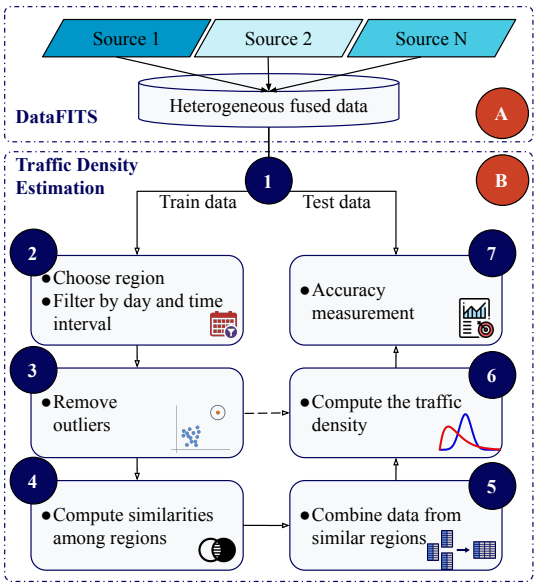

Fig. 1: Traffic density estimation design.

III. Design

In the following, we propose a traffic density estimation framework based on heterogeneous data fused across similar regions, as a solution for the issues raised in the introduction. The goal of this paper is to demonstrate the advantages of heterogeneous data fusion in traffic density measurement.

A. DataFITS

DataFITS, shown in Fig. 1 **(A)**, is a framework implemented in Python and R, capable of collecting, preparing, processing, and analyzing heterogeneous data from transportation systems scenarios to improve the data quality through data fusion. The results of DataFITS can be utilized in a variety of ITS applications, such as traffic estimation.

1) Data Acquisition: DataFITS first stage gathers data using many Application Programming Interfaces (APIs), parsers, and web crawlers based on predefined user criteria such as geographical location, time interval, and predetermined data sources.

2) Data Preparation: In this stage, the data is prepared in a variety of ways. First, all names utilized by each data source are converted into standardized common names to identify the data features. Second, we map distinct values from a certain feature to comparable data types, e.g., descriptive traffic values (normal, increased, congested, etc.) to a numerical representation (0 - 10). We provide an initial mapping in the framework configuration, which can be changed based on the individual user experience. Then, DataFITS transforms the data into a valid internal format for map matching. In addition, a Shapefile (SHP) that contains the road network for the observed region data is acquired.

3) Data Processing: The prepared data is processed, aiming to fuse the heterogeneous data in space and time. The former fusion uses the map matching technique, such as Fast

Map Matching (FMM)[1], where the data from each source is mapped onto the same underlying road network, describing the spatial fusion. Therefore, it can be grouped by unique roads or sets of connected roads, utilizing the information about the traversed path and matched edges, we obtain from the FMM. Furthermore, it conducts a parameter to consider the possible error in the GPS measurement, increasing the data reliability. A temporal fusion can be achieved using aggregations over time (e.g., hour or day) allowing to group the data regarding different temporal aspects. To overcome the issue of differences in timestamps and time granularity between different sources, we apply a floor function on the time value, based on the defined interval in the configuration file, to ensure all sources have the same time value. The data processing step has two outputs: (i) the enriched data in a CSV file, the result of the spatio-temporal fusion, and (ii) a *geojson* file format that enables a pre-visualization of the data by third-party tools such as QGIS.

4) Data Usage: DataFITS processes the CSV file to compute statistics and further visualizations. The spatial coverage increasement after the fusion process can be seen through heat maps, bar plots, and more. For the temporal data fusion, it provides statistics of the information for different time spans (scatter-, line- or correlation-plots). The fused heterogeneous data, outputted by the framework, is used as the input data for the traffic estimation approach described in the next section.

B. Traffic Density Estimation Framework

We describe the methodology of our traffic density estimation framework based on heterogeneous fused data provided by DataFITS, depicted in Fig. 1 **(B)**.

1) The Data Split: To start with the traffic estimation process, we split the available heterogeneous fused database into a training set, used to create a data sample, and a test set, which will be used to evaluate the accuracy of the samples (step 1). Here, we assume that the size of both sets is defined by the user and thus correlated to the amount of data available.

2) Filter the Data: In step 2, we decide on a traffic-estimation region and set the day and time for the observation. First, using data from the map matching process, we establish a list of unique regions. Therefore, we use the *opath* that relates connected road segments on the network to each piece of data input that was matched and group them based on their contained road IDs. To initialize the regions, we take the distinct opath values, with each data entry corresponding to the same region, if more than 50% of the covered roads are identical. This is an initial approach to creating regions and will be improved throughout the framework development, by considering spatial correlation to create regions.

3) Remove Outliers: In the third step, potential outliers in the training data are identified and eliminated. Those erroneous data might bias the sample, lowering the accuracy of the estimate. In this investigation, we define outliers to be observations having only 0-values for traffic and speed, (measurement errors) or record values that are too far apart from the overall regions mean regarding the observed day of the week. Here, we consider as outliers those data observations that lie outside of upper and lower boundaries: total traffic avg $\pm$ standard deviation. This approach may remove a high amount of data in an abnormal traffic scenario and therefore will be changed in the future, regarding our ambition to further include incident information into the prediction model. From this step (3) two directions could be taken based on the evaluation purpose. In one way (steps $3 \rightarrow 4, 5$), we compute similar regions and combine the most correlated ones to increase the amount of data. On the other way (steps $3 \rightarrow 6$), the traffic computation is done without the combinations of data from similar regions.

4) Compute Similarities: The main aspect of this analysis is to identify correlated regions that show a similar traffic behavior (step 4). To define similar regions, we calculate two different metrics, *Pearson Correlation* and Dynamic-Time-Warping [3], between the time series of traffic values $S_i(t)$ for the regions i and j in a correlation matrix (1). Thus, low correlation indicates no dependency on the values between two regions, whereas 1 shows a maximum correlation between them.

$$X_{i,j} = \frac{\sum_{t=1}^{L}(S_i(t) - \bar{S}_i)(S_j(t) - \bar{S}_i))}{\sqrt{\sum_{t=1}^{L-t}(S_i(t) - \bar{S}_i)^2} \cdot \sqrt{\sum_{t=1}^{L-t}(S_j(t) - \bar{S}_j)^2}} \quad (1)$$

To measure the value of distance between two regions we use DTW, which compares the traffic values of two different time series under the assumption that the values are not perfectly synced, but follow a similar pattern. Regarding our context, this metric shows similarities between regions, even if the values are shifted by a low amount of time (e.g., 10 min.).

5) Combine Similar Regions: We use the correlation of different regions to increase the traffic information by grouping their data. Moreover, in case of the absence of historical data in a particular region, we can use this data to improve the traffic density estimation. Using a given threshold for each metric, we can choose the regions that we consider similar based on the following formula. The variables a and b represent the *Pearson Correlation* and DTW respectively, with th_x denoting the threshold of a given metric x as shown in (2):

$$a \geq th_a \wedge b \leq th_b \quad (2)$$

The data from all regions that fulfill this equation is chosen to increase the information for the observed region (step 5).

6) Computing traffic: To create the data sample for estimating the traffic, we are providing a simple prediction algorithm that creates a value for each day of the week in the 10-minute interval. First, we calculate a mean value based on the data from our initial observed region, combining the traffic values from the complete training set for each respective weekday. Next, we compute the average traffic value from the additional available data covering similar regions. Calculating the mean of those two averages results in our final prediction value (step 6). This is an initial version of a prediction algorithm with a

[1]https://github.com/cyang-kth/fmm

low complexity, which is going to be further improved in the future (e.g., adding parameters that control the influence of single values from different data sources).

7) Estimation Accuracy: To evaluate the accuracy of the traffic estimation, we select the test data (the last data collected) from our heterogeneous fused data (step 7) and calculate the respective mean values for each weekday. We compare it to the estimated traffic by using common metrics for data estimation and forecasting such as R2, DTW, GC, MAE and RMSE.

Our proposed methodology is flexible enough to allow many parameters combinations such as different sizes of training/testing datasets; choose which data samples are used throughout the process; other metrics to draw regions similarity or evaluate the results can be plugged in. This leads to many opportunities for different analyses. Here, our focus is on evaluating the benefits of heterogeneous data fusion and grouping data from similar regions regarding the traffic estimation approach.

IV. Evaluation

In this section, we conduct different experiments to quantify the fusion of heterogeneous traffic-related data by the DataFITS and use the fused data to train a model to estimate traffic density.

A. The Raw Dataset

The data acquisition process was set up to collect data of three different categories: traffic, incident, and intra-vehicular. Those data are collected from different commercial providers (HERE maps and BING) or open platforms (OpenData (OD) and Envirocar). HERE and BING provide data combining fixed sensors and probe vehicles, OD uses data from fixed sensors and Envirocar produces probe data. We gathered data from those providers in intervals of 10 minutes for 8 weeks over the city of Bonn in Germany. A non-exhaustive list of collected data are, for example, flow level and speed related to traffic, type of incident, and an identifier of the event are examples of incident-related data. Also, intra-vehicular data are gathered such as speed, fuel consumption, and CO_2 emission. All reported data contain geolocation and timestamp. Combining the different data types can show correlations within the features, especially to analyze traffic patterns concluding from an incident event, and is an aspect we want to further research in the future.

B. The Potential of Data fusion

To analyze the general benefits of heterogeneous data fusion in the context of ITS, we quantify the numbers of road segments covered by each data source, displayed in Table I. It shows the single source exclusive coverage information and the number of overlapping segments that are covered by multiple sources. Therefore, we observe that the sources vary in their coverage, with Traffic HERE holding $20,94\%$ of the unique roads, indicating a higher coverage in comparison to the open data traffic information. Concerning incidents,

TABLE I: Covered road segments by data source.

Source	Total Roads	Unique Roads	Single source exclusive coverage in (%)
Traffic HERE	684	339	20,94%
Traffic OD	581	195	12,04%
Incident HERE	206	53	3,27%
Incident BING	597	256	15,81%
Construction OD	52	31	1,91%
Envirocar	433	178	10,99%
Overlapping	**567**	**567**	**35,02%**
		Total 1619	

BING covers a much higher amount of unique roads, 256 in comparison to 53 reported from HERE and 31 available due to OD. The number of overlapping road segments ($35,02\%$) reveals the potential of using data from multiple sources to enrich the available information on those roads through data fusion. We limit the framework, assuming that the data is reliable without evaluating the quality of sensors. However, this will be considered in future states of the fusion process, changing the influence of a single source to the fused data. Overall, comparing the 684 roads reported by HERE, which is the data source with the highest quantity and coverage data, to the 1619 unique roads covered after the fusion process, the ITS was able to increase by 137% on the data coverage.

C. Traffic Density Estimation

To demonstrate the benefits of fused data on our traffic density estimation, we draw comparisons between estimations made by: i) Fused Data with Grouped Regions (Fused GR) that consider all steps described at Sec. III; ii) Raw Data from a single source (RAW)-{HERE or OD} and iii) Fused Data from a Single Region (Fused SR) that uses the fused data from DataFITS but skips steps 4 and 5 of Fig. 1 **(B)**. The following comparisons are made to show that our methodology can improve the traffic density estimation, by showing the benefits of using Fused GR instead of RAW.

The fused data (step 1 of Fig. 1) was split into 75% of the dataset for training and 25% for testing the model. The training data encompasses a time frame of 6 weeks (28.06.2021 – 08.08.2021) with the testing data two weeks ahead (09.08.2021 – 22.08.2021). The goal is to estimate the traffic density for working days (Monday to Friday) in rush hours (14h to 18h).

We carried out the steps presented in Sec. III over the fused data, like removing data points containing only zero values or outliers. To create the Fused GR dataset, we assume that similar regions have Pearson Correlation $\geq 0,85$ and $DTW \leq 0,3$. By selecting all similar regions, we create a traffic density data sample, calculating the mean traffic level and comparing it against the test data.

We used five different metrics to measure the traffic estimation accuracy, where: i) Coefficient of determination (R2) express the proportion of variance, mainly used in the context of regression models. Here, it is used to evaluate the density estimation quality. Values close to 1 indicate a perfect fit, while negative values indicate a model that fits worse than

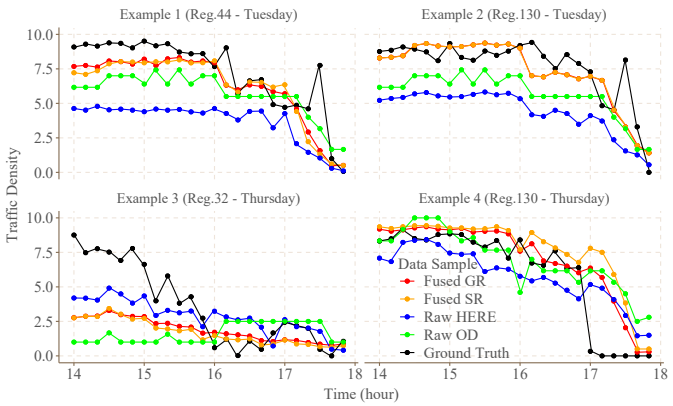

Fig. 2: 4 examples for traffic estimation.

a horizontal straight line, indicating a bad model; ii) The Dynamic-Time-Warping (DTW) is a similarity metric between time series. We use DTW to compare the time series of traffic density estimation and the testing data. Low DTW values close to 0 indicate high similarity between the series; iii) Granger Causality (GC) is a statistical test that indicates if one time series is useful to forecast another. Here, we consider that if the GC value is below the threshold of 0.05, it indicates a useful traffic density estimation; iv) Mean absolute error (MAE) sums up the average difference of each estimated value and the actual observation, its values vary from 0 to ∞. The lower the value is, the better the estimation result; v) The Root-Mean-Square Error (RMSE) is the square root of the standard deviation from all residuals. It's similar to the MAE but adds more weight on larger errors.

Fig. 2 shows the traffic density estimations and the ground truth (i.e., the testing data) curves for a subset of examples (due to limited space). The traffic density values vary from 0 to 10 (non-traffic to jammed traffic) and the displayed data is gathered from different highways in 2 days (Tuesday and Thursday). Table II complements the visual information by tabulating, for the same regions, all estimation accuracy metrics, highlighting the best results in bold. Taking a look at Example 1 of Fig. 2, the traffic density estimation using Fused GR (red line) and Fused SR (orange line) present values close to the ground truth (black line). The samples are nearly equal, but Table II is showing that Fused GR is superior in most of the tested metrics. This indicates that using data from correlating regions slightly improves the traffic estimation in this case. Moreover, the benefits of heterogeneous data fusion (red and orange line) suggest better accuracy compared with the samples using raw data (blue and green).

Similarly, Example 2 in Fig. 2 shows that the fused data achieves the best estimation result with Fused GR and Fused SR being the same values, due to no other similar region available. A R2 value of 0,59 indicates a good fit of the model, supported by the low DTW value of 0,44 and small error values. The RAW samples are worse, indicating that just using the process of fusion and outlier removal without further enrichment through data from similar regions, can

TABLE II: Accuracy metrics for traffic density estimations. A lower value indicates a better result (bold), except by R2.

Ex.	Metric	Fused GR	RAW HERE	RAW OD	Fused SR
1	R2	**0,58**	-1,22	0,3	0,5
	DTW	0,75	1	1,23	**0,74**
	GC	**0**	0,13	**0**	**0**
	MAE	**1,2**	3,29	1,87	1,34
	RMSE	**1,69**	3,69	2,18	1,85
2	R2	**0,59**	-1,49	-0,14	**0,59**
	DTW	**0,44**	1,77	1,14	**0,44**
	GC	**0**	**0**	**0**	**0**
	MAE	**1,01**	3,26	2,15	**1,01**
	RMSE	**1,41**	3,48	2,36	**1,41**
3	R2	0,08	**0,44**	-0,81	0,04
	DTW	1,16	**0,74**	1,45	1,15
	GC	0,25	0,98	0,53	**0,09**
	MAE	2,15	**1,77**	3,03	2,19
	RMSE	2,78	**2,17**	3,89	2,84
4	R2	**0,66**	0,6	0,46	0,27
	DTW	0,36	0,77	0,76	**0,33**
	GC	0,15	0,06	**0,05**	0,24
	MAE	**1,29**	1,82	1,72	1,78
	RMSE	**2,05**	2,21	2,58	2,77
	Legend	Fused Data Group reg.	Raw Data		Fused Data Single reg.

provide huge benefits to the accuracy of the estimation. The R2 measurements are negative indicating a worse fit of the model, complemented by all other metrics, like a DTW of 1,77 for RAW-HERE.

The results presented in Example 3 demonstrate an opposite case, where the data fusion leads to a worse estimation compared to using RAW-HERE. The plot shows that data samples from RAW-OD is very coarse and presents biases when compared to the ground truth. Fusing the data from OD and HERE increases the estimation accuracy, but is still lower compared to using only data from RAW-HERE, shown through all metrics except GC. A strong benefit is measured especially through the R2 metric of 0,44 and a DTW of 0,74. We observe, that using data from correlating regions slightly improves the estimation, especially considering the R2 (0,08 to 0,04), MAE (2,15 to 2,19) and RMSE (2,78 to 2,84). This example indicates the problem of fusing fine-grain and coarse-grain data together leading to imprecise results. However, comparing the fused samples to RAW-OD (coarse-grain data), the fused data shows a significant traffic estimation accuracy improvement.

The fourth Example shows that all estimated values are close with Fused GR achieving the overall best result. Fusing the data of both sources and one more correlating region improves the metrics R2 about of 0,66, MAE about of 1,29 and RMSE about of 2,05. Considering Fused SR, it shows the best DTW measurement about of 0,33. Overall both fused samples have a better accuracy compared to RAW-OD and especially RAW-HERE, furthermore revealing the benefits of our proposed approach.

D. Overall Accuracy

To compare the overall accuracy of our traffic estimation based on data fusion, we expand the number of regions considered to 15 overall observed working days. Table III

TABLE III: The overall accuracy of our traffic estimation.

Data Sample	R2	DTW	GC	MAE	RMSE
Fused GR	**-0,11**	**0,53**	**0,23**	**1,19**	**1,50**
RAW-HERE	-0,25	0,58	**0,23**	1,31	1,58
RAW-OD	-1,36	0,81	0,38	1,67	2,06
Fused SR	-0,14	**0,53**	0,25	1,21	1,52

shows the average for each metric and highlights the best results in bold.

By comparing the use of Fused GR (row 1) against RAW-{HERE or OD} (rows 2 and 3) to estimate the traffic density, the overall best results are achieved using Fused GR. We see that for all averaged metric values our approach achieves the best results. In the case of GC, RAW-HERE scores the same average value, however, the improvements made by Fused GR compared to RAW-HERE in all other metrics are significant. For DTW and MAE we have an improvement of 9%, followed by a 5% better performance regarding RMSE. In comparison to RAW-OD, the benefits of heterogeneous data fusion are even visible due to an accuracy improvement of up to 40% for GC and 35% in DTW. The average error metrics could be improved by over 25%.

Comparing the average metric values of Fused GR and Fused SR we see that they are quite similar, scoring the same average values for the DTW. On all other metrics the Fused GR achieves better results, improving the metrics up to 8% for GC and 1% regarding the MAE and RMSE. This indicates that the usage of correlating regions leads to a minor performance increase. In summary, we were able to quantify the benefits of a heterogeneous data fusion together with the grouping of similar/correlated regions, especially compared to only information from one data source (RAW-OD). For all metrics, we see better results on average, increasing the accuracy of our traffic estimation.

V. Conclusion

In this paper, we proposed a traffic density estimation approach using heterogeneous fused data through the DataFITS framework which increases and enrich the available data, in a given urban area, by fusing them. According to the number of unique reported road segments in Bonn, DataFITS was able to increase by a factor of more than two times. Moreover, 35% of these road segments overlap, meaning that we can enrich the information by fusing the data from multiple sources.

With the fused data, our traffic estimation was able to consider more road segments and identify similar regions used to improve the estimation accuracy. The traffic estimation was tested against using a single data source and fused data from a single region. The fused data led to better accuracy in traffic estimation, especially compared to using only data provided by a single source. Fusing the more coarse data reported from OD with the detailed data from HERE benefits the traffic estimation, leading to way better estimations in certain cases and a performance increase of up to 9% on average. Comparing Fused SR with Fused GR, the average metric results are quite similar, slightly better with the enriched information. However, using data from correlated regions improved the overall accuracy for most examples. This indicates that enriching the data with information from similar regions is beneficial in some cases, and results in a better performance on average, compared to only using the fused data.

As future work, we plan to add vehicular data allowing the analysis of driving behavior, emission, and fuel consumption in traffic areas and add more cities to the data acquisition. Furthermore, we plan to design a methodology to include the incident information to the traffic prediction model, allowing the correlation between them.

References

[1] Rettore, P. H. L., G. Maia, L. A. Villas, and A. A. F. Loureiro, "Vehicular Data Space: The Data Point of View," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2392–2418, thirdquarter 2019.

[2] Rettore, P. H. L., B. P. Santos, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "Towards intra-vehicular sensor data fusion," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2016, pp. 126–131.

[3] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, USA:, 1994, pp. 359–370.

[4] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[5] G. Vítor, P. Rito, and S. Sargento, "Smart city data platform for real-time processing and data sharing," in *2021 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2021, pp. 1–7.

[6] Rettore, P. H. L., A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "A method of eco-driving based on intra-vehicular sensor data," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, July 2017, pp. 1122–1127.

[7] A. B. Campolina, Rettore, P. H. L., M. D. V. Machado, and A. A. F. Loureiro, "On the Design of Vehicular Virtual Sensors," in *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, Ottawa, Canada, June 2017, pp. 134–141.

[8] Rettore, P. H. L., A. B. Campolina, A. Souza, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Driver Authentication in VANETs based on Intra-Vehicular Sensor Data," in *2018 IEEE Symposium on Computers and Communications (ISCC)*, June 2018, pp. 00 078–00 083.

[9] Rettore, Paulo H. L. , B. Pereira, R. Rigolin F. Lopes, G. Maia, L. Villas, and A. Loureiro, "Road Data Enrichment Framework based on Heterogeneous Data Fusion for ITS," *IEEE Transactions on Intelligent Transportation Systems*, 01 2020.

[10] Rettore, P. H. L., R. Rigolin F. Lopes, G. Maia, L. Aparecido Villas, and A. A. Ferreira Loureiro, "Towards a Traffic Data Enrichment Sensor Based on Heterogeneous Data Fusion for ITS," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, May 2019, pp. 570–577.

[11] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 2, pp. 653–662, 2014.

[12] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848–3858, 2019.

[13] C. Zheng, X. Fan, C. Wang, and J. Qi, "Gman: A graph multi-attention network for traffic prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1234–1241.

[14] S. Wang, M. Zhang, H. Miao, Z. Peng, and P. S. Yu, "Multivariate Correlation-aware Spatio-temporal Graph Convolutional Networks for Multi-scale Traffic Prediction," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 3, pp. 1–22, 2022.

[15] W. Jiang and J. Luo, "Big data for traffic estimation and prediction: a survey of data and tools," *Applied System Innovation*, vol. 5, no. 1, p. 23, 2022.

[16] R. A. Anand, L. Vanajakshi, and S. C. Subramanian, "Traffic density estimation under heterogeneous traffic conditions using data fusion," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2011, pp. 31–36.