From GPS Traces to Individual Emission Exposure: A Data-Driven Four-Step Process

Gurban Aliyev[1,2(✉)] and Mirco Nanni[2]

[1] University of Pisa, Pisa, Italy
gurban.aliyev@phd.unipi.it
[2] ISTI-CNR, Pisa, Italy
mirco.nanni@isti.cnr.it

Abstract. Vehicular traffic is one of the major sources of air pollution in urban settings, making it essential to clearly understand how much and where vehicle emissions impact residents. Estimating vehicular pollution using GPS trajectories and microscopic models is getting more popular as this method has several advantages compared to other approaches. However, GPS data sources usually cover only a small sample of actual traffic, making current approaches unable to provide emission estimates for the whole road network. Moreover, to understand how much of these emissions reach different locations, a dispersion model should be applied, and quantifying their effect on individuals requires considering where they stay and/or how they move. Therefore, in this paper, we propose a four-step process that elaborates on raw, incomplete emission estimates and (i) first, estimates initial emissions from GPS data, (ii) estimates emission concentrations for the missing road segments, (iii) further processes the emission data to consider air dispersion, and (iv) computes the expected exposure to emissions of individuals in several use cases, involving both public buildings (e.g. schools) and pedestrian mobility. The experiments are based on a sample of vehicular GPS data in two Italian cities.

Keywords: road networks · vehicular emissions · missing data imputation · emission dispersion · emission exposure

1 Introduction

Ambient air pollution is one of the main barriers to the sustainable development of urban areas, which are expanding fast recently [22]. Air pollution is a more serious concern in cities than in rural areas as cities are densely populated. Also, primary sources of anthropogenic emissions, such as energy production and transportation, are concentrated around urban clusters. As a result, the concentration of air pollutants causes low air quality in cities, with more and

A. Kocian et al. (Eds.): INTSYS 2024, LNICST 608, pp. 64–82, 2025.
https://doi.org/10.1007/978-3-031-86370-7_5

more people exposed to air pollution every year. Because of the effects of this problem on public health and the economy, the United Nations calls to reduce the adverse per capita environmental impact of cities [1].

Our focus is on the quantification of vehicular emissions and mitigation of these emissions, as with the number of cars in cities increasing, the total amount of vehicular pollutants rises [12]. Studies using GPS traces offer the best trade-off between highly-detailed human mobility and representative vehicle fleets to estimate car emissions [2]. Recent studies [2,7,23] assess spatial and temporal distributions of vehicular emissions using vehicular trajectories. These studies help to quantify the impact of existing green mobility policies [23], next-generation routing principles for vehicles [7], or electrification of gross polluters [2] in terms of the decrease in emissions.

However, the GPS-data-based approach has some gaps: firstly, most of the currently available trajectory datasets cover only a portion of all vehicles and roads in the road network, while – to have a complete view of spatial and temporal emission patterns – network management requires an inventory of vehicular emissions that represent the whole population. Secondly, works focused on emissions mitigation policies only consider changes in total emissions and ignore the average exposure levels of the population. For example, the air quality in some parts of a city can get worse while total vehicular emissions decrease, thus calling for assessment methods at a more granular scale. Another issue is that most previous studies based on vehicle data assume that emissions generated on one road segment will not affect other segments, or areas outside streets, yielding potentially skewed results. For instance, secondary roads running parallel to large ones (a common setting around highway junctions) might receive emission levels from the other roads comparable to or higher than the endogenous ones. In addition, dispersion estimates beyond the road network are essential to evaluate the actual exposure of people staying in buildings or walking/cycling through the city, as both modes of transportation can go where cars cannot.

In this work, we tackle the main open issues discussed above, providing the following contributions:

- we introduce a 4-step process that covers all required phases of a (big) mobility data-based approach to estimate individual exposure to vehicle emissions;
- we provide a validation of the raw emission estimates against an indirect ground truth measured from the air quality monitoring network in Rome;
- we introduce the emission imputation problem, aimed to infer the emissions on road segments not covered by the input mobility data. Since there is no literature directly addressing the specific problem, we explore existing solutions for general data imputation over graphs and networks, adapting several approaches to our context and identifying the best candidate;
- we implement a standard emission dispersion model efficiently, enabling the computation of a city-scale map of resulting emissions at a fine granularity;
- we introduce a simple approach to individual exposure estimation for static and dynamic settings, presenting one use case for each;
- finally, all steps are applied and evaluated on a dataset of real vehicular GPS data traces covering the city of Pisa, Tuscany (Italy).

In the rest of the paper, we discuss the relevant literature (Sect. 2), introduce our 4-step pipeline describing all its components (Sect. 3), show experimental results assessing performances of imputation and showcase the pipeline on two use cases (Sect. 4), and finally provide conclusive remarks (Sect. 5).

2 Related Work

In this section, we discuss existing approaches related to the three main technical problems involved in the paper: estimating emissions at the road level from mobility traces, inferring emissions on road segments where the information is not available, and measuring the exposure of places/people taking into account actual dispersion of emissions.

2.1 Emissions Inference

Several works in the literature can estimate vehicular emissions by exploiting the spatial precision and relative abundance of mobility data coming from vehicle fleets [23]. Researchers of the last decade [2,22] have proved the usefulness of GPS sensors paired with microscopic emission models to track vehicles' mobility in real-time mode and calculate vehicular emissions. An important parameter used in microscopic models to estimate emissions is the emission factor (EF), which is a functional relation shown as pollutant emitted per kilometer traveled or liter of fuel consumed [12]. EFs are useful in the sense that they can project deviation of speed and acceleration in a small temporal scale and increase the accuracy of emission estimate [10].

However, it is worth mentioning that developed emissions models have been suffering from the lack of validation/calibration methods, and that it is not possible to calculate on-road emissions on many road segments as most of the GPS data sets do not cover the whole road network. Some inference/imputation methods need to be implemented in applications where data scarcity is an issue.

2.2 Validation of Emission Estimates

Methods to quantify vehicular air pollution in cities are increasing with the development of technology and changes in municipal air quality management (AQM). Several studies focus on the validation of on-road estimations by using on-road measurements as ground truth data [9,15,26]. This way is accurate, but not convenient, as on-road measurements are not easily available. Meanwhile, satellite and air quality station data are more accessible.

Remote sensing (including satellites) is a useful method to evaluate measurement results of emission models and provide complementary spatial information [24]. Satellite data is represented by maps and is more useful for large-scale analysis. Meanwhile, an air quality station measures the air quality of a neighborhood it is located in (of a certain radius). Given a network of stations scattered across the city, it is possible to analyze urban-level emission levels.

In terms of measurement accuracy, AQM stations are more advantageous than satellites as they are located close to the car emission sources (roads). Satellites show air quality by returning a value of the lowermost 15 km of the atmosphere ('column amounts'), and they are the least sensitive close to the surface, which is the most important part of calculating emissions [24].

The common problem of remote sensing measurements is that they are not weather-proof [13,24] as interference of high humidity or other gases can be an obstacle for sensors. Also, remote sensors measure amounts from different sources (transportation, manufacturing, etc.). Nevertheless, vehicles are the only source of some pollutants (such as NO_x) in non-industrial areas of cities. Based on such pollutants, estimated road-link emissions can be linked to station measurements by comparing relative changes, instead of focusing on absolute values.

2.3 Imputation Strategies

Because of the limited amount of data collected, the problem of emission data sparsity on urban roads is common. However, according to recent surveys on traffic-related data imputation [5], there are currently no specific methods to impute missing emission values, while there are various attempts using state-of-the-art [6,14,20,27] approaches to estimate other road features.

Some studies [20,27] claim that missing traffic data imputation methods should be task-specific to allow the model to learn features better. Among task-specific models, those that are built to estimate intensities of traffic flows [20,27] can be useful to predict vehicular emissions as traffic flows and vehicular emissions are intuitively correlated. Meanwhile, other studies [6,14] are task-agnostic and their authors claim that it is possible to create feature embeddings of the input data to use them in any task after fine-tuning [6].

Another classification of existing models is based on a graph representation of road networks. The most natural representation of road networks, where road segments are links and crossroads are nodes, is called a *primal graph*: here, node-related knowledge (location, network metrics, number of various POIs around, aggregated characteristics of road segments linking to the node, etc.) is brought in feature representation. Conversion of original node features into a vector representation is called node embedding. Earlier studies [20,27] rely on node embedding, while more recent works [6,14] adopt a *dual graph* representation of road networks, where road segments are transformed into nodes, while crossroads between road segments become links connecting nodes.

2.4 Dispersion and Exposure Models

Studying individual and population exposure to air pollution is important to quantify the effect of changing travel behavior and traffic policies. Initial studies [25] focus only on the assessment of exposure in static locations (workplace, school, and home) using population census data. Later studies [8,21] use more

accurate mobile data to quantify average mobility exposure. More recent studies investigate the impact of travel behavior on traffic-induced emissions and subsequent exposure to traffic emissions [18].

However, these studies have some limitations. Firstly, these works disregard the dispersion of vehicular emissions both vertically and horizontally. There are various dispersion models to be considered, and the main approaches used in these models include Computational Fluid Dynamics (CFD), Gaussian, Lagrangian, and Box models [17,19].

Depending on the dispersion levels, exposure to car emissions on roads and around roads can change. The concentration of on-road vehicular emissions in certain places can increase or decrease depending on different factors, such as weather conditions (temperature, precipitation, etc.) and building profile data. Depending on these factors, one approach or the other can be more appropriate, eventually yielding more accurate results [19].

3 Methodology

In this work, we propose a 4-step pipeline[1] covering all the phases that allow to conversion of raw input GPS traces of (a subset of) vehicles circulating in a city into exposure estimates for individuals residing or moving in the area. The overall process is depicted in Fig. 1, and consists of an emissions estimate for roads covered by the input GPS data, emissions imputation for the remaining ones, modeling dispersion to distribute them on the territory, and finally an exposure estimate of static or moving entities. In the rest of this section, we illustrate the four steps in detail.

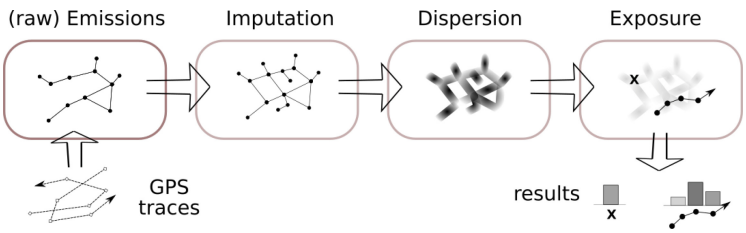


Fig. 1. Overview of the proposed process: estimating initial emissions from GPS data; imputing uncovered roads; simulating dispersion of emissions; computing exposure for static and moving entities.

[1] The code that implements the pipeline is available at https://github.com/baygaliyev/eide.

3.1 Raw Emission Estimates

Our work adopts the approach introduced in [2], which makes use of GPS traces of vehicles moving in a city to estimate their contribution to the emissions over the road segments traversed during their trips. Following the microscopic emissions model in [22], instantaneous emissions associated with each trajectory point p are computed through equation:

$$E_p^j = f_1^j + f_2^j v_p + f_3^j v_p^2 + f_4^j a_p + f_5^j a_p^2 + f_6^j v_p a_p \quad (1)$$

where j denotes the pollutant type (in our case $j = CO_2$), and v_p and a_p are the instantaneous speed and acceleration of the vehicle. Emission factors f_i are determined by the type of pollutant and the engine type (petrol, diesel, etc.).

Single point estimates are then aggregated at the level of the road segment to obtain its average emission concentration. Here, the road segmentation provided by OpenStreetMap (OSM)[2] is adopted, assigning each emission point to the closest segment. In Sect. 4.1 we will see – through the comparison against air quality stations data – that this step yields reasonable local estimates. For each road segment, we divided the CO_2 amount assigned by the length of the road segment to normalize pollution amounts (as longer roads will probably have higher CO_2 pollution). As discussed in later sections, the coverage of GPS data can be incomplete, since they usually describe only a sample of the circulating pool of vehicles and the sampling rate is sometimes low (e.g. in our experiments it is around 1 point per minute), thus smaller or lower-traffic road segments might be particularly ill-represented. This calls for the next stage of our process, trying to reconstruct the expected emission concentration in missed segments.

3.2 Imputation

For the missing data imputation (MDI) step we considered all available graph imputation approaches in the literature, selecting those that could fit our use case (though none of them natively addressed emissions imputation) and identifying a subset of representative solutions. The selection also considered the availability of an open-source implementation of the methods.

The simplest approach to graph imputation consists of assigning a fixed and predefined set of features to each entity (node or edge), which encodes into a standard vector representation all its characteristics and relations to other entities deemed relevant. Then, such representation becomes the input for standard prediction models (regression or classification models, depending on the task).

The most sophisticated models for graph imputation are based on computing some intermediate embedding vector representation of the road network, focusing either on nodes embedding or on edges embedding, and applying either generic graph embedding [6,20] or deep graph learning methods, in particular graph convolutional networks [14,27]. In particular, methods that produce edge embeddings typically represent the road network through a dual graph, where

[2] https://www.openstreetmap.org/.

nodes represent road segments, thus allowing to adoption of standard node-oriented embedding strategies. As in the basic case of fixed node/edge features, the resulting vectors yielded by generic embedding methods are used as input data for task-specific prediction models. On the contrary, graph convolution methods integrate both steps (embedding and prediction task) in the learning.

Among the existing embedding and deep learning approaches we decided to use two methods: SARN [6], a task-agnostic graph embedding model, which was designed to create embeddings of road segments; and the task-specific convolutional SI-GCN [27], which works with nodes and was originally developed to impute traffic flows – intuitively well correlated with emissions.

Graph Features. Besides being fundamental for the basic fixed-features approach mentioned above, also SARN and SI-GCN can take advantage of features describing, respectively, edges and nodes in the input graph. This information is expected to be extremely important, thus we included some characteristics derived from the road network:

- node features can be grouped in 4 sets with a total of 15 features: i) node coordinates (*latitude* and *longitude*), ii) network measures (*node degree*, *betweenness*, *closeness* and *harmonic centralities*), iii) the number of points of interest (POI) around the node, grouped into 6 categories (*cultural*, *education*, *food*, *health*, *service* and *transportation*), and iv) *the number of road segments connected to the node which have emission data*, aggregations of emissions of the road segments connected to the node (*mean value*, *standard deviation* of the mean value, *minimum* and *maximum* values).
- edges have the same features with the additions of 4 coordinate features (both of start and end nodes) and 3 features related to roads' physical characteristics (*length*, *radian*, *highway class*), making 20 road features in total.

Predictive Models. While SI-GCN directly provides an imputation model that can be applied to road segments with missing target values, the fixed-features approach and SARN's task-agnostic road embeddings are used as input for standard regression models. In particular, we adopt some baseline and state-of-art ones: a basic linear regressor; XGBoost; a support vector machine regressor (SVM); and a multi-layer perceptron (MLP).

A Classification-Regression Approach. An inspection of the different models' behaviors revealed that they typically stringy overestimate low emission amounts and underestimate higher CO_2 amounts. To address this issue, we propose a CO_2 imputation consisting of a two-phase process where we first train a classifier to discriminate highly polluted roads from the others, and then we build a regressor specific for high-emission segments and another one for the others. Notice that the distinction between high and low values depends on a threshold value to be decided during the training phase. Predictions are then performed by

applying the classifier to decide which regressor to use, whose output is returned as the final result.

3.3 Dispersion of Emissions

Dispersion models for air pollutants have been studied for many years, yielding several approaches of different complexity that can capture various factors that impact dispersion, such as wind speed, wind turbulent fluctuations, temperature, etc. In this work we selected a basic Gaussian model, in particular, a simplification of the standard Gaussian plume model proposed in [3] and used in several recent works (e.g. [11], which also provides implementations), computing the pollutant concentration in a given position (x, y) relative to the emission location and at height z, and wind direction along the x axis:

$$C\,(x, y, z) = \frac{Q}{2\pi u \sigma_y \sigma_z} \cdot e^{-\frac{y^2}{2\sigma_y^2}} \cdot \left[e^{-\frac{(z-H_e)^2}{2\sigma_z^2}} + e^{-\frac{(z+H_e)^2}{2\sigma_z^2}} \right]$$

Besides considering the emission rate Q of the point source and the dispersion coefficients (lateral σ_y and vertical σ_z), the formula can account for wind scalar speed u and its direction (implicit in the x-axis alignment). Yet in our setting (based on long-term aggregates) that is not useful, thus we fixed u and assumed homogeneous wind in all directions.

For each emission source, we estimate its impact on the surrounding area. Each road segment is divided into small chunks (default length is 1 m) which are considered as punctiform sources emitting at a constant rate. Their emission rate is thus computed by normalizing the road raw emissions concentration by the road length and the time interval covered by the dataset:

$$e(c) = \frac{emission(r)}{len(r) \cdot T}$$

for each chunk c of road segment r, where T is the dataset time horizon in seconds (in our experiments, equal to one year). To compute the dispersion of chunk's emissions, we partition the region of analysis into a regular grid of cells having a small size (e.g. in our experiments the default is 5×5 m) and for each cell, we sum up the contributions from all sources.

Efficiency Aspects. The computational complexity of the approach mentioned above is potentially very large. Indeed, it requires applying the dispersion formula for each pair $(chunk, cell)$, whose number is proportional to the total road network length and the surface area of the territory under analysis. In most cases that means a cubic growth w.r.t. the territory size (e.g. the diameter). However, without considering the effects of winds and other location-dependent factors, the dispersion patterns become symmetrical in all directions and with the same shape for all sources, simply rescaled by the emission rate Q specific to the emission location. Moreover, the dispersion patterns vanish beyond large

distances (e.g. in our experimental setting, a reasonable threshold appears to be around 250 m), thus a basic dispersion pattern can be computed once for all within a fixed limited spatial range and then applied to each emission source through rescaling. That reduces the computational complexity from $O(n^3)$ to $O(kn^2)$, where n is the territory size (diameter) expressed as several cells, and k is the size of the dispersion pattern discussed above, where typically $k \ll n$. Moreover, computational "constants" are also smaller, since single operations are much cheaper and computable through vectorization.

3.4 Exposure Estimation

The first three steps of the process described above yield a (currently static) mapping of vehicle emission concentrations in all locations within the study area. This allows us to study how much citizens are exposed to emissions in at least two ways, that we will explore and next test in the experiment section: a static location-based one, and a dynamic movement one.

Static location-based studies (e.g., [25]) are the simplest ones, and they consist of comparing the estimated emissions that reach a set of locations, for instance, hospitals, schools, etc., to rank existing places in terms of potential health risk or to identify best candidates for new buildings.

Dynamic movement-based studies (e.g., [18]) consider the movement of individuals and the expected amount of emissions that they absorb while moving. Mathematically speaking, they can be estimated as the integral of the emissions the individual is exposed to in each of the points they visit along their movement trajectory. From a more computational viewpoint, that is obtained by approximating the trajectory T with the sequence S of cells visited and, for each cell $s \in S$, its traversal time $time(s)$, exploiting our cell-based approximation of the emissions' dispersion:

$$exposure(T) = \int_{\bar{x} \in T} e(\bar{x}) d\bar{x} \quad \simeq \quad \sum_{s \in S} e(s) \cdot time(s)$$

Our reference application is estimating exposure for pedestrians in a simplified scenario where the movement speed v_p of pedestrians and their breath intensity are constant – typically true for small and flat cities, while complex scenarios involving hills, stairs, traffic lights, etc. might require a different setting. The first assumption allows us to estimate $time(s)$ as $length(T \cap s)/v_p$, i.e. it is proportional to the length of the trajectory segment contained in the cell. The second one says that the ratio of emissions absorbed by the individual is constant and the overall absorption depends only on the concentration.

4 Experimental Results

In this section, we present experimental results over the four steps of our pipeline, namely raw emission generation, imputation, dispersion, and exposure, each

described in a separate subsection. In particular, the first subsection will provide a comparison of raw estimates against air quality data; the second will evaluate the accuracy of various imputation methods; the third will describe the output of the dispersion phase, also showing the impact of imputation on the result; the last one will present two use cases adopting our exposure strategy. Experiments are based on two GPS datasets of private vehicles: one in the area of Rome, Italy (used for raw estimate validation), and one in Pisa, Italy, similar to the one used in [2] and covering respectively ca. 39000 and 11000 vehicles over one year.

4.1 Validation of Model Estimations

The air quality monitoring network of Rome has historical data on several pollutants measured hourly from 1999 (provided by ARPA Lazio[3]). Inside the borders of the municipality of Rome, there are 11 stations (Fig. 2(left)) where traffic is a significant source of emissions. We obtained (mean) daily NO_x concentrations of those stations during the year 2017 (as our emissions estimates are for 2017).

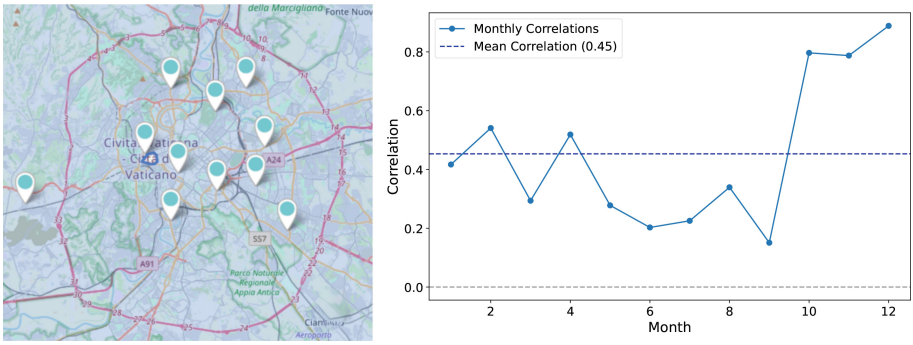


Fig. 2. (left) The air quality monitoring network in the municipality of Rome; (right) Monthly Correlation between CO_2 estimation and NO_x measurements. Correlation values are represented with Pearson's correlation coefficients.

Our validation is based on the comparison between the concentrations measured by the monitoring stations and the emission estimations of our model. More formally, for each station s, day d and week w, we have: i) mean concentrations C_s^d measured by the station (in μ/g^3); ii) aggregated (summed) emissions E_s^w computed by our model (in g). E_s^w is the sum of all the emissions estimated on road segments within a radius r from the monitoring station s during day d. To increase the sample size of cars passing by the stations, we aggregate (summed) road-level emissions by month (E_s^m). Likewise, we aggregate daily station measurements into mean concentration amounts for each month (C_s^m). As a result, we have 11 pairs of estimation-measurement values for each month.

[3] https://www.arpalazio.net/main/aria/sci/basedati/chimici/chimici.php.

Since CO_2 emissions are not directly measured by the stations, our estimates are validated against aggregated NO_x values of station measurements. We can conduct this comparison as NO_x and CO_2 are reported in the literature to correlate significantly (especially among gasoline cars) [4]. The radius of 200 m (according to the guidelines of the European Environment Agency [16]) is chosen to aggregate estimated emissions around each air quality station.

Correlation results are shown in Fig. 2(right). The correlation was computed for each month as Pearson's coefficient between our CO_2 estimates over the 11 stations and the corresponding 11 measurements of NO_x. The mean Pearson's coefficient is 0.45, indicating a significant overall correlation. This result exceeds that of a previous study [28], which compared simpler model estimations – proxy emission scores based solely on traffic volume and roadway type – to air quality station data. A seasonal pattern is evident, with the correlation dropping in the warm season (reaching 0.15 in September). The reason for this is an increased number of warm, sunny, and rainy days which leads to more intensive reactions between NO_x and O_3 and thus lower levels of NO_x in the air, heavily affecting correlations with CO_2 emissions. On the opposite, as seen in previous studies [28], the correlation strengthens towards the end of the year, peaking at 0.88 in December, likely due to the accumulation of emitted NO_x gases in colder, more stable atmospheric conditions.

4.2 Accuracy of Emission Data Imputation

As discussed in previous sections, we evaluated the capability of several approaches to infer the emission concentration of road segments that miss it in the input data. In particular, we evaluate the mean average percentage error (MAPE) yielded by SI-GCN and by several standard regression tools applied both to the raw features of road segments and to an embedding computed by the SARN approach. Also, the classification-regression schema proposed in Sect. 3.2 is applied to all basic regressors. In this case, a manual exploration of the training set identified 4.2 g/m as the best threshold to discriminate between the two classes, marking around 85% as low-emission roads.

Table 1 summarizes the results obtained over 10 runs. For classification-regression, we show both the overall MAPE for each configuration (all) and separately the results for test instances classified as low (L) or high (H). For each algorithm and feature subset pair, the results correspond to the parameter configuration that minimizes the MAPE. The first clear observation is the generally poor performances of all models, with MAPE $> 100\%$, highlighting the difficulty of the task as compared to other similar ones treated in literature. Second, performances are not improved by adopting segment embeddings in place of raw features, nor by SI-GCN, signifying that the structural information they can bring is not relevant for emission imputation. Finally, the best approach for overall results is XGBoost within our classification-regression schema, applied over raw features. In particular, we can see that the classification-regression improves XGBoost's MAPE values by 15%–69% over the basic approach – especially for high-emission segments.

Table 1. Imputation results of MDI models over raw features and SARN embeddings. Values represent average MAPE and its standard deviation (in the brackets) over 10 runs.

Model	Inputs	
	Raw Features	SARN Embeddings
SI-GCN	793.40% (–)	–
Direct Regressors:		
Linear Regressor	215.26% (11.85%)	255.51% (16.80%)
XGBoost Regressor	143.25% (5.34%)	185.17% (11.60%)
SVM Regressor	255.57% (7.13%)	217.51% (24.12%)
MLP	198.65% (0.74%)	300.77% (50.82%)
Classification-Regression:		
Linear Regressor	168.91% (all) 173.05% (L) 128.91% (H)	216.19% (all) 198.85% (L) 383.43% (H)
XGBoost Regressor	**129.88% (all)** 130.50% (L) 124.46% (H)	160.85% (all) 156.64% (L) 207.72% (H)
SVM Regressor	186.76% (all) 194.53% (L) 111.60% (H)	181.32% (all) 181.21% (L) 183.42% (H)
MLP	256.41% (all) 267.33% (L) 150.11% (H)	262.89% (all) 268.60% (L) 201.80% (H)

Error Distribution over Road Types. After obtaining final data imputation results, we analyzed MAPE values considering the most important characteristics of road segments, namely the length of the road segment length and the speed limit on it. Prediction errors of CO_2/m grouped by road lengths are reported in Table 2. We can see that the model tends to have larger errors for longer roads with low pollution per meter. The opposite can be noticed for roads with high pollution per meter, where the MAPE across the longest roads (top 20%) goes below 43%. This result supports the intuition that longer segments generally belong to more important and traffic-intensive roads, and vice versa, thus the model makes more mistakes on roads that violate this trend.

We also aggregated errors by the speed limit of roads (Table 3). Results are consistent with Table 2. Indeed, prediction errors are lower (MAPE below 100%) when a road has a low speed limit and low pollution, as shorter roads tend to be located in residential areas and have lower speed limits. The same applies to roads with higher speed limits and higher pollution amounts (a higher speed limit is typical for highways, which are longer than roads in a residential area). It should also be noted that the majority of roads do not have speed limit information (results for such roads are given in the last row of Table 3).

Applying Imputation to the Whole Road Network. After the validation of results, where we identified the classification-regression approach with

Table 2. Performance analysis (average MAPE with the standard deviation) across different road lengths.

Range	Roads		
	Low Pollution	High Pollution	All roads
0%–20%	102.35% (132.20%)	227.74% (303.86%)	103.33% (131.64%)
20%–40%	110.29% (181.97%)	122.18% (171.38%)	126.65% (210.57%)
40%–60%	117.28% (170.33%)	174.71% (231.43%)	111.54% (152.96%)
60%–80%	125.40% (256.94%)	114.28% (186.58%)	127.71% (255.45%)
80%–100%	163.89% (385.46%)	**42.54%** (42.86%)	155.06% (372.81%)
All ranges	123.79% (243.46%)	136.02% (213.69%)	124.82% (241.14%)

XGBoost as the best solution, we applied it to impute the emission data on the 36000 road segments in Pisa that missed that information.

Table 3. Performance analysis (average MAPE with the standard deviation) across different road speed limits.

Limit	Roads				
	All Roads	Low Pollution	Count	High Pollution	Count
10 km/h	287.67% (400.96%)	287.67% (400.96%)	11	–	0
20 km/h	354.91% (302.93%)	354.91% (302.93%)	3	–	0
30 km/h	162.95% (517.46%)	156.54% (513.14%)	151	646.81% (606.09%)	2
40 km/h	107.45% (88.96%)	**97.95%** (75.52%)	36	122.31% (104.93%)	23
50 km/h	**99.77%** (147.56%)	**98.61%** (149.56%)	239	112.95% (121.74%)	21
60 km/h	105.14% (203.50%)	**87.72%** (74.18%)	10	110.59% (229.15%)	32
70 km/h	**73.64%** (83.36%)	**69.57%** (12.71%)	3	**73.97%** (86.59%)	37
80 km/h	**69.36%** (69.02%)	–	0	**69.36%** (69.02%)	11
90 km/h	145.35% (342.12%)	151.09% (347.67%)	25	**1.89%** (–)	1
130 km/h	**83.11%** (–)	–	0	**83.11%** (0.00%)	1
Not Given	126.36% (210.20%)	123.33% (206.90%)	1393	226.69% (281.93%)	42

4.3 Dispersion of Car Emissions

The dispersion approach described in Sect. 3.3 was applied to the central area of the city of Pisa, obtaining a fine-grained map of emissions. Concentrations are log-transformed and they represent accumulated CO_2 emissions over a one-hour period. The results are shown in Fig. 3, where we compare the concentrations obtained using the (incomplete) emissions based only on the original GPS data (left) with those obtained over the data extended through imputation (right).

In both cases, results are rather consistent with common sense, as concentrations are higher around industrial zones (e.g. lower-left and top-right) and some traffic-intensive roundabouts (e.g. on the sides of the rightmost bridge on the river). Yet, the imputed data better captures areas erroneously assigned to very low emissions by the original estimates (in particular in the top-right corner), and some high emissions spots are now slightly more pronounced.

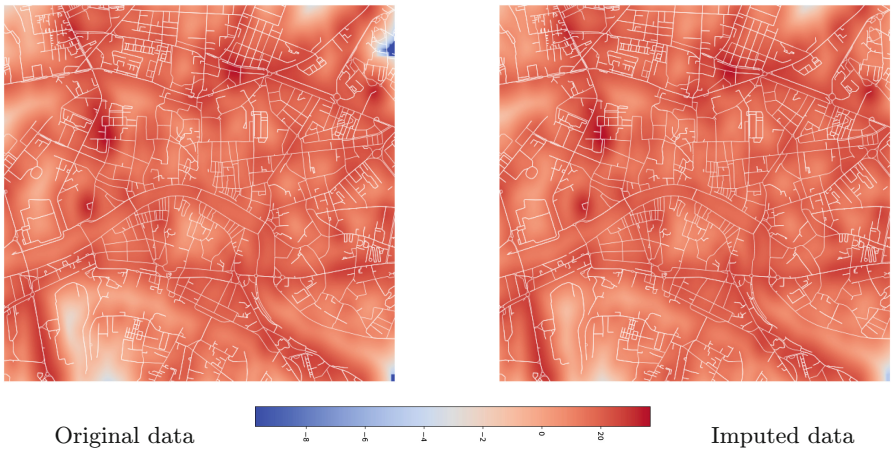


Fig. 3. Log-transformed vehicular CO_2 concentration in Pisa: Original (left) vs. Imputed Data (right).

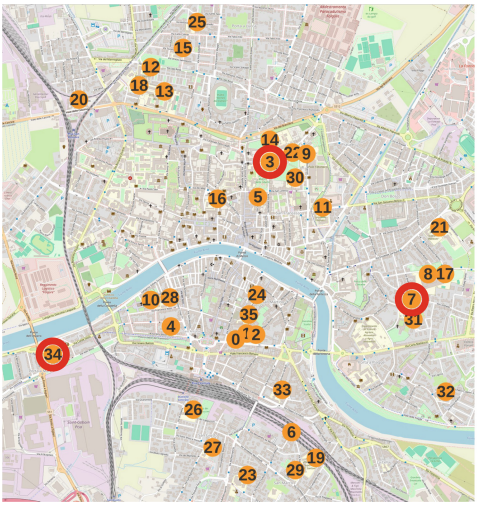
4.4 Exposure to Car Emissions

In this section, we describe two use cases that exploit the emission map obtained in the previous steps to infer the exposure of individuals in two settings: 1) exposure from a static place; and 2) exposure while walking. For the first case, we estimated exposure levels at schools in the central area of Pisa. In the second case, we estimated exposure level to emissions while taking several routes to walk from a university building in Pisa to the central train station, which is a very common task for students.

Static Location-Based Exposure in Schools. Estimation of exposure levels at schools is extremely important because children, who are more vulnerable to air pollution, spend a significant part of their day at school. There are 36 schools in the area, all of which are plotted on the map in Fig. 4(left). There is also a table next to the map which shows the school numbers from the map associated with the corresponding concentrations. The values represent the amount of CO_2 emitted by cars that accumulates in the cell where the school is located over the

course of one year. These numbers can help us to interpret exposure to traffic emissions at schools.

We can see from the table that the highest concentration level is around school number 7, which is indeed located in via Cisanello, the busiest street in the eastern part of Pisa. The second most exposed school (34) is instead located next to State Highway 1 (via Aurelia) and a factory. Meanwhile, the third most exposed school (3) is next to one of the largest parking areas in the center of Pisa. Interestingly, most schools outside the city walls in the south and southeast (19, 23, 32) are the least exposed ones.



School #	Exposure	School #	Exposure
0	39.93	18	9.47
1	23.34	19	1.44
2	44.62	20	7.60
3	**142.70**	21	48.46
4	21.18	22	3.32
5	3.70	23	1.00
6	5.31	24	1.57
7	**958.82**	25	1.89
8	2.10	26	2.75
9	5.75	27	5.03
10	4.61	28	7.01
11	87.03	29	1.57
12	27.93	30	2.93
13	4.55	31	34.68
14	11.35	32	1.45
15	9.73	33	27.55
16	1.90	34	**796.99**
17	2.41	35	1.92

Fig. 4. Map showing the schools located in the center of Pisa (left) and their annual exposure values (right). The 'top' 3 schools are circled.

Exposure While Walking. The travel between the University and the Central train station is one of the popular commuting activities in Pisa. Therefore, we considered a scenario when a person takes a walk to the train station (in the South) from a relatively close university building (in the North). To simulate realistic paths, we used TomTom's API[4] and got 4 alternatives with the shortest travel times for pedestrians (Fig. 5). Exposure values are calculated considering accumulated CO_2 pollution over a one-hour period.

We can see from the figure that the travel time difference between any two alternative paths is less than 2.5 min or 11%. However, we can see that a pair of routes that pass through the central streets lead to lower exposure. The red route (route 1) passes through pedestrian streets (Borgo Stretto and Corso Italia,

[4] https://developer.tomtom.com/.

denoted with a 'pedestrian' sign on the map), consequently having the lowest CO_2 exposure. The blue route (route 2), passes through Borgo Stretto, but not through Corso Italia, which translates to a relatively higher exposure. Most of the travel through Green (route 3) and yellow (route 4) routes pass through driving roads, which results in 2–3 times higher CO_2 exposure compared to the red path. Especially, the yellow route passes through the busiest parts of Pisa after crossing the river: lungarno Fibonacci, piazza Guerrazzi (roundabout sign on the map), and viale Bonaini ('traffic light' sign on the map).

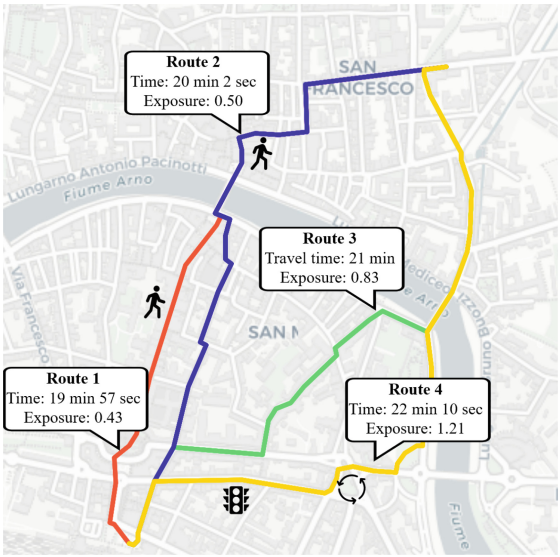


Fig. 5. Alternative walking routes, their travel times and CO_2 exposure amount (mg), from a University building to the Central Train Station of Pisa.

5 Conclusion

In this paper, we introduced a 4-step process to derive estimates of individual exposure to emissions (CO_2 in particular) from raw GPS vehicle data, introducing an imputation phase to account for the limited coverage of GPS data, and exploring existing compatible tools. Finally, emission estimates obtained through a simple dispersion phase are exploited to infer exposure at static locations and along trips with two small use cases.

The proposed pipeline provides the first viable and easy-to-use end-to-end solution for the task under study, yet several issues and open problems emerged, waiting for further studies. First, the best solution for emissions imputation is still far from optimal, suggesting that the problem has peculiarities that call

for ad hoc approaches different from other graph-based methods. Also, a factor not considered so far is that we expect low-emission road segments to miss emission data more often than high-emission ones, thus introducing a bias that needs proper treatment. Second, the spatial density of the initial emission estimates might be improved, for computing map-matching with shortest paths between consecutive points, inferring someway the speed and acceleration in reconstructed points. This improved density might allow to have time-dependent estimates, useful to obtain more precise exposure estimates. Finally, more sophisticated dispersion models should be included, for instance taking into account weather data, made possible by time-dependent estimates.

References

1. Transforming Our World: The 2030 Agenda for Sustainable Development. Springer (2018). https://doi.org/10.1891/9780826190123.ap02
2. Böhm, M., Nanni, M., Pappalardo, L.: Gross polluters and vehicle emissions reduction. Nat. Sustain. **5**(8), 699–707 (2022). https://doi.org/10.1038/s41893-022-00903-x
3. Briggs, G.A.: Diffusion estimation for small emissions. Preliminary report (1973). https://doi.org/10.2172/5118833
4. Carslaw, D.C., Beevers, S.D., Tate, J.E., Westmoreland, E.J., Williams, M.L.: Recent evidence concerning higher NOx emissions from passenger cars and light duty vehicles. Atmos. Environ. **45**(39), 7053–7063 (2011). https://doi.org/10.1016/j.atmosenv.2011.09.063
5. Chan, R.K.C., Lim, J.M.Y., Parthiban, R.: Missing traffic data imputation for artificial intelligence in intelligent transportation systems: review of methods, limitations, and challenges. IEEE Access **11**, 34080–34093 (2023). https://doi.org/10.1109/ACCESS.2023.3264216
6. Chang, Y., Tanin, E., Cao, X., Qi, J.: Spatial structure-aware road network embedding via graph contrastive learning. In: Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, 28–31 March 2023, pp. 144–156. OpenProceedings.org (2023). https://doi.org/10.48786/edbt.2023.12
7. Cornacchia, G., Böhm, M., Mauro, G., Nanni, M., Pedreschi, D., Pappalardo, L.: How routing strategies impact urban emissions. In: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2022. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3557915.3560977
8. Dewuf, B., et al.: Dynamic assessment of exposure to air pollution using mobile data. Int. J. Health Geograph. **15** (2016). https://doi.org/10.1186/s12942-016-0042-z
9. Ekström, M., Sjödin, A., Andreasson, K.: Evaluation of the COPERT III emission model with on-road optical remote sensing measurements. Atmos. Environ. **38**, 6631–6641 (2004). https://doi.org/10.1016/j.atmosenv.2004.07.019
10. Gately, C.K., Hutyra, L.R., Peterson, S., Wing, I.S.: Urban emissions hotspots: quantifying vehicle congestion and air pollution using mobile phone GPS data. Environ. Pollut. **229**, 496–504 (2017). https://doi.org/10.1016/j.envpol.2017.05.091

11. Gavros, A., Karatzas, K.: Air pollution due to central heating of a city-centered university campus. In: Wohlgemuth, V., Naumann, S., Behrens, G., Arndt, H.K. (eds.) ENVIROINFO 2021. Progress in IS, pp. 117–133. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-88063-7_8
12. Huang, Y., et al.: Remote sensing of on-road vehicle emissions: mechanism, applications and a case study from Hong Kong. Atmos. Environ. **182**, 58–74 (2018). https://doi.org/10.1016/j.atmosenv.2018.03.035
13. Huang, Y., et al.: Emission measurement of diesel vehicles in Hong Kong through on-road remote sensing: performance review and identification of high-emitters. Environ. Pollut. **237**, 133–142 (2018). https://doi.org/10.1016/j.envpol.2018.02.043
14. Jepsen, T.S., Jensen, C.S., Nielsen, T.D.: Relational fusion networks: graph convolutional networks for road networks. IEEE Trans. Intell. Transp. Syst. **23**(1), 418–429 (2022). https://doi.org/10.1109/TITS.2020.3011799
15. Kousoulidou, M., et al.: Use of portable emissions measurement system (PEMS) for the development and validation of passenger car emission factors. Atmos. Environ. **64**, 329–338 (2013). https://doi.org/10.1016/j.atmosenv.2012.09.062
16. Larssen, S., Sluyter, R., Helmis, C.: Criteria for EUROAIRNET: the EEA air quality monitoring and information network. Technical Report No. 12, European Environment Agency, Copenhagen, Denmark (1999). https://www.eea.europa.eu/publications/TEC12
17. Leelőssy, Á., Molnár, F., Izsák, F., Havasi, Á., Lagzi, I., Mészáros, R.: Dispersion modeling of air pollutants in the atmosphere: a review. Cent. Eur. J. Geosci. **6**(3), 257–278 (2014). https://doi.org/10.2478/s13533-012-0188-6
18. Li, Q., Liang, S., Xu, Y., Liu, L., Zhou, S.: Assessing personal travel exposure to on-road PM2.5 using cellphone positioning data and mobile sensors. Health Place **75**, 102803 (2022). https://doi.org/10.1016/j.healthplace.2022.102803
19. Liang, M., Chao, Y., Tu, Y., Xu, T.: Vehicle pollutant dispersion in the urban atmospheric environment: a review of mechanism, modeling, and application. Atmosphere **14**(2) (2023). https://doi.org/10.3390/atmos14020279
20. Liu, Z., et al.: Learning geo-contextual embeddings for commuting flow prediction. In: AAAI Conference on Artificial Intelligence (2020). https://api.semanticscholar.org/CorpusID:211037977
21. Nyhan, M., et al.: Exposure track - the impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. Environ. Sci. Technol. **50**, 9671–81 (2016). https://doi.org/10.1021/acs.est.6b02385
22. Nyhan, M., et al.: Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model. Atmos. Environ. **140**, 352–363 (2016). https://doi.org/10.1016/j.atmosenv.2016.06.018
23. Rahman, M.N., Idris, A.O.: TRIBUTE: trip-based urban transportation emissions model for municipalities. Int. J. Sustain. Transp. **11**(7), 540–552 (2017). https://doi.org/10.1080/15568318.2016.1278061
24. Schneider, P.: Alternative technologies for monitoring urban air quality – from satellites to sensor networks (2019). https://nilu.com/publication/1756394/. Lecture presented at Tekna - Faggruppen for Energi, Industri og Miljø, Oslo
25. de Souza, P., et al.: Quantifying disparities in air pollution exposures across the United States using home and work addresses. Environ. Sci. Technol. **58**(1), 280–290 (2024). https://doi.org/10.1021/acs.est.3c07926

26. Wu, Y., Song, G., Yu, L.: Sensitive analysis of emission rates in moves for developing site-specific emission database. Transp. Res. Part D: Transp. Environ. **32**, 193–206 (2014). https://doi.org/10.1016/j.trd.2014.07.009
27. Yao, X., Gao, Y., Zhu, D., Manley, E., Wang, J., Liu, Y.: Spatial origin-destination flow imputation using graph convolutional networks. IEEE Trans. Intell. Transp. Syst. **22**(12), 7474–7484 (2021). https://doi.org/10.1109/TITS.2020.3003310
28. Yoon, S., Moon, Y., Jeong, J., Park, C.R., Kang, W.: A network-based approach for reducing pedestrian exposure to PM2.5 induced by road traffic in Seoul. Land **10**, 1045 (2021). https://doi.org/10.3390/land10101045