URBANITE

Supporting the decision-making in urban transformation with the use of disruptive technologies

Deliverable D4.1

Strategies and algorithms for data modelling and visualizations

Editor(s):	Alina Machidon, Maj Smerkol
Responsible Partner:	"Jožef Stefan" Institute
Status-Version:	Final – v1.0
Date:	01.12.2020
Distribution level (CO, PU):	Public

Project Number:	GA 870338
Project Title:	URBANITE

Title of Deliverable:	Strategies and algorithms for data modelling and visualizations
Due Date of Delivery to the EC:	30/11/2020

Workpackage responsible for the Deliverable:	WP4 – Algorithms and simulation techniques for decision-makers
Editor(s):	"Jožef Stefan" Institute
Contributor(s):	Matjaž Gams, Maj Smerkol, Alina Machidon, Nuša Muršič (JSI) Giuseppe Ciulla (ENG) Gonzalo Lázaro, Ignacio Olabarrieta, Sergio Campos, Ibai Laña (TECNALIA)
Reviewer(s):	Sonia Bilbao (TECNALIA)
Approved by:	All Partners
Recommended/mandatory readers:	WP3, WP5, WP6

Abstract:	This document presents a survey of existing methods for data modelling and visualizations that could be applicable to the URBANITE domain. The document provides an overview of data models and analytical models, and identifies references for specific problems present in urban mobility planning: the transport demand modelling, the analysis of noise and air pollution and the analysis of traffic estimation and prediction and identification of traffic jams. Additionally, the concept of actionability is introduced as a characteristic that any data-analytics method should present for application and use in real decision environments. This deliverable is the result of tasks T4.1 and T4.4.
Keyword List:	Data analytics, Data visualization, Prediction, Actionability
Licensing information:	
Disclaimer	This document reflects only the author's views and neither Agency nor the Commission are responsible for any use that may be made of the information contained therein

Document Description

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	08/06/2020	Draft ToC and contents	JSI
V0.2	11/11/2020	Contribution to 3.4 Implementation methods and tools for data visualisation	ENG
V0.3	19/11/2020	Contribution to section 3. Data analytics strategy for urban planning and 5.	ENG
V0.4	25/11/2020	Actionability, Data analytics strategy and Integration of sections	TEC
v0.5	27/11/2020	Integrated contributions from all partners. Ready for internal QA.	JSI
V0.6	30/11/2020	Integrated the comments received in the QA	JSI
V1.0	01/12/2020	Ready for submission	TEC

Table of Contents

List of Figures

Terms and abbreviations

AFC	Automatic Fare Collection
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AQI	Air Quality Index
ATIS	Advanced Traveller Information System
ATMS	Traffic Management System
BNN	Bayesian Neural Networks
BPNN	Back propagation Neural Networks
BSD	Berkeley Software Distribution
CO2	Carbon Dioxide
D3	Data Driven Document
DOM	Document Object Model
DL	Deep Learning
DB	Decibels
EC	European Commission
FCD	Floating Car Data
GIS	Geographic Information System
GPS	Global Positioning System
HMM	Hidden Markov Model
IDW	Inverse Distance Weighting
IJESD	International Journal of Environmental Science and Development
ISO	International Organization for Standardization
ITS	Intelligent Transportation Systems
JS	JavaScript
ML	Machine Learning
MLP	Multilayer Perceptron
OD	Origin-Destination
PCA	Principal Components Analysis
PM	Particulate *Matter*
RNN	Recurrent Neural Networks
SNN	Stochastic Neural Networks
SOM	Self-Organizing Map
SUMP	Sustainable Urban Mobility Plan
SVM	Support Vector Machine
UI	User Interface
xAI	Explainable Artificial Intelligence

Executive Summary

This document is the first deliverable of WP4, the work package focused on the development of advanced AI algorithms for analysis of big data and simulation techniques, as support for policy-makers to tackle complex policy problems. This deliverable presents a survey of existing methods for data modelling and visualizations that could apply to the URBANITE domain.

The document has four sections devoted to: data analysis and modelling methods, data analytics strategy for urban planning, data visualization methods, and the real applicability of such methods. The document begins with an overview of data models and analytical models, and then identifies references for analysis and visualization in specific problems present in urban mobility planning: the transport demand modelling, the analysis of noise and air pollution and the analysis of traffic estimation and prediction and identification of traffic jams. Additionally, the concept of actionability has been introduced as a characteristic that any data-analytics method should present for application and use in real decision environments.

This document constitutes the first step on the selection and design of such methods, by identifying existing references and technical options. In conclusion, there are many available methods and tools, covering different data, analysis and forms of visualization and interaction and the real challenge for their successful application on the project use cases, is the appropriate selection and integration, according to case needs and the available data sources, in terms of availability and quality.

WP4 activities run in parallel with the detailed definition of use cases, so once these are detailed, the next step will be the refining, adaptation and extension of the methods identified in D4.1, according to the specific needs of the cities. Deliverables D4.2 and D4.4 (both due the twenty-fourth month of the project) will provide the implementation, extension and adaptation of the adopted algorithms, and the whole traffic flow model, supporting such methods and analysis. The insights of D4.1 will also be questions to discuss on the WP2 stakeholder sessions.

1 Introduction

Deliverable D4.1 presents a survey of existing methods for data modelling, analysis and visualizations that could be applicable to the URBANITE domain. The methods are discussed and evaluated in the context of urban mobility and policy support.

This document is part of work package WP4 "Algorithms and simulation techniques for decision-makers" and is the first outcome of tasks T4.1 "Methods for Exploratory Data Analysis and User Interaction" and T4.4 "Advanced visualizations methods".

1.1 About this deliverable

This state-of-the-art report is generated as a research in the field of strategies and algorithms for data modelling, analysis and visualizations and reflects opportunities that exist for the development of such algorithms and strategies. It also identifies strategies for progressing this important area of learning. It must be noted that the data provided does not claim to be complete, neither presents an empirically grounded research, but it reflects estimations.

Furthermore, this state-of-the-art report will inform the next steps of the project, which seek to develop an online platform providing several selected algorithms and strategies (selected upon the analysis performed and summarised in this report) for data modelling, analysis and visualisations in the area of traffic data and urban mobility.

1.2 Document structure

This document is organised into four main sections and a conclusion chapter.

This introductory chapter provides a brief summary of this deliverable, together with the structure of the present document.

The second section is the overview of relevant data modelling and analysis methods, including a discussion of their applicability and role in the context of the URBANITE solution and their evaluation in terms of data and user requirements. The third section identifies the main issues and relevant results on the application of data modelling and analytics methods to some relevant problems, identifying references for the analytics strategy in the urban planning process, evaluated in terms of accuracy, data requirements and interpretability for end-user. The fourth section is the overview of visualization methods, both in general and specifically, in the domain of urban mobility, important event detection and policy validation. It includes discussion on their applicability in the context of the URBANITE solution, and the evaluation criteria.

The fifth section introduces the concept of actionability as a characteristic that any data-analytics method should present to be successfully deployed in a real operating environment and analyses its application in the context of urban mobility planning.

Finally, the conclusion section summarises the most important points of the document and identifies the challenges of the application of data-based techniques in the project. The document ends with the references section.

2 Data analysis or modelling methods

This section describes motivation for using relevant data modelling methods, explaining the rationale for usage and evaluation of the methods.

2.1 Objectives

Main objectives of the data modelling tools are to enable the policy makers to interactively explore available heterogenous data, discover related data attributes that may indicate causal relations and enable the deeper understanding of complex and so-called 'wicked' mobility-related problems.

2.2 Context

The tools and methods described will be used in the environment of the collected data that will include many heterogenous data sets harvested from sources as described by the Deliverable D3.1.

The environment consists of the traffic network with other geo-referenced data sets and other data sets, which can be used simultaneously. The geo-referenced data sets will include:

- The city's map and traffic network representation.
- Traffic flow meters, bicycle count stations and pedestrian counters.
- Air pollution sensor data at each metering site.
- Noise metering data at each metering site.
- Positions of certain facilities such as schools, public transport stops, parking facilities and related data.
- Schedules of public transport providers.
- Demographic data of city districts.

These methods are also relevant for analysis of the traffic simulation results. Most of the simulation results are geo-referenced as well, including the simulated trips, generated air and noise pollution.

The tools developed aim to exploit large amounts of heterogenous data while providing common methodology and cohesive analysis.

2.2.1 Projection methods

Projection methods are used in the context of big data analysis in multiple ways, listed below.

1. **Projection methods for dimensionality reduction** help to decrease the computational difficulty of methods used for further analysis of results. Due to high computational complexity of big data analysis reducing the dimensionality of the data often provides large efficiency benefits with low costs in terms of analysis accuracy.
2. **Projection methods for visualization** use similar techniques to reduce the cognitive overload visualizing data sets by projecting high dimensional data sets in 2- or 3-dimensions that can be easily visualised in an understandable way. While some of the detailed information available is lost, the most important aspects of the data can be understood easily.
3. **Projection methods for data exploration** enable us to perform analysis of correlations between different attribute groups much more efficiently. Data attributes that get projected into the same component are usually somewhat correlated which can be leveraged to speed up correlation discovery and confirmation.

Some relevant methods are the following:

Principal Components Analysis (PCA) is a commonly used method of dimensionality reduction that maximises the data's variation while projecting the data to lower dimensional space of first few principal components – co-orthogonal and pointing into the direction of largest variance. It allows visualization of different clusters of data on a 2- or 3-dimensional plot. To allow for best results, tools for preprocessing the data will be provided. We are considering providing common preprocessing methods such as mean centering, common unit checks, and scaling the data appropriately in an intuitive and guided way to guide the user away from making common data-analysis mistakes and guiding the entire process.

Factor analysis is a statistical method that aims to discover the hidden variables that explain the data in lower dimensional space. It can be used on time series data and commonly used to discover trends in mobility patterns.

Targeted projection pursuit is used for finding projections of interest in an interactive manner, promoting data exploration, while projection pursuit techniques are used for discovery of wider projection candidates. Combined, these methods guide the user to play with the data and discover new possible relations. These methods are also used for human-guided intelligent feature selection and data modelling.

Other projection methods will be considered as new needs of the use cases are presented.

2.2.2 Clustering methods

Clustering methods allow the users to discover clusters of similar data samples. This allows us to approximate a typical data point from each of the clusters, compare clusters and find differences and similarities between different clusters.

Different clustering methods are appropriate for different data types and use cases as discussed below.

K-Means clustering is a quick way of clustering data based on Euclidean distance. It is very effective for clustering big data sets. This method is most appropriate when the number of clusters is known, and clusters are somewhat balanced. However, this method cannot be used on series data, such as time-series or sequences of positions.

Affinity propagation is a clustering method that is applicable to smaller data sets due to high computational complexity. This method can handle more sophisticated data such as time-series and heavily unbalanced data sets. This method will be used for clustering of simulated trips and constructing traffic flow data models.

Spectral clustering is a method of clustering based on the spectrum of a similarity matrix of a data set, specifically appropriate for clustering graphs. This allows for clustering of simulated travel paths and thus discovering traffic patterns. This clustering method is also potentially useful for discovery of specific noise sources and air pollution hot spots.

Hierarchical clustering is appropriate to segment the demographical data of a population, analysis of the rolling stock and similar data. Hierarchical clustering generates trees that can be easily read and understood.

2.2.3 Learning methods

This section introduces some of the most popular learners or learning methods, both supervised and unsupervised.

With a focus on supervised approaches**, Decision Trees (DT)** has become one of the most popular methodologies, due to its algorithmic foundation, easy to comprehend and implement. Summarizing, a decision tree is a set of nodes organised hierarchically in layers or categories, from up to down, from a root node to leaves nodes, where nodes identify a test or logic questions, involving some of its attributes each time. After a first training, new measures or data sets will end up in one of the tree leaves, and the process and reasoning are human-readable as a decision flow.

The **Artificial Neural Networks (ANN)** family of techniques is possibly the most popular Machine Learning (ML) technique, being widely adopted. Along the last decades, from the single-layer perceptron to recent Deep Learning (DL), there is a multitude of variants as Multilayer perceptron (MLP), Recurrent Neural Networks (RNN), Bayesian Neural Networks (BNN), Back propagation Neural Networks (BPNN), Stochastic neural networks (SNN), among many others. A good survey can be found in [1]. In general, their behaviour starts with an external stimulus, which propagates along with the network, according to weighting inputs and activation function at each node. The adjustment of the structure or training, through the comparison of the expected and current output, as bases for an incremental adjustment of link weights, functions and the layout itself.

Support Vector Machines (SVM) are another supervised technique, whose aim is at creating the best possible separating border between the classes to classify, maximising the distance between instances from each side of that border, by using a highly non-linear model. SVM can be used for classification and also regression problems. Bayesian models, Hidden Markov model (HMM) and other probabilistic approaches.

Finally, the called **ensembles**, which are combinations of learners, can combine both same kind of learners (bagging, boosting or staking approaches), complementing their capabilities.

On the other hand, **unsupervised (unlabelled) algorithms** infer knowledge from datasets without explicit knowledge of the labels assigned to the input data, by seeking hidden structures in datasets. These techniques are applied to different problems of functionalities as clustering, outlier detection or dimensionality reduction, among others. The next section focuses on Self-Organizing Map (SOM), as the most relevant case of unsupervised ANN.

2.2.4 Self-Organizing Map

A self-organizing map (SOM) is an artificial neural network that is trained unsupervised to project data samples on a grid in accordance to a certain metric. The self-organizing map maps a data space onto a commonly 2- or 3-dimensional grid, organizing the data samples by distance or similarity of the data points. It enables users to overview a large data set and see how the data groups according to different attributes, whether some attributes are correlated to other attributes.

SOMs have some disadvantages regarding the understandability of visualizations. The points mapped using a SOM are displayed on a grid however the Euclidic distance between the points does not represent any actual metric. A recent improvement algorithm called ViSOM [2] allows for displaying the points mapped according to a chosen metric. Using the ViSOM method is therefore also well suited for cluster analysis, complementing clustering methods mentioned above.

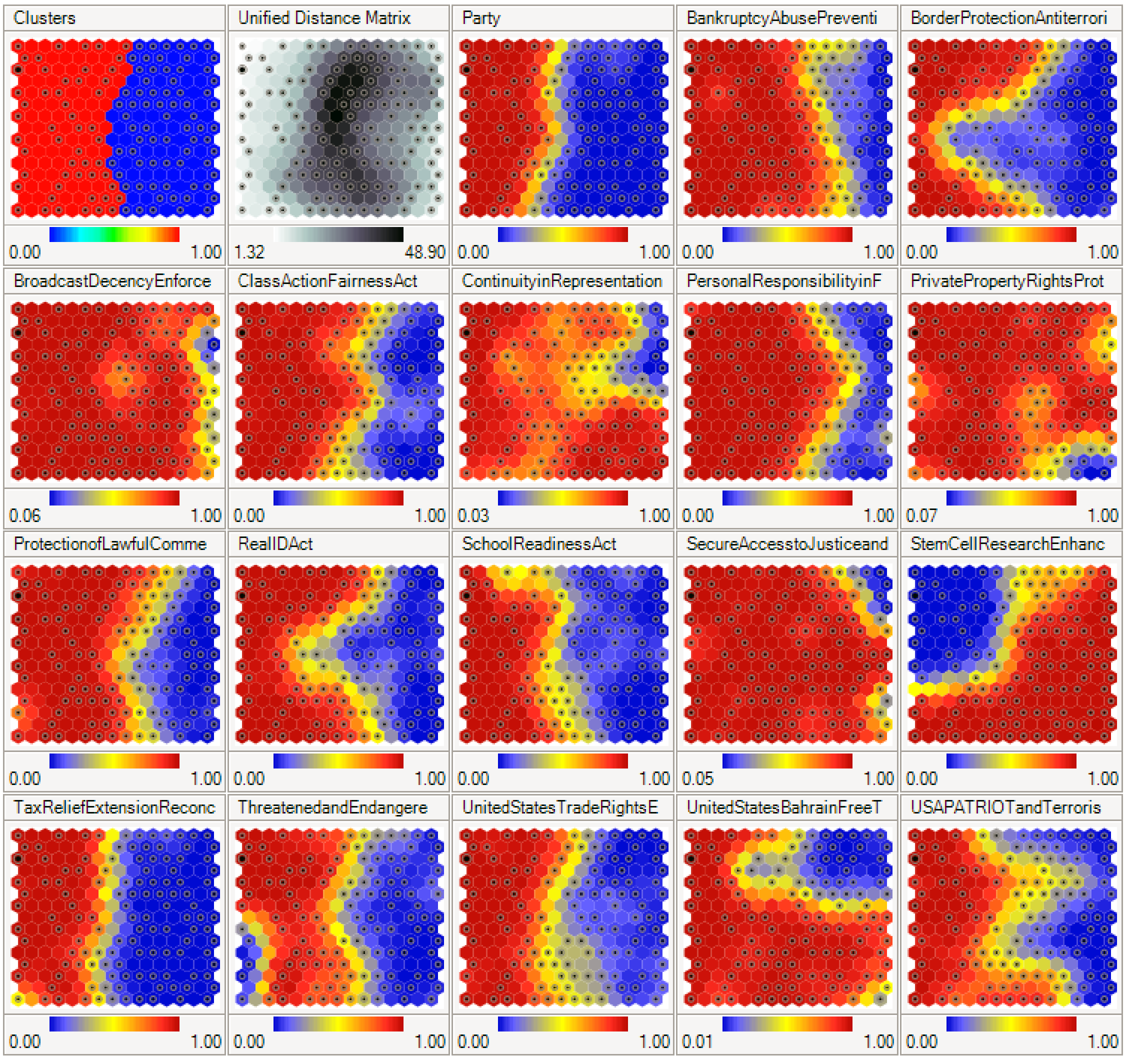

Figure 1 Example of a self-organizing map.

Figure 1 shows an example of a self-organizing map, representing stances of members of a senate on different topics. Each of the images of the SOM shows a specific attribute. The second image of the top row shows the distance between neighbouring points – it is not uniform, as mentioned above.

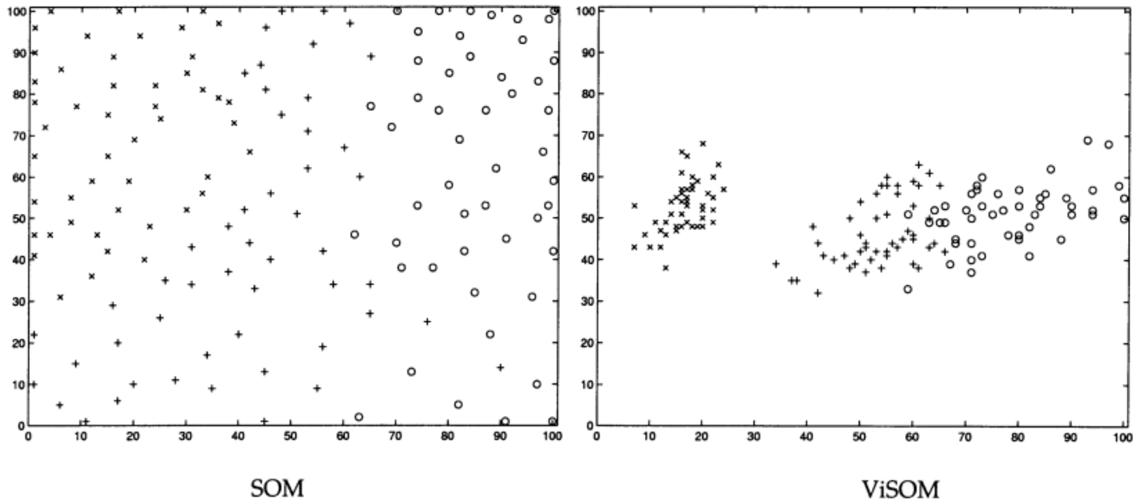

Figure 2 Mappings of Iris data set left SOM, right ViSOM. (Doi: https://doi.org/10.1109/72.977314)

2.2.5 Correlation discovery methods

Correlation discovery enables users to sieve through big data and find correlated data attributes that may be worth investigating.

Linear correlation discovery methodology consists of quick discovery of many potential groups of attributes by sampling available data, measuring the correlation using the appropriate method depending on attributes' data types.

Using partial correlation analysis and linear correlation discovery we can approximate a graph of related attributes and attribute groups. Thus, we can identify potential direct correlations.

3 Data analytics strategy for urban planning

This section identifies main issues and selected strategies on the application of data analytics to some domain specific or applications in urban planning to URBANITE related domains.

3.1 Transport demand model estimation

Knowing the exact way public transportation is being used by citisens within a given city has a crucial importance to take the correct strategic and/or operational decisions. The origin destination matrices, O/D matrices, that record how many trips are being taken from a given origin location to another destination in a given period of time are the most precise and complete structure that describe how people use a given transportation system or in other words the transport demand estimation [3] [4]. In the popular four-step planning model, the O/D matrices, are computed in the second of the steps.

The construction of these O/D matrices in general involves a very large accounting problem (actually, a big data problem). In this problem all the trips from all the locations, all the destinations and at different times are considered. In the best scenario there is exact information of from where and to where the passengers travel but in general that is not the case. Most of the cases the O/D matrices can be only estimated by means of indirect methods. These classical methods to tackle the problem can be divided in three main groups [5]:

- Survey-based methods (e. g. survey of car-owners, public transport travellers, etc). Typically these methods imply the use of large number of resources in order to obtain statistically relevant information.
- Trip distribution models (e. g. gravity models). These imply taking assumptions about the mobility which one wants to measure.
- Updating methods (which update old versions of the O/D matrices using new or updated relevant information as for example growth measures or flow values at different points)

The most modern of the classical approaches uses combinations of these methods.

In recent times the improvements of computer power, big data techniques and automatic methods to capture information have produced the emergence of methods that are based directly on the data. One of these methods has been made possible by the fact that a huge percentage of the population use mobile phones (>80% in European countries). Mobile phones need to be always in contact with the carrier antennas in order to be able to connect. Therefore, computation of O/D Matrices now can be obtained my analysing the antenna contact information and CDR (Call Detail Record) obtained and stored by cell phone carriers. Privacy issues disallow the general use of these data, the access is heavily restricted and for the city to be able to use these data implies having a strong alliance between the cell phone carrier and the city. Moreover, the processing of the cell phone related data produces general O/D Matrices that in their first approximation do not distinguish between different modes of transportation.

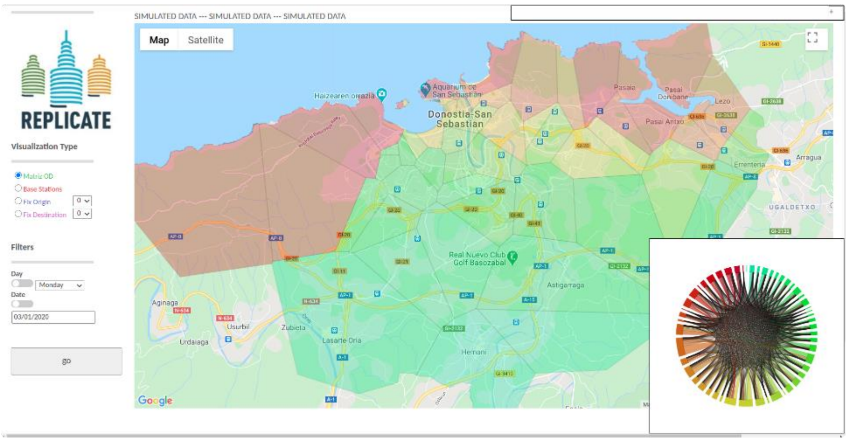

Figure 3:Image from EU project Replicate where O/D Matrices have been computed using CDR. Provided by Tecnalia.

Another data approach is to use the available information from the fare system of the public transport that wants to be analysed. Automatic Fare Collection (AFC) systems have a widespread use over most of the public transport systems nowadays [6]. Many AFC systems produce valuable information from the use of travel cards (Oyster in London, Barik in Bilbao, OV-chipkaart in Amsterdam, etc...) typically referred as smart cards which in many cases are personalised and used by a single traveller. The processing of data from these cards have several advantages:

- The mode of transportation is always known since the use of the card is tied up with the transportation system.
- All the transactions using the travels cards are stored in order to keep track of payments.
- The data produced by the transportation system is used in order to obtain the O/D matrix of the same system. No third party involved.

- The origin/destination locations are directly mapped to stations in the transportation system.
- The penetration of the smart cards system usually is deep among the typical travellers due to the important discounts and other benefits that these systems provide. Some authors claim that up to 90% and during the Covid-19 pandemic in some cases the penetration has gone up to 100% as in Bilbao Bizkaibus and Bilbobus bus systems.
- Information from personalised cards imply that the mobility patterns can be identified up to a single traveller in the system.

According to the data stored by the AFC system, two main situations are found:

- It stores the boarding and the alighting locations for each trip; this is the case for some of the subway systems and specially when the fare varies from station to station in longer transport systems.
- It only stores the boarding location, the alighting one always being disregarded.

The first situation which has more complete information and which is easier to process is highly difficult to be found in urban environments. The normal situation is that the traveller is not forced to check the exit from the vehicle and therefore it falls into the second main situation described. Lacking the alighting locations imposes a serious drawback for computing the O/D matrices and presses to use more advanced analytical work. Specifically, it comes into relevance the work presented in [7] where making use of the fact that the data is obtained from personalised smart card systems can produce estimation of the exit location for every trip based on the individual mobility patterns for each traveller. This method is nowadays known as trip-chaining method and it can be used to estimate the alighting location by associating it with the boarding location of the following trip for the same person.

3.2 Analysis of noise and air pollution

Air quality in cities has become one of the main environmental issues by both citizens and the administration, due to its adverse effects on health. It is known that one of the main sources of urban pollution is transport, proof of this has been the general high impact on emissions during periods of confinement and mobility restrictions in different European cities due to Covid-19. Likewise, the effect or dependence on factors such as the orography of the area and meteorological conditions is recognised. Given the importance and urgency of greater control of emissions, cities have articulated different traffic containment measures along the time with variable impact in pollutant levels.

Pollution models are inputs for the decision of adequate traffic management strategy [8]. Recently, big data have become widely available, in the form of traffic flows, general urban mobility traces, meteorological data, and even 3D city models, enabling us to solve such models from a data perspective. In the existing (extensive) literature, while some researchers use only traffic data [9], some researchers build their models, including both traffic and meteorological data as inputs [10]. Some ensemble supervised learning model have been introduced [11], providing also quantitative measures of the different variable importance for a better understanding of their dependencies. Several machine learning techniques have been applied, being neural networks ANN the most frequently used, with variations on the model and improvements in the pre-processing stages [12] [13]. Methods as decision trees [14], support vector machines (SVMs), and linear regression has also been used to model PM concentrations. [15] [16]

The locations of air quality stations and traffic loops, as contextual information for predictions, is also a key factor to consider, due geographical and topological relations. We can find detailed guides [17] [18] of the application of bigdata and ML techniques on IoT sensor for estimations and predictions. The IJESD also published a summary [19] of the research results relating to the application of ML and AI to air quality evaluation, identifying the challenges and future research needs: the improvement of big data quality and its assurance, and the modelling issues for dynamic air quality monitor. An advanced approach considers two different stages, splitting among spatial and temporal models: a model of the spatial correlation between air qualities of different locations by analysing spatially-related features (e.g., POI information and the navigable network segments), and the second model for the temporal dependency of air quality in each location. According to this approach, suitable combinations of ANNs and linear-chain conditional are proposed [20]. In detail, relevant reports deal with the assessment [21] [22] of road vehicle emissions analysing how emission factors are derived, and clearly explains what an emission factor represents and why it often differs from the emissions of an individual vehicle. Emission models provide an emission estimate, becoming more sophisticated. This is reflected in the level of resolution, the level of details for input data and the extent of incorporating driving behaviours.

With respect to noise, existing approaches are similar, always considering that the physical phenomena are clearly different. As in the case of CO2 emissions, traffic is a major contributor to overall urban noise pollution, impacting on surrounding areas. Along the last two decades, different traffic noise prediction models have been developed, based on field measurements, mainly with the application of statistical methods. A good introduction to the experience on different countries can found in [23], remarking that the specific scenario, challenges and supporting methods, depends on the vehicle park characterization, the predominance of specific transport modes (e.g. two-wheeled in India), city layout and urbanism, and street and buildings elevation models.

In general, the early linear regression analysis has moved to more advanced modelling techniques, mainly ANN, whose features and parameters are the number of vehicles (for categories), average height of the surrounding buildings and width of the street to achieve a better prediction. Other exciting approach is the application of numerical methods for approximating complex ANN, as the Levenberg–Marquardt (L–M) [23] application to model predicting highway traffic noise.

Nevertheless, some studies obtained that no significant performance differences are obtained among different models, but the key factor releases on the detail of street canyons to predict sound propagation [24]. In a study presented recently [25], new models for predicting traffic-noise annoyance based on noise perception, noise exposure levels, and demographics were developed, combining different ML techniques, also compared in terms of errors.

This report also identifies other related functionalities and experiences with regards to noise modelling and prediction:

- Classification of urban locations for environmental noise assessment based on the traffic composition, using expert systems, addressing an early identification of potential road-traffic-noise related problems [26].
- Analysis of noise annoyance and effect models proposed the application of fuzzy logic and rules, allowing predictions for specific groups or collectives, instead of managing a great population. This solution provided mobility managers a ‘‘noise annoyance adviser” and for social science a test of hypotheses around noise sensitivity or the degree of urbanization [27].

- Prediction of the temporal structure and spectral composition of road traffic noise [28].
- Prediction or classify different types of road surfaces, the effect of pavement distresses on rolling noise levels and detection of roads wetness from tire-surface interaction [29].

Also, some initiatives promote citizen collaboration, by using the mobile phone as a mobile sensor, for the acquisition of ground noise and ambient information, but presenting high sensibility on the quality of the device.

3.3 Traffic estimation and identification of traffic Jams

There exist multiple ways to obtain measurements of the state of traffic. Here we only list few of the ways to get information from the traffic:

- Traffic camera images/video: this can be considered as one of the more direct ways for identifying traffic jams. For the automatic identification an artificial image processing software is needed.
- Drone images/video: this case is similar to the previous case, in the sense that the automatic identification relies on images captured by cameras and needs for an image processing step. The main difference is that the drone can move its point of view, its overall location, and it can have a wide field of view.
- Very High-Resolution Satellite imagery: this is a very new way to capture information that has applicability specially for strategic decision-making process in regions where there is a lack of other sensors [30]
- Floating car data (FCD) [31]: in the last decade this has become one of the most useful methods for automatic identification of the state of traffic. It is based on data obtained by cell phones or by other devices that travel with the flow of vehicles, e. g. navigation or fleet tracking devices.
- Traffic flow sensors: typically, these are sensors that are installed below the road asphalt and consist in a loop of wire where a magnetically induced current is generated every time a vehicle passes over the road. These are the sensors that have been widely used by road managers.
- Doppler Radar sensors: these systems assume a measurement of the instantaneous velocity of the vehicles.
- Speed Trap: these systems measure the average velocity of a given vehicle between two different positions. Automatic systems rely in automatic identification of traffic plates which is done in most of the cases by images processing.
- Collaborative apps: (e. g. Waze)[1] This mobile application gathers information from its community of users who actively report the state of the road.
- Bluetooth beacons.

From the previous list two of the capturing methods stand up from the rest due to their popularity: traffic flow sensors and FCD.

Of particular interest is the use of the method FCD. The main advantage is that processing is done from the information gathered from devices that ride with the vehicles, either cellular phones or navigation devices. This implies that there is no need for installing expensive infrastructure in the road network that needs to be maintained and updated as time passes by. The expense of the FCD devices typically lies on the users of the network.

[1] Waze App. https://play.google.com/store/apps/details?id=com.waze&hl=en&gl=US

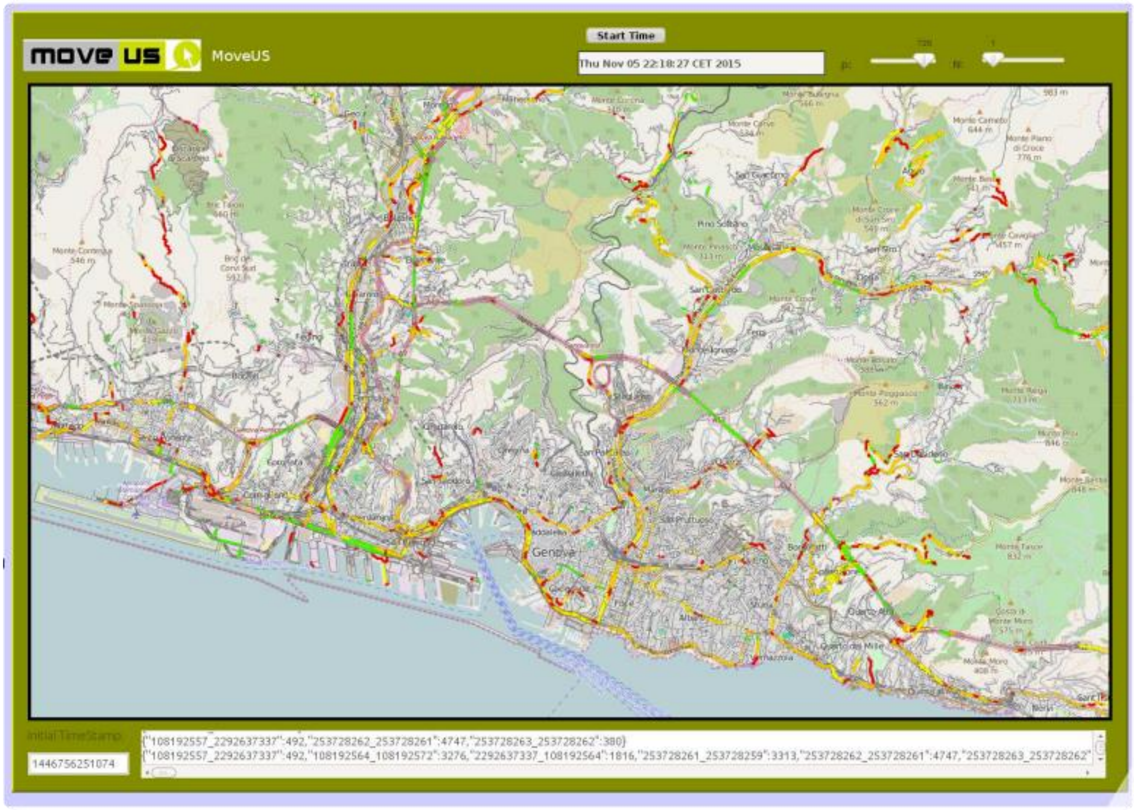

Figure 4: Image from EU project MoveUS where the level of service is computed using FCD. Provided by Tecnalia.

Another important feature of this capturing method is the fact that in order to obtain the level of service of a large amount of the navigation network only a small percentage of the vehicles need to have sensors.

In order to measure the state of a particular road segment in the network several steps need to be performed:

- Capture of the GPS data: The service needs to obtain the GPS data from a given vehicle. In theory, a list of measurements is considered at every time step.
- Cleaning of the measurements: This process, which is usually called as Map-Matching [32] cleans the noise that measurements may contain. The noise could be due to several different reasons including multipath effects, no line of vision, problems with the device, etc. As an output of this important process, the actual road segment where the vehicle is located is returned.
- Averaging and ageing process: this process computes the average of all the vehicles that are on the same road segment. Values that have been captured some time ago should be weighted out in order to give them less importance.
- Compares the average result with the nominal speed of the road segment in free flow and computes a level of service.

Automatic identification of traffic jams has a crucial importance for traffic management. From the operational point of view, it allows the management to act as soon as the traffic jam is identified without the need of human reporting. From the more strategic point of view, it can be used to reduce congestion, increase safety and improve traffic forecasting accuracy.

Several types of data sources can be used in order to perform this identification. All of them rely on measuring the speed of the vehicles and comparing that speed with the typical velocity of the road during free flow state.

3.4 Traffic prediction

The problem of predicting the traffic state at a future time can be tackled by different means. Simulation has been one of the most popular ways especially due to its capability of giving prediction in cases that haven't been tested in real life beforehand. That process is introduced elsewhere in this document. Other option is to base the prediction on the history of data captured.

Data-driven traffic forecasting has been a research topic since the late 1970s. The first attempts at predicting traffic flows consisted mainly of time-series approaches with different techniques [33] [34] [35], as well as early explorations of Kalman filtering methods [36]. In any of the above, or in posterior attempts to predict traffic features, short-term prediction horizons were used. Since then, data availability, analysis methods, and computational capacity have evolved and grown remarkably, along with the interest of the research community in this field.

Nowadays, predicting traffic features [37] [38] based on data is one of the key components of Intelligent Transportation Systems (ITS). Data from road sensors (e. g. magnetic loop sensors) are available with fine-grained resolution. Forecasting methods have evolved at a regular pace; although a great share of the latest research contributions still relies on time-series analysis, there is also a wide focus on machine learning methods.

Mostly all the literature is focused on short-term prediction, despite the usefulness of long-term estimations for road administration purposes and/or for enhancing short-term models as an additional input feature [37] [39].

Van Arem et al. [40] explored applications of traffic forecasting to dynamic traffic management by analysing the overall process from an economical demand and supply approach. This review is concentrated on forecasting methodologies, performance evaluation techniques and real-world application examples.

In [37] a deeper review can be found. The authors considered three main aspects of traffic forecasting: scope of application, output specification, and modelling features. The first aspect refers to the kind of roads for which the prediction is made and the type of application the prediction will be used for (mainly Advanced Traffic Management System, ATMS, and Advanced Traveller Information System, ATIS). On the other hand, the output specification involves the prediction horizon and step concepts and elaborates on which traffic parameters should be considered for developing the predictive model. The authors also provide a profound analysis of the modelling aspects of traffic forecasting. A classification of prediction models is presented, comparing their characteristics and performance. The authors contribute a methodological workflow to select and tune the model parameters that has been extensively referred.

Although most of the reviewed literature is focused on road features forecasting, especially traffic flow or volume obtained from magnetic loops, travel time is an alternative variable to predict. A survey on travel time prediction was published [41], exposing the techniques and main drawbacks and difficulties to estimate this traffic feature. According to this study, in 2005 the main handicap for this specific traffic forecasting scenario was concluded to be the lack of data, which the authors proposed to overcome with simulation techniques. Nowadays, the widespread proliferation of GPS-enabled devices and connected vehicles that can supply floating car data, allows researchers to model and predict travel-times without resorting to simulation methods. A more recent review on this field [23], delves in the latest travel-time prediction from floating car data (FCD) sources, and highlights the relevance of forecasting this feature for ATIS and operational planning of mass transit. Concise work [42]on prediction modelling delved into understanding and examining forecasting models proposed by previous research works. The authors considered a naïve category (predictions based in historic values or an average of them) in addition to the parametric/non-parametric category, and classified traffic simulation as a parametric method that can be used for traffic forecasting.

The main comparison features were prediction horizon, scope of application, computational speed and accuracy, most of them aligned with previous surveys. Interestingly, network-wide predictions and a comparison method between simulation and the rest of models were first suggested as open challenges in this work. More recently, [43] shows an agreement on the latter, stating the difficulties to compare not only simulation to every other forecasting method, but all methods to each other.

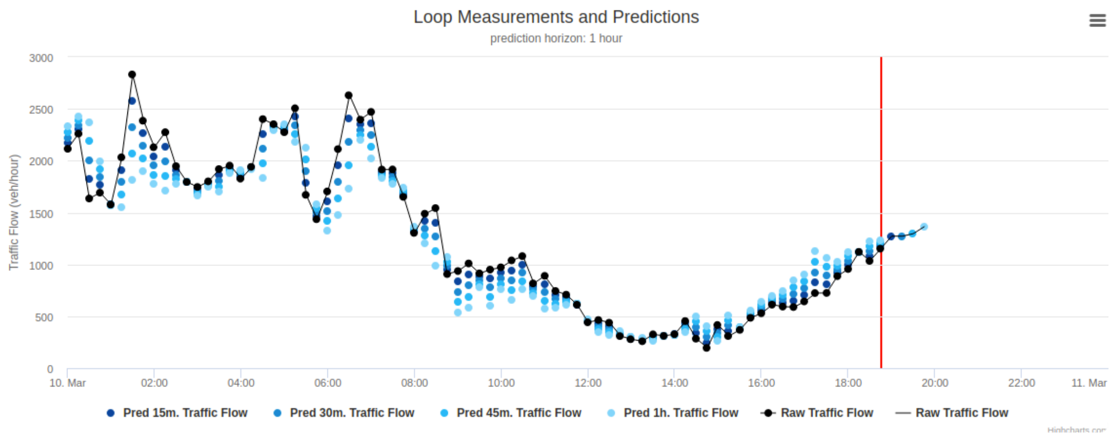

Figure 5: Image Random Forest application for predicting the traffic flux measure by a given magnetic loop. The prediction is performed at 4 different horizons (15, 30, 45 and 60 minutes). Provided by Tecnalia.

4 Data visualization methods

This section describes data visualization methods, motivation for visualizing data and evaluation of the methods. More specifically, we discuss visualization of traffic data and visualizations of other data types.

4.1 Motivation

Data Visualization plays an important role in decision-making processes allowing users to visually represent raw data and/or results of analysis. To visualise information, several methods and approaches can be followed each of which depends on specific requirements and technological stacks. Moreover, choosing the best visualization for data depends also on different factors, for instance, the target audience of the visualizations and/or the content of data itself.

4.2 Objectives

The main motivation for visualizing data is to increase the usability of the data and present it in an understandable and intuitive way. Since the project will make use of different heterogenous data types we provide a variety of methods to present different data in the most efficient way.

Concerning the visualizations, building on the above-mentioned main motivation, the objectives are to make them:

1. Understandable - meaning all stakeholders and target audience will be able to comprehend the data and use them. To achieve understandability different visualization methods will be used as most appropriate, aiming to be as clear and simple as possible without losing informativity.
2. Relevant - the visualised data will be of use and significant to the stakeholders and target audience. This will be achieved by identification of the most appropriate method automatically when possible and allowing users to manipulate parameters interactively.
3. Interactive - meaning the visualizations will be a reciprocal and participatory venture allowing for a two-way flow of information where appropriate in order to allow analysis of the data from different points of view as well as limiting the visualizations to relevant data only and experimenting with different methods and thus enabling data exploration.

4.3 Context

We consider the following data types:

- **Microscopic traffic simulation results** provide positions and statuses of specific vehicles over time, including their acceleration, position within traffic lane, air pollution and noise generated. From these results we can obtain further derivative data such as traffic flow or traffic density. These data provide a global view of the city-wide traffic conditions immediately. Due to high granularity of the data it can also be used for in-depth analysis of specific locations in the city.
- **Macroscopic traffic simulation results** provide more general results, namely traffic flow or density on roads, split by traffic modality. Some of the methods used for results of microscopic traffic simulation can be reused as described later.
- **Georeferenced data** can most effectively be visualised using combined methods that provide direct geographical representation as a map. Maps and examined overlay based visualization methods are described below.
- **Time series data** can be represented in many ways including the use of animated visualizations and graphs or charts that include a time axis. Combined time series georeferenced data can also be represented as an animated map.
- **Tabular data** can be visually presented using many methods. General methods for visualization of such data are discussed below, as are some of the methods that are more targeted towards certain relevant contexts.

Visualizations discussed are categorised into the following categories:

- **Map based methods** are based on a geographical map of the city area and allow the geo-spatial component of the data to be intuitively expressed. A common element of these methods is the city map, over which other information can be layered. Elements of the overlay include heat maps for visualization of traffic congestion and high density pedestrian crowds
- **Animation based methods** use animation of the graphics to express the time axis, providing intuitive understanding of changing states of the data over time. Examples of commonly used animated visualizations are animated satellite weather images, or in the context of the URBANITE project the changes in road usage during a workday.

- **Animated map based methods** use animated overlays over the city map, combining the advantages of map based methods (intuitive expression of spatial relations) and animation based methods (intuitive expression of time).
- **Other methods** include commonly used charts and graphs that are highly understandable due to common usage. Some of the data that will be presented using classic methods are sizes of subpopulation (using pie charts), trends (using line charts and bar charts) and composition of populations (using hierarchical trees).

4.4 Implementation methods and tools

The possible visualizations that can be built, as already mentioned, strictly depend on the chosen technological stack, starting from JavaScript (JS) libraries to complete data analysis frameworks. Some (non-exhaustive) JavaScript libraries to build visualisations are:

- **Chart.js[2]:** a JS library to build and configure visualizations (e.g. Bar charts, Line charts, Pie charts, etc.); it is an open source project released under MIT License.
- **D3.js[3]:** (D3 stands for Data Driven Documents) a JS library for realising interactive data visualizations (**¡Error! No se encuentra el origen de la referencia.**); it offers functionalities for DOM (Document Object Model) manipulation within HTML Web pages, advanced data interactions, animations and geographic functionalities. It is released under BSD license.
- **Apache Echarts[4]:** a JS library, that includes visualization for statistics, multi-dimensional data and relationships. It is released under the Apache-2.0 License.
- **Leaflet.js[5]:** a JS library that provides functionalities to build interactive maps. It is released under the BSD 2-Clause "Simplified" License.

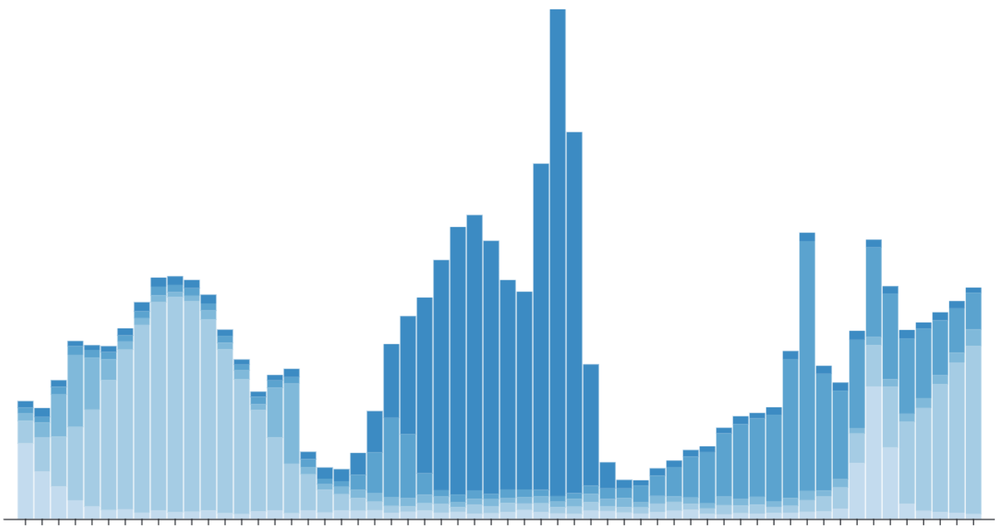

Figure 6: D3.js, Stacked-to-Grouped Bars example[6]

[2] https://www.chartjs.org/
[3] https://d3js.org/
[4] https://echarts.apache.org/
[5] https://leafletjs.com/
[6] https://observablehq.com/@d3/stacked-to-grouped-bars

Analytical frameworks in many cases follow a freemium approach, providing the user with a free version with limited capabilities and a premium version with all the functionalities. Examples of data analysis frameworks that also provide data visualization capabilities are:

- **Kibana[7]:** part of the "ELK" Stack[8] (Elasticsearch, Logstash, and Kibana) representing its data visualisation component; Kibana allows to build visualization taking advantage of the analytical features offered by Elasticsearch. It is possible to build interactive dashboards, filters and to share the dashboard creating a direct link or embedding the dashboard in a web page.
- **Grafana[9]:** this framework allows to query, visualise and explore data from different data sources. To interact with a specific data source the user should configure a plugin and build queries following the data source's specific language. The framework provides capabilities to build dashboards (**¡Error! No se encuentra el origen de la referencia.**) that can be shared or embedded into web pages.

Figure 7: Grafana, Grafana Play Home example[10]

Starting from one or more tools and libraries, a dashboard can be built combining different visualizations, charts and graphs. It is possible, of course, to build dashboards starting from the single visualization built through JS libraries or taking advantage of the built-in functionalities provided by the analytical frameworks.

In the context of URBANITE project, the central access point for the data visualizations is the URBANITE-UI, described in URBANITE deliverables "*D5.3 Integration Strategy*". This component will be used by developers to build their custom visualization taking advantage of JS libraries and/or embedding external dashboards built with analytical frameworks. It is important to underline that URBANITE-UI is not intended to be an analytical framework, but a tool that provides a uniform point of access to the project's technical tools and visualization developed into the different WPs and tasks. Moreover, providing as one of the integration approaches to

---

[7] https://www.elastic.co/kibana
[8] https://www.elastic.co/what-is/elk-stack
[9] https://grafana.com/
[10] https://play.grafana.org/d/000000012/grafana-play-home?orgId=1&from=1605104298090&to=1605107898091

embed visualization and/or dashboards into the UI, does not limit partners to use a specific framework to build their visualizations and/or dashboards.

4.5 Data visualization strategy for urban planning

Data visualization methods covered are chosen with currently recognised use cases in mind and discussed from a thematic point of view.

4.5.1 Noise Pollution Analysis

4.5.1.1 Strategic noise map

In order to accurately visualise noise pollution data, (strategic) noise maps are used. These are a graphical representation of the sound level distribution and the propagation of sound waves in a given region, for a defined period. [44] A strategic noise map, furthermore, is a graphical representation of the predicted situation with regards to noise in a particular area with different colours representing different noise levels in decibels [dB(A)]. [45]

Noise maps are based on two main noise indicators: Lden and Lnight. Lden is the day-evening-night noise indicator, representing the noise indicator for overall annoyance. It takes into account higher levels of noise pollution in the evening and night periods. Lnight is the night time noise indicator and is used in the assessment of sleep disturbance [45].

In order to develop a noise map, a noise calculator is needed, such as the one proposed by Farcas and Sivertun [46]. There are also other models for generating the maps, some of which using empirical methods, while most are based on the physics of propagation of sound outdoors.

The quality of noise mapping is dependent on the input data, which can be generated with the knowledge of environmental acoustics and use of noise pollution/sound propagation models (for instance, the Nordic propagation model Nord2000) [47]. In most cases those rely on an adaptation of the Geographic Information System (GIS) to provide effective data. The results of interpolation techniques, such as the inverse distance weighting (IDW) interpolation can then be visualised [48].

4.5.2 Air pollution analysis

4.5.2.1 Air Quality Index (AQI) map

There are numerous visualization methods used for air pollution analysis. Ranging from S-shaped parallel coordinates, a weighting map, used by Qu et al, to multi-dimensional views, network visualisations and use of splattering by Liao et al., and two-dimensional diagrams applied by Li et al. [49]

Lu et al. propose a visualisation method meant for large air quality monitoring data, however for smaller scale analysis (such as those done on individual cities) [50], the most effective map visualisation method is a map.

4.5.3 Bicycle traffic congestion analysis

4.5.3.1 Congested section distribution via line map

Since traffic congestion (bicycle based or other vehicle based) occurs in the road network, the most appropriate method of presentation is a visualisation of traffic congestion on the road network, since it provides analysis for spatial distribution and identification of traffic congestion patterns. Traffic congestion level is assigned a class, distinguished by a specific colour. Temporal changes can thus be observed by the colour change of the segment. A dot marker can additionally be used for identifying heavily congested intersections [51].

Figure 8 Example of visualization of traffic jams using darkened road streches and dots.

4.5.4 Mobility pattern identification

4.5.4.1 Flow map based on OD data

For many use cases in this type of context, information is stored in the form of Origin-Destination (OD) data, which represents aggregated movement count between pairs of geographic locations. [52] To visualise such data, Origin-Destination Matrices are used most frequently. However, in order to present spatial data on a map, flow map visualisations are more suitable. Each location (point or surface) can be either an origin or a destination. The interaction data between the origin and the destination is usually expressed by a line that connects the geometric center of the unit, while the width or colour (or opacity) of the line indicates the flow direction value between the origin and the destination. [53] Similarly, as with congestion analysis, the temporal change can be observed through the change in an attribute of the line, e.g. its opacity.

Figure 9: Air Quality Data Collected at Outdoor Monitors Across the US [53]

4.5.4.2 Chord diagram

When it comes to data for specific districts, flow maps are a viable visualisation. However, a chord diagram could be an additional one. In such a diagram, the flow between two nodes is represented by an arc, whose size (width) at an entity is proportional to the importance of the flow. To add a spatial dimension, each district or neighbourhood can be given a distinct colour in the chart, as well as on the map of the city.

4.5.5 Use of public transport analysis

4.5.5.1 Network distribution and heat dot map

Public transit lines take the form of a network; thus, a suitable visualisation method is a network distribution map. Different lines are coloured differently for easier visual distinction, while the width/thickness of the line indicates the volume of moving traffic, based on traffic count. Thus, the temporal and spatial information about the prevalence of specific public transit line routes is identifiable. Additionally, individual stations can be visualised with a multivariable heat distribution dot map, where each dot marker represents a station, while its colour visualises the traffic volume at the station at a certain point in time. With these two methods, public transport traffic on roads as well as at individual stations is presented for further analysis.

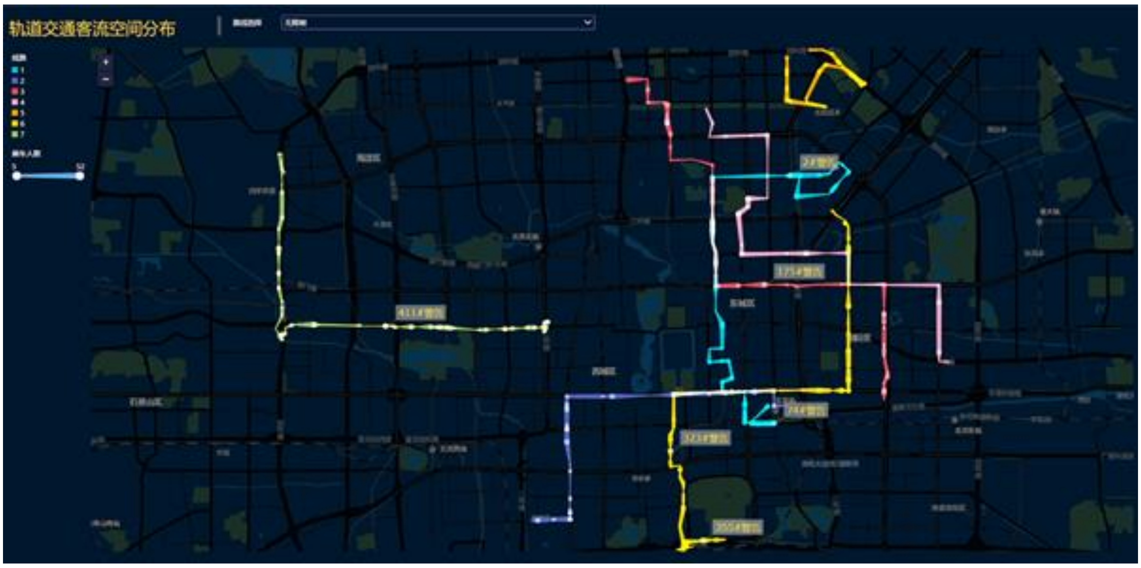

Figure 10 Example of traffic distribution and heat dot map

4.5.6 Public transport delay and traffic congestion analysis

4.5.6.1 Congestion and delay distribution via line map

Similarly, as with the use case of bicycle traffic congestion, public transport delay and congestion can be visualised with a distribution line map. The level of delay at a certain road section is represented by a class (e.g. major delay, minor delay, no delay), and the line is appropriately coloured based on the class. The level of congestion can be visualised in the same way. Heavily congested intersections may be presented with a marker.

4.5.7 Accident analysis

4.5.7.1 Visualisation method: Hotspot distribution heatmap

Common locations of traffic accidents can be presented with the use of a heatmap, where areas with high density of collisions are coloured usually red, while areas with low density of collision get assigned a light blue colour. Thus, identification of points with greater volume of incidents i.e. most common location of incidents, is clearly visually presented.

In the case of showing temporal information, traffic incidents at a certain point in time can simply be visualised with a marker at the location of the collision.

4.5.8 Potential traffic congestion and scheduled events

4.5.8.1 Kernel density map and point map

Potential traffic congestions can be presented in the same way as concrete congestions - with a line distribution map, where each road section is coloured appropriately based on the congestion level. However, if the question is about the probability of congestion and accidents, Kernel density maps could be used, too. In such maps, the Kernel density estimation - an estimation of the probability density function of random variables, is visualised with heat mapping. The resulting visualisation is a heatmap showing line sections with greater probability of traffic congestions and incidents.

Scheduled events disturbing traffic flow could simply be visualised with markers at the point of the event, while their consequence is appropriately presented in the traffic flow or congestion visualisation.

4.5.9 Traffic flow analysis

4.5.9.1 Point map

A common way of tracking and modelling traffic flow of cities is with ‘events’ or individual trips taken by a person/agent. To show the temporal and spatial information these events are then visualised on a point map, where each moving dot represents one trip, coloured depending on the type of vehicle used (car, bike, bus, train, etc.). This visualisation method is suitable for density analysis, as well. However, a better presentation of the volume of traffic at certain locations in space, is based on density mapping of trajectories. The resulting heat line map thus shows the density of the trips taken on a specific location of the road, colours distinguishing between the density levels.

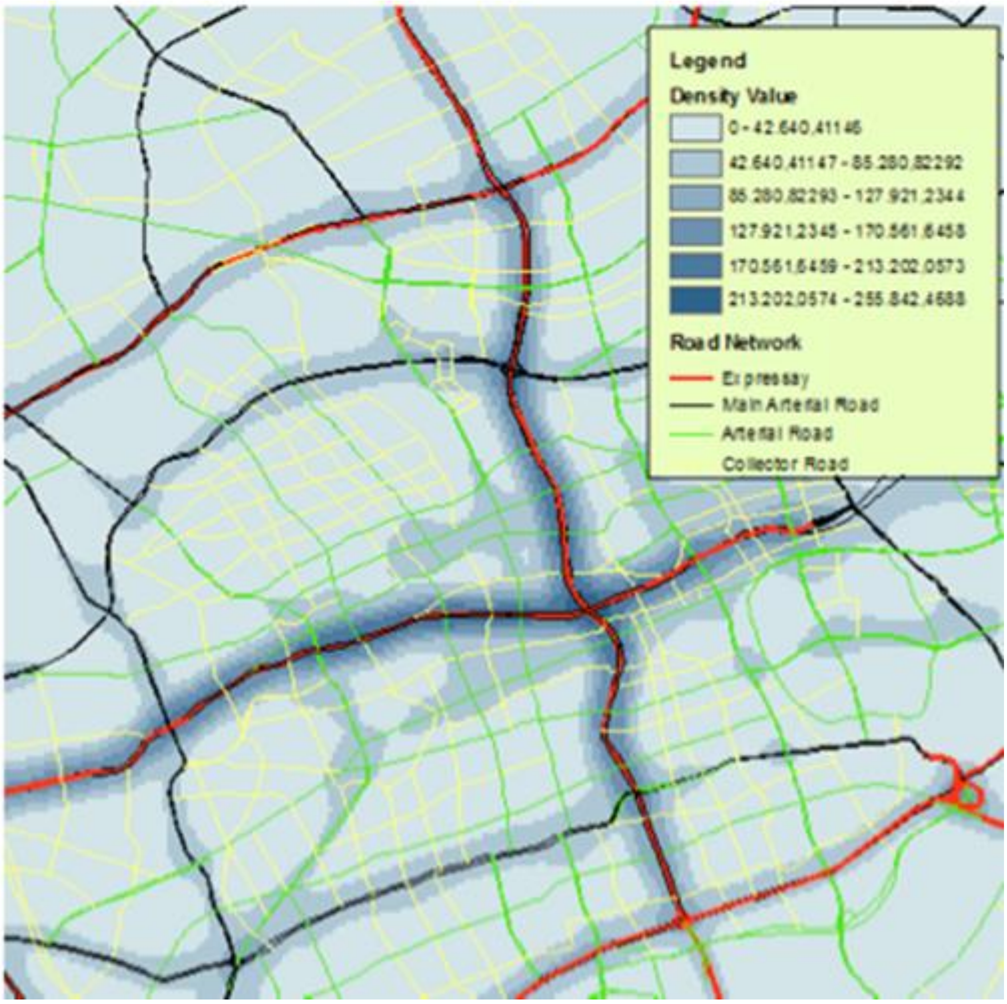

Figure 11 Example of a point map

5 Actionability of algorithms

Based on previous experience in the context of Intelligent Transport Systems [54], we find that aspects as confidence, accuracy, among other non-functional properties are essential for the predictive and analytical techniques be practical in use. In this section, we introduce the term *Actionability*, as the characteristic that any system based on data-analytics or artificial intelligence must present to be successfully deployed and used in a real operating environment. This term includes features like confidence, explainability, sustainability, adaptability, robustness, stability, transferability, applicability and usability, for such data-based methods to yield insights of practical value, for decision-makers. Below, such features are introduced in the context of urban mobility planning and will be desirable for the URBANITE solution.

In most of the cases, the data-driven models are often subject to uncertainty, involving stochastic non-deterministic processes, both on the learning mechanisms, execution or training/input data, also present on the results. During the design and experimentation phase, it will be possible to contrast the results with the opinion of an expert in a supervised way. However, once deployed, it is essential to provide an objective measure of the reliability and precision of the results, gaining in terms of *Confidence*. It is not frequent to find in a specific way the treatment of this aspect in academic research, mainly since such a measure is usually challenging to obtain and the estimation procedure usually model specific. Additionally, it requires initial statistical analysis of input data to understand its variability and characteristics properly [55] [56].

The need to explain and to do *interpretable* the underlying analytical models is undoubtedly one of the research fields with the greatest impact, being considered under the concept of Explainable Artificial Intelligence (xAI). This field of study comprises different techniques and methods, taking into account three fundamental factors: the nature of the model to be explained, from inherently transparent to completely opaque and unintelligible; the user of the algorithms; and finally, the way in which such explanation is to be prepared and presented to the decision-maker, which will depend on their degree of knowledge, as well as the intrinsic possibilities offered by the model to be explained in one way or another. A complete taxonomy and analysis of recent contributions related to the *Explainability* of different Machine Learning models are published [57], showing the difference between interpretable models and model interpretability techniques, transparent models and post-hoc explainability [58]. In fact, going into more in-depth detail, three levels are contemplated: algorithmic transparency, decomposability [6] and simulatability [58] [59] [60] and the post-hoc techniques we may distinguish among text explanations, visualizations, local explanations, explanations by example, explanations by simplification and feature relevance.

URBANITE's planning context is not as sensitive as other operational environments. Still, in any case it is important to ensure that data-based models can be understood by non-AI expert and can appropriately trusted on the inclusion of such models in their decisional processes, providing reliable information and arguing the different steps, treatments and sensitivity of the results to uncertainty in the input. In general, as we have commented previously, one of the usual problems is the qualitative jump in terms of performance and non-functional properties of the methods, between the experimentation phase and the actual deployment in operation. It is crucial to establish a trade-off or balance in terms of usability and accuracy in the processes and predictions so that we ensure their practical application. While for Intelligent Transport Systems, especially in traffic management and control, the need for real-time reaction is evident; in planning it is not so critical, since the information has been previously stored, with the possibility of executing the calculations offline. However, it is still necessary to respond in an adequate time to the calculations of KPIs and the presentation of diagrams and synoptics to decision makers.

By the term of *Self-sustainability*, we refer to the ability of the models to adapt to a dynamic environment subject to change, so that the system can remain in time, without the need for continuous adjustment and maintenance; in the case of models, an offline adjustment or re-learning of the model. In the case of mobility, the context is intrinsically dynamic, subject to different trends or exogenous variables, such as socio-demographic (including population displacement, economic organization, migration), technological or social, which generate alterations over time and space. The changes and expected impact on the model, to be taken into account, deploying particular designs and methods, must be analysed in detail for each specific application.

Adaptation is the reaction of a system, model or process to new circumstances, with the idea of maintaining its performance or reducing its loss, compared to the ideal conditions that were taken into account in its design and initial adjustment. The main problem in scenarios whose underlying phenomena changes over time, without being addressed by the model itself, is that the conclusions, predictions or categorizations will not be reliable. This phenomenon is called concept drift [61] [62] [63]. Thus, learning in nonstationary environments requires adaptive or evolving approaches that can monitor and track the underlying changes and adapt the model to accommodate those changes accordingly. From the perspective of urban planning, these changes will refer to social-economic reasons, external events (e.g. Covid-19 pandemic impact in traffic and mobility in general) and emerging vehicle/transportation technologies and modes innovation.

Robustness refers to the ability of a system to maintain service before external incidents. In the case of urban planning, it will not be so critical, since the decisions to be made will not be made in real time; However, the ingestion of data from the different datasets and warehouses available, if it must be operational, to minimise the loss of input data, as well as being robust data algorithms in such loss situations, fluctuations in frequency of them, low quality data, etc, [64]. In URBANITE project, data quality is addressed explicitly by the data curation-associated components deployed.

Stability means to ensure no surprises to the user in terms of functionality. Beyond the capability of algorithms to adapt to external environment changes (Adaptability), once exceeded statistical stability limits, it is necessary to restore the system, to minimise the impact of external changes on the quality of its output [65] [66] . Usually, the algorithms are worked in a specific geographical area and according to the available datasets. However, for their deployment in a real environment, it is necessary to project them to larger areas and volumes of data. This issue must be taken into account from the design stage of the algorithms, to optimise their algorithmic complexity, which represents the number of resources (temporary, execution time and space, required memory) that an algorithm needs to solve a problem. This characteristic allows determining the efficiency of this algorithm, not in terms of absolute measures but measures relative to the size of the problem. Currently, the availability of new technologies and paradigms of parallel and distributed processing of massive volumes of data, e.g. batch processing (map reduce), stream computing (Spark), hybrid (lambda architectures), Edge / Fog computing or using advanced hardware as GPUs / Physics, allows an escalation of the methods, obtaining adequate response times. However, its exploitation requires the adequate implementation and tailoring of the algorithms according to the architecture in which they will be deployed, as well as optimizing this deployment of analytical workloads across the different layers [67], by evaluating their computing resources. Other criteria when assessing the slavery of algorithms are the easiness to introduce new variables, the complexity of tuning if applies, and the execution time and the adequacy to migration to previously mentioned advanced architectures.

Another key feature is the *Compliance of new methods with Transport Engineering*. Existing traffic and mobility engineering practices are well established, with a powerful knowledge base. They have for a long time leveraged the use of information technologies, including simulation tools, but it is time to maximise the potential of data analytics and other data-based technologies throughout the engineering process. We find in general, that existing data-driven traffic models have been implemented as a proof of concepts [68], focusing on the techniques, instead of fitting the knowledge of traffic and transportation models, as related to traffic Flow analysis or transportation demand modelling. Here, a better understanding of hybridisation of simulation and data analytics methods, data-driven and model-driven approaches, by combining the strengths of each side, will help us to improve models by identifying more complex underlying assumptions and including additional aspects as those related to socio-demographic) and forecasting changes on the demand and supply sides, and their expected impact in terms of urban KPIs.

Usually, the models are obtained from specific datasets, regardless of whether they are real or synthetic. Therefore, the results are closely linked to the experiments carried out. *Transferability* is a desirable characteristic for algorithms and any model, to be able to present adequate performance and functionality in other contexts and starting data, different from those used in learning. The key in this aspect is the generation capacity of the methods, and the definition of the considered environment, so there will be calculations, easily extrapolated to other situations. In contrast, others are particular and sensitive to these contour variables and data of the specific use case. Examples of this behaviour can be found under the concept of soft sensing or virtual sensing [69] [70], where the measurements of a sensor can be estimated based on other sensors, for which it is necessary to add complexity to the model, to take into account those context variables. In particular, in the field of urban planning, this technique is applicable, by calculating posterior probability models, assigning greater relevance to those with acceptable average performance in many contexts.

From the *operational perspective*, the availability of large volumes of unprecedented data, available in the form of mobile phone traces, transactional information and from sensors deployed in the city, can be applied to transport planning and urban infrastructure, opening exciting possibilities to better understand flows and mobility guidelines and their complementarity with traditional data collection methods, allowing to have continue and more complete information. Currently, in most cities and regions, annual data consolidation is carried out with insufficient frequency to assist in the operation. The information is not exploited in a coordinated way and data sources are not merged, so it isn't easy to assess whether the information that is already available is sufficient or not. This point is especially relevant when planning urban mobility, where the coordination of the different departments, urban planning, mobility, environment and sustainability, must work aligned in their plans and regulations. If we move specifically to the Sustainable Urban Mobility Plan (SUMP) preparation. The availability of analytics and simulation tools to support initial diagnostics and the pre-evaluation of the different candidate actions has a significant impact on the decision process.

The ISO 9241-11 standard defines *Usability* as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in the specific context of use”, thus the completeness of goals, resources allocated and the acceptability of use. In practice, usability refers to the adequacy of the tool, in terms of interaction, process and result visualization to the decision maker skills. There exists different metrics for usability from the perspective of user interfacing [71] [72]. The System Usability Scale (SUS) [73] proposes general software usability measuring, most of them, dealing with the interaction process, the quality of the provided information and many other requirements as an intuitive design, ease of learning, memorability, etc.

Finally, we cannot forget the contextual aspects identified by the EU for the definition of policies, measures and sustainable mobility solutions, as part of the SUMPs and any supporting tool, aiming to contribute to urban regeneration, transport sustainability, social inclusion and social empowerment through active participation.

6 Conclusions

This deliverable presents a survey of existing methods for data modelling and visualizations that could apply to the URBANITE domain. Data modelling methods and tools support policy-makers to interactively explore available data, apply advanced data analytics and machine learning algorithms and visualise the results, simplifying their use for non-experts. There are many libraries and tools available, capable of handling different types of data and forms of visualization and interaction. Specifically, methods with application to the project use cases from a functional point of view are identified.

The document also identifies main issues and relevant results on the application of data analytics algorithms to some specific applications in urban planning, determined as appropriate by the project cities: mobility demand characterization, emission and noise estimation and traffic related analysis. Finally, it ends with the introduction of the actionability feature, a key property of any data-based modelling process to yield insights of practical value so that managers can harness them in their decision-making processes.

Given the existing references and options, the challenges are the selection of adequate tools and libraries ensuring the usability and applicability of such methods in a real planning scenario and process, and the specific research on new ad-hoc actionable algorithms for solving particular problems not previously addressed efficiently and/or managing the use case data-sources, in terms of availability and quality, including accuracy, completeness, frequency, volatility, etc.

7 References

[1] A. Esther, “State-of-the-art in artificial neural network applications: A survey,” Afribary.com: Retrieved November 30, 2020, from https://afribary.com/works/state-of-the-art-in-artificial-neural-network-applications-a-survey, 20118.

[2] H. Yin, “ViSOM - a novel method for multivariate data projection and structure visualization. doi: 10.1109/72.977314.,” *IEEE Transactions on Neural Networks,* vol. 13, no. 1, pp. 237-243, 2002.

[3] A. Peterson, “The Origin–Destination Matrix Estimation Problem - Analysis and Computations,” Dissertations, University of Linköping, 2007.

[4] A. Abderrahman and J. Eliasson, “Dynamic Origin-Destination-Matrix Estimation Using Smart Card Data: An entropy maximization approach,” *arXiv: Optimization and Control,* no. Corpus ID: 202538470, 2018.

[5] J. Doblas and F. Benitez, “An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix,” *Transportation Research Part B: Methodological,* vol. 39, no. 7, pp. 565-591, 2005.

[6] C. P. Marcela A. Munizaga, “Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile,” *Transportation Research Part C,* vol. 4, pp. 9-18, 2012.

[7] M. Trépanier, N. Tranchant and R. Chapleau, “Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System,” *Journal of Intelligent Transportation Systems,* vol. 11, no. 1, pp. 1-14, 2007.

[8] B. Barratt, R. Atkinson, H. Ross, S. Beevers, F. Kelly, I. Mudway and P. Wilkinson, “Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme,” *Atmospheric Environment,* vol. 41, no. 8, pp. 1784-1791, 2007.

[9] R. H. Keller, “A machine learning model of Manhattn air pollution at high spatial resolution.PhD thesis,” Massachusetts Institute of Technology. Department of Earth, Atmospheric, and Planetary Sciences., 2014.

[10] G. Ibarra-Berastegi, J. Saenz, A. Ezcurra, A. Elias and A. Barona, “Using neural networks for short-term prediction of air pollution levels. 498-502,” in *International Conference on Advances in Computational Tools for Engineering Applications*, 2009.

[11] I. Laña, J. Del Ser, P. A., M. Vélez and C. Casanova-Mateo, “The Role of Local Urban Traffic and Meteorological Conditions in Air Pollution: A Data-based Case Study in Madrid,” http://hdl.handle.net/20.500.11824/1184 (2016), 2016.

[12] J. Hopfield, “Artificial neural networks,” *IEEE Circuits and Devices Magazine,* vol. 4, no. 5, pp. 3-10, 1988.

[13] K. Wan and K. Lei, "Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau," in *2010 IEEE International Conference on Automation and Logistics, ICAL 2010*, 2010.

[14] I. Gonzalez-Aparicio, J. Hidalgo, A. Baklanov, A. Padro and O. Santa-Coloma, "An hourly PM10 diagnosis model for the Bilbao metropolitan area using a linear regression methodology," *Environmental Science and Pollution Research,* vol. 20, no. 7, pp. 4469-4483, 2013.

[15] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986.

[16] EPA, "Air Data: Air Quality Data Collected at Outdoor Monitors Across the US https://www.epa.gov/outdoor-air-quality-data," 2020.

[17] GSMA, "Quality Monitoring with IoT. Big Data A Technical Guide for Data Processing and Analytics," GSMA, 2018.

[18] DOEE, "Air Quality Data Assessment and Analysis. https://doee.dc.gov/service/air-quality-data-assessment-and-analysis," 2020.

[19] G. Kang, J. Zeyu Gao, S. Chiao, S. Lu and G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," *International Journal of Environmental Science and Development,* vol. 9, no. 1, 2018.

[20] Y. Zheng, L. Furui and H. Hsun-Ping, "U-Air: When Urban Air Quality Inference Meets Big Data. Microsoft," in *KDD'13, August 11–14*, Chicago, Illinois, USA, 2013.

[21] J. Shang, Y. Zheng, W. Tong, E. Chang and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proceedings of the 20th ACM SIGKDD international conference on Knowl edge discovery and data mining - KDD '14*, 2014.

[22] J. L. Reyna, M. V. Chester, S. Ahn and A. M. Fraser, "Improving the accuracy of vehicle emissions for urban transportation green house gas and air pollution inventories," *Environmental Science and Technology,* vol. 49, no. 1, pp. 369-376, 2015.

[23] P. Kumar, S. Nigam and N. Kumar, "Vehicular traffic noise modeling using artificial neural network approach," *Transportation Research Part C,* vol. 40, p. 111–122, 2014.

[24] S. Givargis and H. Karimi, "Mathematical statistical and neural models capable of predicting LA, max for the Tehran-Karaj express train," *Appl. Acoust,* vol. 70, no. 2009, p. 1015–1020.

[25] L. Bravo-Moncayo, J. Lucio-Naranjo, M. Chávez, I. Pavón-García and C. Garzón, "A machine learning approach for traffic-noise annoyance assessment," *Applied Acoustics,* vol. 156, p. 262–270, 2019.

[26] A. Torija, D. Ruiz and A. Ramos-Ridao, "Use of back-propagation neural networks to predict both level and temporal-spectral composition of sound pressure in urban sound environments.," *Build. Environ.,* vol. 52, pp. 45-56, 2012.

[27] D. Botteldooren, A. Verkeyn and P. Lercher, "Noise annoyance modelling using fuzzy rule based systems," *Noise Health,* vol. 4, p. 27–44, 2002.

[28] A. Torija and D. Ruiz, "Automated classification of urban locations for environmental noise impact assessment on the basis of road-traffic content," *Expert Syst. Appl.,* vol. 53, pp. 1-13, 2016.

[29] I. Abdic, L. Fridman, E. Brown, W. Angell, B. Reimer and E. Marchi, "Detecting road surface wetness from audio: a deep learning approach," in *23rd int. conf. pattern recognit..3458–63. https://doi.org/10.1109/ICPR.2016.7900169.*, 2016.

[30] S. Ø. Larsen, H. Koren and R. Solberg, "Traffic Monitoring using Very High-Resolution Satellite Imagery'," *Photogrammetric Engineering and Remote Sensing,* vol. 75, no. 7, pp. 859-869, 2009.

[31] L. Xu and Y. L. Yue, "'Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data," in *Procedia - Social and Behavioral Sciences 96:2084-2095*, 2013.

[32] I. Olabarrieta, A. Torre-Bastida and B. Molinete, "Autopistas Inteligentes. 6ª monografía ASEPA Sistemas de asistencia al conductor y de gestión inteligente del tráfico," 2014.

[33] M. S. Ahmed and A. Cook, Analysis of Freeway Traffic Time-series Data by Using Box-Jenkins Techniques, 1979.

[34] M. Levin and Y. Tsao, "On forecasting freeway occupancies and volumes," *Transportation Research Record,* no. 773, pp. 47-49, 1980.

[35] C. K. Moorthy and B. Ratcliffe, "Short term traffic forecasting using time series methods," *Transportation Planning and Technology,* vol. 12, no. 1, pp. 45-56, 1988.

[36] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transportation Research Part B: Methodological,* vol. 18, no. 1, pp. 1-11, 1984.

[37] E. I. Vlahogianni, J. C. Golias and M. G. Karlaftis, "Short−term traffic forecasting: Overview of objectives and methods," *Transport Reviews,* vol. 24, no. 5, p. 533–557, 2004.

[38] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transportation Research Part C: Emerging Technologies,* vol. 19, no. 3, p. 387–399, 2011.

[39] F. Su, H. Dong, L. Jia, Y. Qin and Z. Tian, "Long-term forecasting oriented to urban expressway traffic situation," *Advances in Mechanical Engineering,* vol. 8, no. 1, pp. 1-16, 2016.

[40] B. Van Arem, H. R. Kirby, M. Van Der Vlist and J. C. Whittaker, "Recent advances and applications in the field of short-term traffic forecasting," *International Journal of Forecasting, vol. 13, no. 1, pp. 1–12, 1997,* vol. 13, no. 1, pp. 1-12, 1997.

[41] H. Lin, R. Zito and M. Taylor, "A review of travel-time prediction in transport and logistics," in *Proceedings of the Eastern Asia Society for Transportation Studies, vol. 5, Bangkok, 2005, pp. 1433–1448*, 2005.

[42] C. van Hinsbergen, J. van Lint and F. Sanders, "Short term traffic prediction models," in *14th World Congress on Intelligent Transportation Systems*, 2007.

[43] E. Bolshinsky and R. Friedman, "Traffic flow forecast survey," Computer Science Department. Technion Tech. Rep., 2012.

[44] Wikipedia, "Noise map," in *En.wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Noise_map. [Accessed: 04- Nov- 2020]*, 2020.

[45] TII, "Noise Maps," *[Online]. Available: https://www.tii.ie/technical-services/environment/noise-maps/. [Accessed: 05- Nov- 2020],* 2020.

[46] F. Sivertun and Å. Farcaş, ""ROAD TRAFFIC NOISE: GIS TOOLS FOR NOISE MAPPING AND A CASE STUDY FOR SKÅNE REGION" [Online]. Available: https://www.isprs.org/proceedings/xxxviii/2-W11/Farcas_Sivertun.pdf. [Accessed: 05- Nov- 2020]".

[47] SINTEF, "Environmental Acoustics," [Online]. Available: https://www.sintef.no/en/digital/acoustics/environmental-acoustics/. [Accessed: 05- Nov- 2020], 2020.

[48] V. Poposki, K. Mitreski, D. Davcev and A. Naumoski, "NOISE POLLUTION MODELLING AND VISUALISATION – THE CASE STUDY FOR THE CITY OF SKOPJE[Online]. Available: https://stumejournals.com/journals/sbs/2016/2/48.full.pdf," *INTERNATIONAL SCIENTIFIC JOURNAL SCIENCE. BUSINESS. SOCIETY,* no. 2534-8485, 2020.

[49] P. Chen, "Visualization of real-time monitoring datagraphic of urban environmental quality," *", EURASIP Journal on Image and Video Processing,[Online]. Available: https://doi.org/10.1186/s13640-019-0443-6. [Accessed: 05- Nov- 2020],* vol. 2019, no. 1, 2020.

[50] W. Lu, T. Ai, X. Zhang and Y. He, "An Interactive Web Mapping Visualization of Urban Air Quality Monitoring Data of China," *Atmosphere,* vol. 8, no. 12, p. [Online]. Available: https://www.researchgate.net/publication/319116324_An_Interactive_Web, 2017.

[51] K. C. S. Djati, " Visual Exploration of Spatial-Temporal Traffic Congestion Patterns Using Floating Car Data," Master's, Technische Universität München, 2020.

[52] I. Boyandin, "Visualizing mobility data: the scalability challenge, [Online]. Available: https://medium.com/teralytics/visualizing-mobility-data-the-scalability-challenge-2575fe819702. [Accessed: 06- Nov- 2020]," Teralytics, 2019.

[53] L. Chou, "Top 10 Map Types in Data Visualization," Towards data science. [Online]. Available: https://towardsdatascience.com/top-10-map-types-in-data-visualization-b3a80898ea70. [Accessed: 06- Nov- 2020] , 2019.

[54] I. Lana, S.-M. J.J., E. Vlahogianni and J. Del Ser, "From Data to Actions in Intelligent Transportation Systems: a Prescription of Functional Requirements for Model Actionability," arXiv, 2020.

[55] E. Mazloumi, G. Rose, G. Currie and S. Moridpour, "Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction," *Engineering Applications of Artificial Intelligence,* vol. 24, no. 3, p. 534–542, 2011.

[56] C. I. van Hinsbergen, V. L. J. and H. Van Zuylen, "Bayesian committee of neural networks to predict travel times with confidence intervals," *Transportation Research Part C: Emerging Technologies,* vol. 17, no. 5, p. 498–509, 2009.

[57] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. arXiv:1910.10045," 2020.

[58] D. Doran, S. Schulz and T. R. Besold, "What does explainable AI really mean? a new conceptualization of perspectives. arXiv:1710.00794," 2017.

[59] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical. Society: Series B (Methodological),* vol. 58, no. 1, p. 267–288, 1996.

[60] Y. Lou, R. Caruana and J. Gehrke, "Intelligible models for classification and regression," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 150–158*, 2012.

[61] K. Kawaguchi, "Deep learning without poor local minima," *Advances in neural information processing systems, 2016, pp. 586–594.,* p. 586–594, 2016.

[62] A. Datta, S. Sen and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning system," in *2016 IEEE symposium on security and privacy (SP), 598–617*, 2016.

[63] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR),* vol. 46, no. 4, p. 44, 2014.

[64] J. Lobo, J. Del Ser, M. Bilbao and C. Perfecto, "DRED: An evolutionary diversity generation method for concept drift adaptation in online learning environments," *Applied Soft Computing,* 2018.

[65] S. Wechsler, H. Ban and L. Li, "The pervasive challenge of error and uncertainty in geospatial data," in *Geospatial Challenges in the 21st Century*, Springer, 2019, p. 315–332.

[66] M. D. Lara, "A mathematical framework for resilience: Dynamics, strategies, shocks and acceptable paths," *arXiv preprint arXiv:1709.01389,* 2017.

[67] J. Diaz de Arcaya, R. Miñón and A. Torre-Bastida, "Towards an architecture for big data analytics leveraging edge/fog paradigms. 173-176. 10.1145/3344948.3344987," 2019.

[68] I. Lana, J. Del Ser, M. Velez and E. I. Vlahogianni, "Road traffic forecasting: Recent advances and new challenges," *IEEE Intelligent Transportation Systems Magazine,* vol. 10, no. 2, pp. 93-109, 2018.

[69] H. Habibzadeh, A. Boggio-Dandry, Z. Qin, T. Soyata, B. Kantarci and H. T. Mouftah, ", "Soft sensing in smart cities: Handling 3vs using recommender systems, machine intelligence, and data analytics," *IEEE Communications Magazine,* vol. 56, no. 2, pp. 78-86, 2018.

[70] M. Arslan, "A Roadmap for Usability and User Experience Measurement during early phases of Web Applications Development, Master Thesis Software Engineering," BTH, 2010.

[71] A. Getachew and E. J. Obrien, "Simplified site-specific traffic load models for bridge assessment," *Structure and infrastructure Engineering,* vol. 3, no. 4, p. 303–311, 2007.

[72] W. Barfield and T. Dingus, Human factors in intelligent transportation systems, Psychology Press, 2014.

[73] J. Brooke, "Sus-a quick and dirty usability scale," *Usability evaluation in industry,* vol. 189, no. 194, pp. 4-7, 1996.

[74] M. L. a. Y. Tsao, "On forecasting freeway occupancies and volumes," *Transportation Research Record,* no. 773, p. 47–49, 1980.

[75] K. C. S. Djati, "Visual Exploration of Spatial-Temporal Traffic Congestion Patterns Using Floating Car Data," ", Master's, Technische Universität München, 2020.

[76] M. S. Cook and A. R. Ahmed, Analysis of Freeway Traffic Time-series Data by Using Box-Jenkins Techniques, 722., 1979.

[77] N. L. Nihan and K. Holmesland, "Use of the box and Jenkins time series technique in traffic forecasting," *Transportation,* vol. 9, no. 2, p. 125–143, 1980.