Università degli Studi di Torino
Department of Computer Science
Doctoral School in Sciences and Innovative Technologies

RESEARCH DOCTORATE IN COMPUTER SCIENCE
XXXI Cycle

Smart City management support: an ICT framework for vehicular traffic flow classification. The City of Turin case study.

Doctoral Dissertation of:
Mauro Giraudo

Tutor:
Prof. Luca Console

Supervisor of the Doctoral Program:
Prof. Marco Grangetto

Academic Year: **2018/2019**

Scientific Disciplinary Sector: **INF/01**

Acknowledgments

I would like to thank all the people who have positively influenced my path during these three years of the PhD course.

First, I would like to express my gratitude to people who assisted me in the research path. Thanks to professor Luca Console, for having accepted me as his student, and permitted me to start the doctoral studium. Thanks to professor Paola Pisano. My interest for smart cities and for application of ICT to real problems originates from her passion and engagement as Deputy Mayor for Innovation and Smart City in Turin. Without her help and intercession this thesis would never have seen the light. Thanks to professor Federica Cena, for her support during final steps of my path. Thanks for her availability, and for having spurred me, involving me in her courses.

I would like to thank the reviewers, professor Eleni Vlahogianni and professor Pierfrancesco Bellini, for their suggestions and for their support on reviewing my thesis.Their remarks are fundamental for improve this thesis.

Thanks to all the colleagues of my PhD cycle, Agata, Silvestro, Paolo and Manuel. Thanks to all the people I have known during these years.

Finally, special thanks to my family, my wife Daniela and my son Lorenzo, for their support throughout this period and for their interest in my work. Even if they are involved in exams to complete their own studies, they have endured my absences.

Mauro Giraudo
Torino
July 2019

Abstract

Smart City is a paradigm rapidly evolving which brings with it new ICT issues to address. Scholars described the smartness of a city as the ability to bring together all its resources, to effectively and seamlessly achieve the goals and fulfil the purposes it has set for itself. Public data are a pillar on which its development is founded, and are the basis for enabling real time decisions by stakeholders.

Torino As a Platform is the project that the City of Turin is developing to reach those goals: collecting data from IoT infrastructure, application, utilities, companies and citizens reports to build a data driven decision making platform that will support governance of the smart city.

The project realization bases its development on two main activities: on one side realizing and delivering to communities an SDK with standard API for increasing and facilitating collection and reuse of data and citizens participation. On the other side, developing a machine learning framework with predictive algorithms that helps the governance of the Smart City and visually supports the decision making process.

Inside this second activity, we studied a system to analyse temporal trends of traffic flow within the boundaries of the city, and make predictions about its status in short, medium and long period (15 minutes, 1 hour, and 2 hours).

City governance can use output of these predictions to implement correct actions to address traffic issues.

Following previous work, we use a combined approach to analyse historic data of vehicular traffic flow from the City of Turin: we test different combinations of methods to develop a classifier framework for traffic flow.

The framework analyses new upcoming data in real time, assigns the event to a specific class, and visualizes it in an info-system to help responsible offices to make decision about actions to be taken.

As next step to improve the system, we realize a predictive model about traffic flow status in short, medium and long period, with the aim to show future (possible) problematic events and permit to city governance the treatment of issues in advance.

Future work and further improvement is the integration of the system with a database of actions taken in traffic issues treatment, with analysis of impact by studying the deviation of real flow from the predicted one.

Aim of this integration is to build a knowledge base of possible actions that can be taken in traffic issues with a grade of awaited result based upon contexts and applications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Anthopoulos (2017) described Smart City as a paradigm rapidly evolving which brings with it new ICT issues to address. In (Int, 2014) the International Standards Organization (ISO) described the smartness of a city as the ability to bring together all its resources, to effectively and seamlessly achieve the goals and fulfil the purposes it has set for itself. Public data are a pillar on which its development is founded, and are the basis for enabling real time decisions by stakeholders.

A city can be defined as 'smart' when investments in human and social capital, in traditional transport and modern communication infrastructure point to efficient and sustainable development. Smart Cities are the result of a dynamic process, which develops along six dimensions: smart economy, smart people, smart mobility, smart environment, smart living and smart governance. The day-to-day management of the Smart City activities involves many decisions at the strategic, tactical and operational level. The managers use the facilities of Decision Support Systems (DSS) for complex decision-making (Chichernea, 2014).

Torino As a Platform is the project that the City of Turin is developing to reach the following goal: collecting data from IoT infrastructure, application, utilities, companies and citizens reports to build a data driven decision making platform that will support governance of the smart city.

The project realization bases its development on two main activities: on one side realizing and delivering to communities an SDK with standard API for increasing and facilitating collection and reuse of data and citizens' participation. On the other side, developing a machine learning framework with predictive algorithms that helps the governance of the Smart City and visually supports the decision making process. In Figure 1.1 we report the big picture of the project.

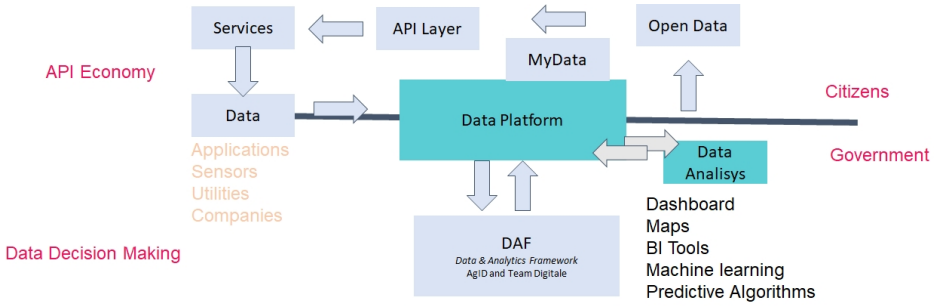

Figure 1.1: TAaP full diagram

Inside the second activity of the project, we studied a system to analyse temporal trends of traffic flow within the boundaries of the city, classify the data and make predictions about its status in short, medium and long period (15 minutes, 1 hour, 2 hours).

As final objective, the result of our work is an InfoVis system that output traffic flows status and forecasts, with the aim of support the City governance to implement correct actions to address traffic issues.

The remainder of this thesis is organized as follows.

The chapter 2 is an overview of previous work on vehicular traffic flow modelling and classification.

In chapter 3, we describe the proposed approach to the problem of traffic flow classification and forecasting.

In chapter 4, we describe the case study of City of Turin, its peculiarity, issues and the datasets used.

In chapter 5, we describe the framework proposed, its development and implementation.

In chapter 6, we analyse the results obtained by the framework in term of classification, prediction and output visualization.

Finally, chapter 7 summarizes the findings of the work and open to new research question.

Chapter 2

Related Works

Analysis of traffic flow is a long time life research theme, started when Lighthill and Whitham (1955) presented a model based on the analogy of vehicles in traffic flow and particles in a fluid. Since then, mathematical description of traffic flow has been a lively subject of research, with studies focused on two main aspect: traffic modelling and traffic forecasting.

2.1 Traffic modelling

Hoogendoorn and Bovy (2001) made a good state-of-the-art analysis of the previous works on traffic modelling, focusing their attention on the nearly fifty years of traffic flow theories and models.

All the studies analysed are mathematical models build to represent the vehicular traffic flow trends in time and space.

They classified the discussed traffic models according to the following:

- Scale of the independent variables (continuous, discrete, semi-discrete).
- Level of detail (submicroscopic, microscopic, mesoscopic, macroscopic).
- Representation of the processes (deterministic, stochastic).
- Operationalisation (analytical, simulation).

- Scale of application (networks, stretches, links, and intersections).

For our studies we put particular attention to: *i)* level of detail and *ii)* scale of application.

Hoogendoorn and Bovy propose the following classification of traffic models according to the level of detail with which they represent the traffic systems:

1. Submicroscopic simulation models (high-detail description of the functioning of vehicles' subunits and the interaction with their surroundings).
2. Microscopic simulation models (high-detail description where individual entities are distinguished and traced).
3. Mesoscopic models (medium detail).
4. Macroscopic models (low detail).

A microscopic simulation model describes both the space-time behaviour of the systems' entities as well as their interactions at a high level of detail.

Similar to the previous, the submicroscopic models describe the characteristics of individual vehicles in the traffic stream. However, apart from a detailed description of driving behaviour, also vehicle control behaviour is modelled in detail.

A mesoscopic model does neither distinguish nor trace individual vehicles, but specifies the behaviour of individuals, for instance in probabilistic terms.

Macroscopic flow models describe traffic at a high level of aggregation without distinguishing its constituent parts. For example, the traffic stream is represented using characteristics as flow-rate, density, and velocity.

Regarding the scale of application, they identified it as the area of application of the model. For instance, the model may describe the dynamics of its entities for a single roadway stretch, an entire traffic network, a corridor, a city, etc...

Hoogendoorn and Bovy concluded that macroscopic models are suitable for large scale, network-wide applications, where macroscopic characteristics of the flow are of prime interest. They are very suitable for application in model-based estimation, prediction, and control of traffic flow.

2.2 Traffic forecasting

On the other side, traffic forecasting, Vlahogianni et al. (2014) presented their analysis of existing works in short-term traffic flow forecasting.

The combination of unprecedented data availability and the ability to rapidly process these data has brought on immense development and acceptance of ITS technologies.

At the same time, a new research area, based on data driven empirical algorithms, has been systematically growing in parallel to the well-founded mathematical models that are based on macroscopic and microscopic theories of traffic flow.

The field of short-term traffic forecasting has a long life; in the first part of its development, most of the research employed statistical approaches to predicting traffic at a single point. Later, applications of data driven approaches were the focal point in the literature.

The weight placed recently on empirical computational intelligence-based approaches, including Neural and Bayesian Networks, Fuzzy and Evolutionary techniques, can be considered as inevitable.

Particularly, most classical approaches have been shown to be 'weak' or inadequate under unstable traffic conditions, complex road settings, as well as when faced with extensive datasets with both structured and unstructured data (Vlahogianni et al., 2014, p. 2).

Vlahogianni et al. clarified that most effort in previous studies has gone into: *i)* using data from motorways and freeways, *ii)* employing univariate statistical models, *iii)* predicting traffic volume or travel time, and *iv)* using data collected from single point sources.

Their work was based on a set of ten challenges stemming from the changing needs of ITS (Intelligent Transportation System) applications.

Vlahogianni et al. concluded that researchers seem to be unprepared to answer two important questions: are they confident that new models are better, in terms of accuracy, than models developed in the past? Moreover, what have they learnt about prediction that has significantly changed the perception for traffic operations and management?

The above imply that both research and practice in short-term traffic forecasting are entering a maturity phase, where models and methods must be critically assessed to produce solid knowledge on the concepts and processes

involved with short-term traffic forecasting.

In this direction, a step into the future is towards enhancing the performance and explanatory power of the prediction models through synergies with classical statistics.

Statistics and artificial intelligence should act complementarily to improve *i)* core model development and goodness of fit, *ii)* analysis of large data sets and *iii)* causality investigation (Vlahogianni et al., 2014, p. 14).

2.3 Studies with particular approach

In the remainder of this chapter, we analyse some scholars' works that approached issues similar to that we exposed in chapter 1.

Yu et al. (2013) pointed out that present study on urban road traffic condition classification recognition mainly focuses on two aspects: one is road traffic condition classification, which emphasizes traffic real-time data classification; the other is traffic condition recognition based on prior classification, which lays stress on the method of traffic condition recognition.

They proposed a system based on SVM (Support Vector Machine) pattern recognition algorithm by considering traffic multidimensional characteristics in urban road traffic, but they focused the study on the performance and accuracy of the method, not applying it to a real world case.

They based the study on three dimension: Traffic flow, Average speed, and Share ratio. According to the indicators, urban road traffic conditions have been studied by cluster analysis combined with existing theoretical research, but they didn't specify either, grouping in four class representing block flow, congested flow, steady flow, and smooth flow respectively.

Under those set classes, a simulation obtained average speed, and share ratio. They generated 50 samples per traffic condition, totalling 200 sets of observational data, among which 40 samples were training data, and 10 sets were detection data (Yu et al., 2013).

Petrovska and Stevanovic (2015) presented in their work an innovative visual tool with the aim of detecting and avoiding road traffic congestion, based on Google Maps data.

Their application focuses on urban roads congestion analysis, with links and nodes (intersections) enclosed. Besides the spatial aspect, the study pays

attention also to the temporal dimension of road link traffic by providing stored time of the day analysis data when the congestion level was highest.

They suggested that knowing that traffic on a particular road is congested, but not how congested it is, is not very much of help in making decision for the traffic managers or road users. Therefore, they proposed an alternative method to quantify actual road traffic congestion based on Google Maps Traffic Layer. The users of the tool developed are supplied with an interface to analyse traffic congestion data interactively, focusing on visualization methods to study and analyse traffic congestion on small parts of a network (Petrovska and Stevanovic, 2015, p. 1491).

Montazeri-Gh and Fotouhi (2011) used k-means clustering algorithm for traffic condition recognition, collecting data from an ad-hoc hardware device installed on a vehicle.

Their focus is on the application of driving and traffic condition information in an intelligent Hybrid Electric Vehicle (HEV) control strategy. In an advanced type of HEV control strategy, the controller adapts itself to the current traffic condition to reduce fuel consumption and exhaust emissions. Traffic Condition Recognition is a critical sub-system of this intelligent HEV control strategy.

Stathopoulos and Karlaftis (2003) work concentrates on developing flexible and explicitly multivariate time-series state space models using core urban area loop detector data, limiting their case study to a 3-lane per direction signalized arterial on the periphery of the core area of the city in a period of 5 months.

Using volume measurements from urban arterial streets near downtown Athens, models were developed that feed on data from upstream detectors to improve on the predictions of downstream locations.

They concluded noting that the multivariate modelling of flow, speed and occupancy data in urban areas is a complex and tedious process. Data from different detectors are not only highly correlated among themselves but are also related to prevailing traffic conditions, which tend to exhibit high short-term fluctuation. In addition, during large parts of the day, traffic is highly congested, approaching unstable conditions.

Thianniwet et al. (2009) proposed a classification system of road traffic congestion based on Decision Tree Algorithm and Sliding Windows, collecting from a notebook date, time, latitude, longitude, and vehicle velocity from GPS

and capturing images of road traffic condition by a video camera, on a test vehicle.

The vehicle passed through overcrowded urban areas approximately 30 kilometres within 3 hours. In the experiment, they gathered the congestion levels from 11 subjects with driving experience up to 10 years. They watched a 3-hour video clip of road survey and rated the congestion levels into three levels, light, heavy, and jam.

Then, the concluded congestion levels from 11 subjects were calculated using majority vote. The judged congestion levels were then synchronized with velocity collected by the GPS device.

Other works focused on particular condition and data, e.g. traffic accident analysis, traffic congestion or car sharing data (Shanthi, 2012; Sohn and Lee, 2003; Pecherková and Nagy, 2017; Vaniš and Urbaniec, 2017; Zhang et al., 2016; Pagani et al., 2017).

Chapter 3

Research approach: issues, context and innovation

In this thesis research we tried to study the following research questions:

1. How implement a classification framework that uses historical data of vehicular traffic flow;
2. Use of results of classification process for the prediction of future status of traffic flow;
3. Focus on the pratical case of a large city and on support to its management;

The research we approached in this thesis presents some peculiarities.

First, the context in which we are operating: we want to analyse the traffic flow inside the city boundaries, in a complex urban scenario. Every road has peculiarities, as different number of lanes, presence of traffic lights, intersections, different speed limits or other limitations, e.g. traffic permitted only in certain hours.

Second, we start the analysis from the data collected by city traffic management, in particular from loops present in city roads, in the last 3 years for

the historical database, and open data service, via XML web download, for real-time data.

Third, our objective is building a framework comprehensive of a model for classifying traffic flow, e.g. from fast to high, of a method to predict future class assignemnt, and of a system that visualises results of previous steps.

At the end, the result we want to produce is an InfoVis system that outputs the traffic status and its forecasts to support city traffic manager to address eventual issues.

In this context major issues we have faced in our studies are: *i)* managing the large amount of data, *ii)* building a correct training set for the classification algorithm, and *iii)* realizing a simple InfoVis system to be presented to public managers.

Point *i)* was partially approached as described in chapter 4, and, in conjunction with *ii)*, guided the first step in the framework's realization.

Indeed, the still high number of records remained after data cleaning phase, the high number of different stations where data were collected, and the differences in road's topology, made very difficult every manual labelling approach.

For every point, we had to analyse about 300000 records, by the two dimensions of speed and flow, to identify possible classes of traffic, and then label a sufficient number of examples to build a good training set.

So, we proposed a combined approach to analyse historical data of traffic flow: *i)* use of an unsupervised algorithm, like clustering methods, to label instance of traffic data, and *ii)* use of a supervised classify algorithm, trained by previously labelled dataset, to assign class to data instances.

With the first step, we simplified the labelling task: we left to the clustering algorithm the identification of coherent groups of speed and flow combinations, and we built with results obtained a training set with dimension of the whole data set.

We then used those training sets to build the classification model, testing two different algorithms.

We developed the framework of classification and forecasting using Weka tools for the machine learning algorithms, and a simple web based application for the info-vis system, realized with the standard AMP stack (Apache, MySQL, PHP).

3.1 Use of an unsupervised algorithm for labelling data instances

As unsupervised learning algorithm we used K-means clustering. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.

The algorithm takes as input the number of clusters and the data set, that is the collection of features for each data point. The algorithms starts with initial estimates for the centroids, which can either be randomly generated or randomly selected from the data set.

The algorithm then iterates between two steps:

1. Data assigment step: Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance.

2. Centroid update step: In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

The results of the K-means clustering algorithm are:

- The centroids of the K clusters,

- Labels for the training data.

Rather than defining groups before looking at the data, clustering allowed to find and analyse the groups that have formed organically. Examining the centroid values it's possible to qualitatively interpret what kind of group each cluster represents.

There is no method for determining exact value of K, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the

extreme of reaching zero when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point", where the rate of decrease sharply shifts, can be used to determine K. (MacQueen, 1967)

3.2 Use of a supervised classify algorithm to assign classes

After having labeled the training set with the algorithm previously described we tested two different classification algorithms.

3.2.1 Decision Trees

Decision trees are simple but intuitive models that utilize a top-down approach in which the root node creates binary splits until a certain criteria is met. This binary splitting of nodes provides a predicted value based on the interior nodes leading to the terminal (final) nodes.

In this classification context, decision tree outputs a predicted target class for each final node produced.

3.2.2 Naive Bayes

Naive Bayes is a simple technique for constructing classifiers: assign class labels to problem instances, the vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. (John and Langley, 1995)

3.3 Info-vis system

Obtained the classification framework, we used it to classify data gathered in real-time from the open data systemof the traffic management.

We proposed a simple web based service that show the places where data are collected, the classes of traffic flow assigned, and a forecast for next classes.

This service is oriented to support monitoring and decision of the traffic management officers of the city.

Chapter 4

Case study and dataset: City of Turin

In this chapter we describe the case study we adopted to apply the research approach presented in chapter 3, while in the next chapter 5 we describe the framework built.

4.1 Data sources

As previously described in chapter 3, we want to analyse the traffic flow inside the city boundaries, in a complex urban scenario. We chose the City of Turin as our case study.

We gathered historical data from '5T Consortium', a public society that manages public and private mobility services for City of Turin.

We received a dataset of about 40 million records, collected from more than 100 loops detector stations, disseminated in the city boundaries, covering a 3 years period from January 2015 to December 2017.

The principal traffic data we took in consideration consist in:

- mean speed (Km/Hr),
- flow (Veh/Hr),

- accuracy (percentage).

The data are calculated by the manager in 5 minutes aggregation, and accompanied by geolocation of stations (longitude and latitude), road information (name and POI identification), and direction of flow (positive or negative).

The accuracy parameter explains in percentage the data validity.

For the real time analysis, we gathered data from 5T OpenData web service, which publishs the same type of information in XML format at predetermined time interval (`http://opendata.5t.torino.it/get_fdt`).

Those data are available about 1 minute after the period considered (e.g., data from start time of '12:40:00' to end time of '12:45:00' have a generation time of '12:46:00').

In table 4.1 we report some examples of historical data dimension.

Table 4.1: Historical traffic data

Station	Longitude	Latitude	Road Name	Records
5	7.62745	45.01608	Corso Unione Sovietica	309019
93	7.64825	45.03448	Corso Unione Sovietica	243629
95	7.63681	45.02413	Corso Unione Sovietica	197759
4	7.62801	45.01641	Corso Unione Sovietica	197065

In table 4.2 we report some examples of basic statistics about historical data, i.e. mean, standard deviation, maximum and minimum values for both speed and flow.

4.2 Data cleaning

We did some preliminary operations on data: first, we excluded those that had accuracy equals to 0 and we marked for a more in-depth analysis those that had accuracy less than 75.

In the second cleaning phase, we took into consideration the numerosity of records in the period: starting from the 5 minutes aggregation, we expected a maximum of 105120 records for year, and a maximum of 315360 in the three years period.

Counting the sum of records per stations we excluded stations which had substantial gaps on data retrieving, following the policies: we must have a

Table 4.2: Basic statistic of traffic data

Station Feature	Min.	Max.	Mean	Std. dev.
Station 4				
Speed (Km/hr)	3	176	53.293	12.808
Flow (Veh/hr)	12	1968	394.527	89.612
Station 20				
Speed (Km/hr)	3	176	53.195	12.901
Flow (Veh/hr)	12	1392	274.632	97.669
Station 40				
Speed (Km/hr)	3	176	64.546	17.738
Flow (Veh/hr)	12	1380	289.964	98.075
Station 73				
Speed (Km/hr)	3	176	75.654	16.204
Flow (Veh/hr)	12	1140	168.993	41.975
Station 90				
Speed (Km/hr)	3	176	54.238	11.667
Flow (Veh/hr)	12	2040	356.991	87.134
Station 106				
Speed (Km/hr)	3	147	29.428	13.981
Flow (Veh/hr)	12	2136	513.801	120.157

number of records greater than 236520 in the three years period, and a number of records greater than 78840 in the last year (2017), both corresponding to a threshold of 75 percent of data expected.

We also took in consideration stations that had a coverage less than 75 percent in the overall period, but considering a focus on more recent data, we accepted those that had a coverage greater than 90 percent in the last year (at least 94608 records).

We excluded in this way stations that did not work for long period (e.g., for faults or road works), or started working only in 2017, but not those which had a high percentage of data in the last year considered.

As third cleaning step, we took in consideration the geographical distribution of the remaining stations: we excluded the stations that are isolated and selected those that are along main roads.

The remaining data guaranteed us good historical and geographical cov-

erage and we stored their data in a MySQL database; figure 4.1 shows the scheme adopted.

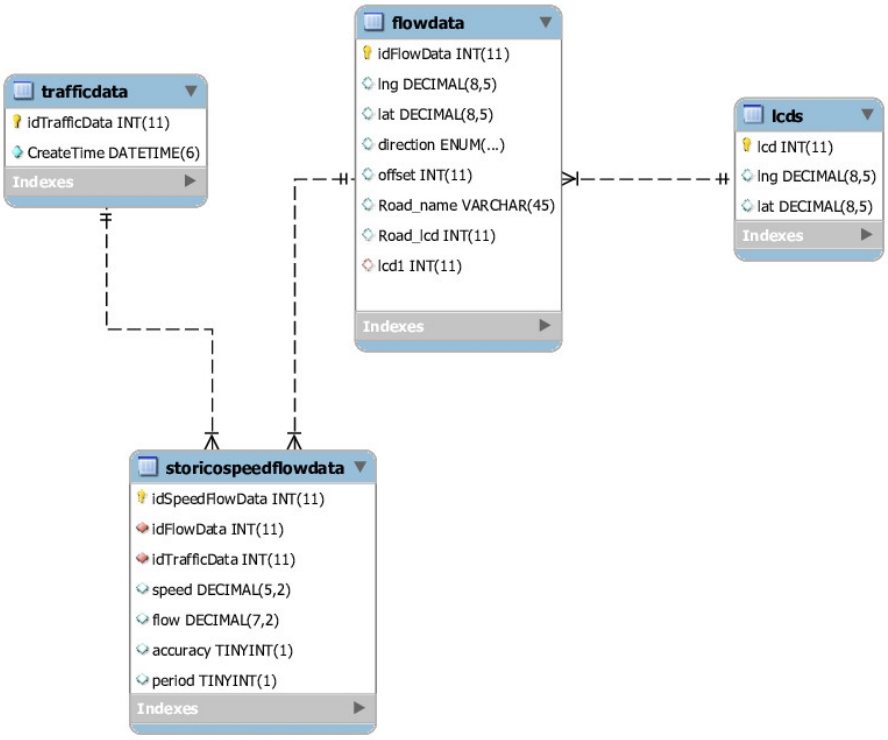

Figure 4.1: MySQL database scheme for storing data from 5T historical dataset

Final dataset consisted in about 15 millions of records from 46 stations; table 4.3 reports the final station selection and number of records, and figure 4.2 shows the distribution along the city road.

Figure 4.2: Final selected stations geolocation

Table 4.3: Final stations selection

Station	Longitude	Latitude	Direction	Road Name	1 year	3 years
49	7.64748	45.04104	negative	Corso Agnelli	100178	301655
68	7.64492	45.03767	negative	Corso Agnelli	91604	295523
48	7.64773	45.04129	positive	Corso Agnelli	99302	298993
106	7.63528	45.0233	positive	Corso Agnelli	83590	275874
84	7.69852	45.09947	negative	Corso Giulio Cesare	104078	281166
90	7.70485	45.109	negative	Corso Giulio Cesare	98716	275059
2	7.7091	45.11606	positive	Corso Giulio Cesare	90290	290352
10	7.70151	45.10447	positive	Corso Giulio Cesare	104297	291288
32	7.68658	45.08135	positive	Corso Giulio Cesare	96608	277135
79	7.69374	45.09244	positive	Corso Giulio Cesare	101263	301265
30	7.69599	45.08495	negative	Corso Novara	103444	310128
37	7.69045	45.08801	negative	Corso Novara	102589	308946
40	7.69273	45.08685	negative	Corso Novara	103986	179822
29	7.69749	45.0831	positive	Corso Novara	103605	296030
36	7.68849	45.08869	positive	Corso Novara	96581	223228
38	7.69109	45.08765	positive	Corso Novara	101773	308759
39	7.69368	45.08621	positive	Corso Novara	103168	178642
15	7.66632	45.08252	negative	Corso Regina Margherita	103093	300604
26	7.70375	45.07013	negative	Corso Regina Margherita	95231	246672
71	7.69882	45.07174	negative	Corso Regina Margherita	102511	291134
73	7.66165	45.08406	negative	Corso Regina Margherita	95723	293591
14	7.66692	45.0822	positive	Corso Regina Margherita	101827	296802
25	7.7046	45.06975	positive	Corso Regina Margherita	94280	243534
72	7.65943	45.08468	positive	Corso Regina Margherita	101330	297613
89	7.66461	45.08299	positive	Corso Regina Margherita	98694	287284
20	7.69161	45.07239	negative	Corso San Maurizio	99577	300314
96	7.69397	45.07017	negative	Corso San Maurizio	99378	274185
97	7.69814	45.06599	negative	Corso San Maurizio	101683	298535
19	7.69154	45.07289	positive	Corso San Maurizio	103770	309413
22	7.69858	45.06579	positive	Corso San Maurizio	101053	297328
56	7.66548	45.04981	negative	Corso Turati	79160	241859
60	7.67117	45.05501	negative	Corso Turati	98057	294389
122	7.66893	45.05299	negative	Corso Turati	84740	248921
55	7.66592	45.05004	positive	Corso Turati	79068	240471
65	7.66814	45.05216	positive	Corso Turati	79680	240336
121	7.66941	45.05323	positive	Corso Turati	89267	240806
5	7.62745	45.01608	negative	Corso Unione Sovietica	102635	309019
93	7.64825	45.03448	negative	Corso Unione Sovietica	78905	243629
95	7.63681	45.02413	negative	Corso Unione Sovietica	100573	197759
4	7.62801	45.01641	positive	Corso Unione Sovietica	101735	197065
94	7.64885	45.03463	positive	Corso Unione Sovietica	79374	244452
103	7.63267	45.0203	positive	Corso Unione Sovietica	84650	259580
105	7.63569	45.02298	positive	Corso Unione Sovietica	102279	303599
119	7.65156	45.03712	positive	Corso Unione Sovietica	98529	287489
1	7.66662	45.06766	positive	Corso Vinzaglio	99594	295989
54	7.66773	45.06927	positive	Corso Vinzaglio	86938	259271

Chapter 5

Framework

In this chapter we present the realisation of the framework presented in chapter 3, usiing as example data from a specific station, selected as test bed, and, where it is necessary, we make comparisons with some other stations. The overall discussion of results for the set of stations is reported in chapter 6.

5.1 First step: clustering for data labelling

As previously presented, our choice for building the training set was to use the K-Means clustering algorithm for identifying coherent groups of data, by the two dimensions of speed and flow.

The first operation we did in this step was the choice of the correct number of cluster to be used.

We analysed individual stations data using the elbow identification method, i.e. the analysis of the graph of the progress of the SSE Sum of Squared Error-based on the number of clusters.

The aim of this method is the identification, by visual analysis of the graph, of the breaking point where the trend smooths. The corresponding number of clusters can be used as a good candidate.

In figure 5.1 and figure 5.2 we report the graph of SSE vs number of cluster for two selected stations, and identify the possible elbow with a red arrow.

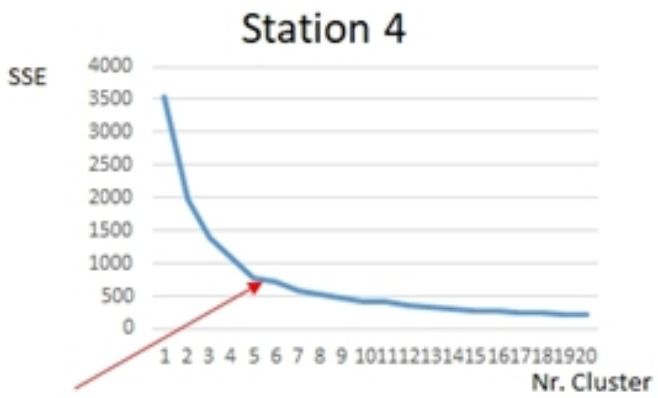

Figure 5.1: SSE vs. number of clusters graph for station 4

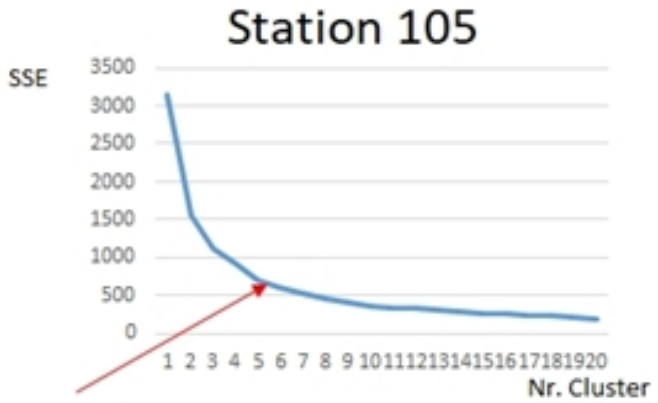

Figure 5.2: SSE vs. number of clusters graph for station 105

For both cases, we can identify a good candidate for number of cluster in 5, and this was the best choice for about all the stations selected, despite some reported problems with the choice, as shown in figure 5.3 for the station with id 106.

In any case, we selected 5 clusters for all the stations, so that there was consistency in the labelling system.

Once selected the best cluster number, we used it as parameter of the algorithm K-Means method in WEKA, obtaining the WEKA command:

```
weka.clusterers.SimpleKMeans -init 0 -max-candidates 100
-periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5
-A "weka.core.EuclideanDistance -R first-last"
```

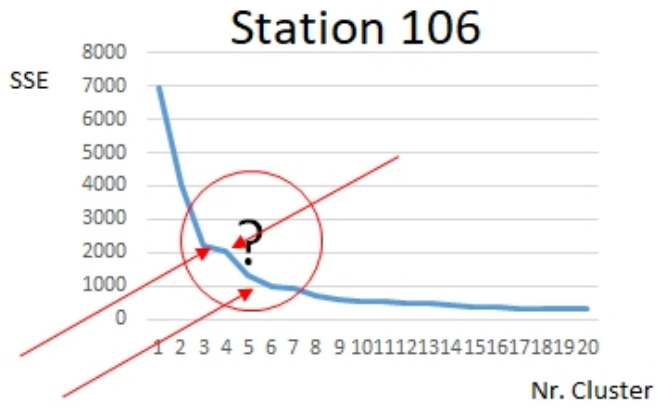

Figure 5.3: SSE vs. number of clusters graph for station 106

```
-I 500 -O -num-slots 1 -S 10
```

In table 5.1 we report result of K-means clustering for the station 4, with final centroids of cluster attribution. Figure 5.4 shows the instances distribution by speed and flow, with cluster association by colours.

Table 5.1: Final cluster centroids for station 4

Attribute Cluster (nr. of instances)	Speed Km/hr	Flow Veh/hr
Full data (197065)	53.2929	394.5504
0 (12960)	82.2556	82.1648
1 (64871)	51.8505	510.0687
2 (55486)	59.1947	176.6407
3 (27185)	38.0510	167.5926
4 (36563)	47.9622	799.7557

Analysing the centroids value for the clusters generated, we identified 5 possible groups of traffic typologies:

- FAST: this cluster is characterized by very low level of flow and very high speed, it represents the condition of absence of traffic, with high speed;
- SLOW: this cluster presents very low level of flow and very low speed, in this class we can consider the condition of absence of traffic, but with slow speed, or also the condition of blocked traffic;

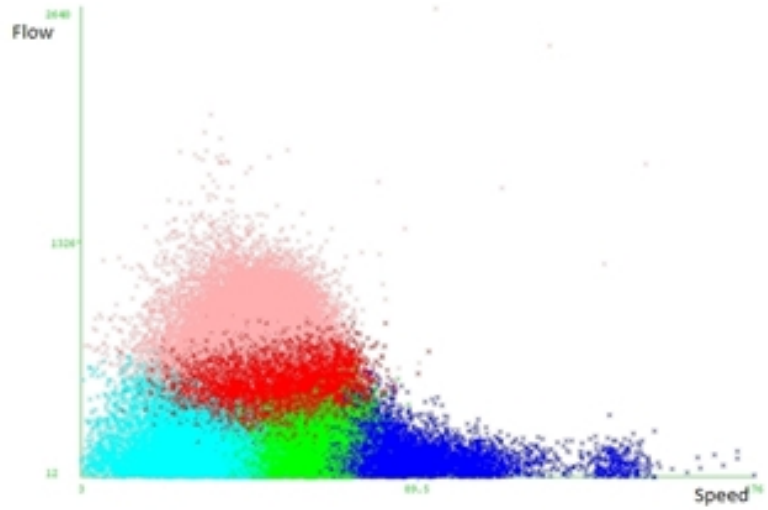

Figure 5.4: Clusters distribution for station 4

- LOW: this cluster presents a low level of flow and a relative high speed, we can identify in it a low level of traffic;
- NORMAL: this cluster presents a medium level of flow and a medium speed, it represents the normal condition of traffic;
- HIGH: this cluster presents the highest level of flow and a relative low speed, it represents a condition of high level of traffic.

Table 5.2 reports the association of cluster obtained for the station with id 4 with human readable labels that identify traffic typologies, based on speed-flow values.

The 5 groups of traffic flow we identified are not strictly related to traffic science. Indeed, the literature of speed-volume relationships in urban road networks relates to signalization and describes what are the traffic states to be met in urban networks.

For example, traffic flow conditions referred to unconstrained (free) flow, to conditions near to capacity, or to congested traffic.(Vlahogianni, 2007; Vlahogianni et al., 2008)

The aggregation of data in 5 minutes interval can not permit to compare our results to that of previous works, in particular for relationships to signalization, that changes in time along the day, within a range of 30 to 90 seconds.

Even the peculiarities of the roads presented at the beginning of chapter 3 can not permit to use the classic definitions for traffic flow status, so we adopted those presented above.

Table 5.2: Cluster labelling for station 4

Cluster	Label (traffic typology)
Cluster 0	FAST
Cluster 1	NORMAL
Cluster 2	LOW
Cluster 3	SLOW
Cluster 4	HIGH

Then, we identified the following possible rules for clusters labeling, and realized a simple algorithm to assing human readable label to cluster:

1. first af all, identify the two clusters with the lowest level of flow: one with the higest level of speed and the other with the lowest one;
2. assign the label "FAST" to the first and "SLOW" to the second;
3. order the remaing three cluster by crescent flow;
4. verify that previous step orders also by decrescent speed;
5. assign in order the labels "LOW", "NORMAL", and "HIGH".

We used this cluster's labelling policy for the results obtained by the clustering of all the stations.

5.2 Second step: classification model

We used the labelled training set obtained from previous step to test two different classification algorithms, in particular we used Decision Tree and Naive Bayes.

The respective WEKA command used are:

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Station-4_clustered
Instances:    197065
Attributes:   8
```

```
                Instance_number
                idSpeedFlowData
                giorno
                ora
                speed
                flow
                accuracy
                Cluster
Test mode:      10-fold cross-validation

Scheme:         weka.classifiers.bayes.NaiveBayes
Relation:       Station-4_clustered
Instances:      197065
Attributes:     8
                Instance_number
                idSpeedFlowData
                giorno
                ora
                speed
                flow
                accuracy
                Cluster
Test mode:      10-fold cross-validation
```

In table 5.3 we report statistics output from model building for station 4 for both methods used, in table 5.4 and table 5.6 the detailed Accuracy by Class for J48 method and Naive Bayes respectively, and finally, in table 5.5 and table 5.7 the confusion matrices for J48 method and Naive Bayes respectively.

We obtained best results with Decision Tree C4.5, in the J48 implementation of WEKA.

Indeed, as shown in table 5.3, the Naive Bayes method was faster, but had errors, both absolute and relative, greater than the Decision Tree method, and it had also a sparse confusion matrix, as shown in table 5.7.

As for the previous step, we repeated the process for all the 44 stations in the dataset, and we obtained classification models for every traffic data collecting point we chose to study.

Table 5.3: Statistics output for station 4 - J48 and NB methods

Parameter	Value (J48)	Value (NB)
Time taken to build model	3.4 seconds	0.27 seconds
Total Number of Instances	197065	197065
Correctly Classified Instances	196894 (99.9132%)	190007 (96.4184%)
Incorrectly Classified Instances	171 (0.0868%)	7058 (3.5816%)
Kappa statistics	0.9989	0.9524
Mean absolute error	0.0005	0.0598
Root mean squared error	0.0162	0.0598
Relative absolute error	0.1769%	19.8056%
Root relative squared error	4.1823%	35.9496%

Table 5.4: Detailed Accuracy by Class for station 4 - J48 method

Class Parameter	FAST	MEDIUM	NORMAL	SLOW	HIGH
TP Rate	0.999	0.998	1.000	0.999	1.000
FP Rate	0.000	0.000	0.000	0.000	0.001
Precision	1.000	0.999	1.000	0.999	0.997
Recall	0.999	0.998	1.000	0.999	1.000
F-Measure	0.999	0.999	1.000	0.999	0.999
MCC	0.999	0.998	1.000	0.999	0.998
ROC Area	1.000	1.000	1.000	1.000	1.000
PRC Area	1.000	1.000	1.000	1.000	1.000

Table 5.5: Confusion Matrix for station 4 - J48 method

Classified as Class	FAST	MEDIUM	NORMAL	SLOW	HIGH
FAST	12947	0	1	0	0
MEDIUM	13	64761	10	27	1
NORMAL	0	0	55475	0	0
SLOW	0	19	0	27151	2
HIGH	0	91	0	7	36560

We then labelled all the 3 years historical data traffic using an appropriate station model, we stored results in the same database, and at the end of the process, we had a complete set of traffic classes associated to stations, related to their date and time.

The framework analyses new upcoming data in real time, assigns the event to the relative class, and stores it in the database.

Table 5.6: Detailed Accuracy by Class for station 4 - NB method

Class Parameter	FAST	MEDIUM	NORMAL	SLOW	HIGH
TP Rate	0.851	0.980	0.979	0.956	0.959
FP Rate	0.007	0.019	0.021	0.001	0.001
Precision	0.899	0.962	0.949	0.994	0.993
Recall	0.851	0.980	0.979	0.956	0.959
F-Measure	0.874	0.971	0.964	0.974	0.976
MCC	0.866	0.957	0.950	0.971	0.971
ROC Area	0.993	0.997	0.988	0.998	0.998
PRC Area	0.903	0.996	0.984	0.976	0.984

Table 5.7: Confusion Matrix for station 4 - NB method

Classified as Class	FAST	MEDIUM	NORMAL	SLOW	HIGH
FAST	11026	30	1904	0	0
MEDIUM	4	63577	1006	66	218
NORMAL	877	258	54343	8	0
SLOW	351	842	0	25979	13
HIGH	11	1378	1	91	35082

5.3 Third step: forecasting of traffic class

Using classification data obtained from the two-steps process above we built a statistical trend, analysing, for each station, data by weekday and time. With this operation, we had as an output a table stored in the database that associates station, weekday, and hour with the statistic distribution of single traffic class in percentage, e.g.:

```
4, Monday, 05:25, FAST 0.85, NORMAL 0.00, LOW 0.08, SLOW 0.07, HIGH 0.0
```

We used this information as first attempt to forecast new traffic flow data at prefixed time distance, i.e. 15 minutes, 1 hour, and 2 hours.

In figure 5.5 we show the flow used for making forecasts of traffic flow classes.

We also added a filter to historical data selection: we considered weather condition as a driver for better forecasting.

We gathered historical weather data from Open Weather Map History Bulk download (`https://openweathermap.org/`) and we added that information to instances stored in the database.

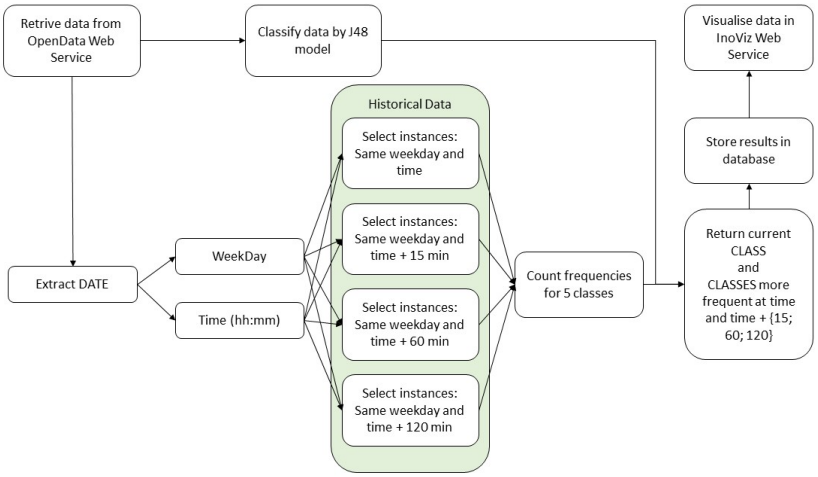

Figure 5.5: Flow for classes forecasting

We used the same Web Service, which publishes weather forecast in open data format (XML standard) accessible by API, to retrieve data for current and future weather simultaneously with the traffic data, and we filtered data selection from historical database for counting frequencies of traffic flow classes.

5.4 Fourth step: information visualization

Final step of the implementation of the system was the realization of the web-based visualization system for the traffic management officers of the city.

We utilized the MapSplit software for the extraction of tiles from OpenStreetMap data file including City of Turin at zoom level from 12 to 18.

We then used the OSM2World software to transform 2D tile images in 3D images.

The web-based service presents, using OpenLayers library, a map in 3D of Turin with stations location where sensors collect traffic data, circle tags with colours representing real-time status of traffic (by class), and a 3 column coloured histogram representing forecast at short, medium and long period.

A table summarizes the same information in human readable format.

Below the table, a chart can be seen with historic series for both model expected and real data classification, in a selectable range from 1 hour to 6

hours in the past.

In figure 5.6 we show an example of the output visualization.

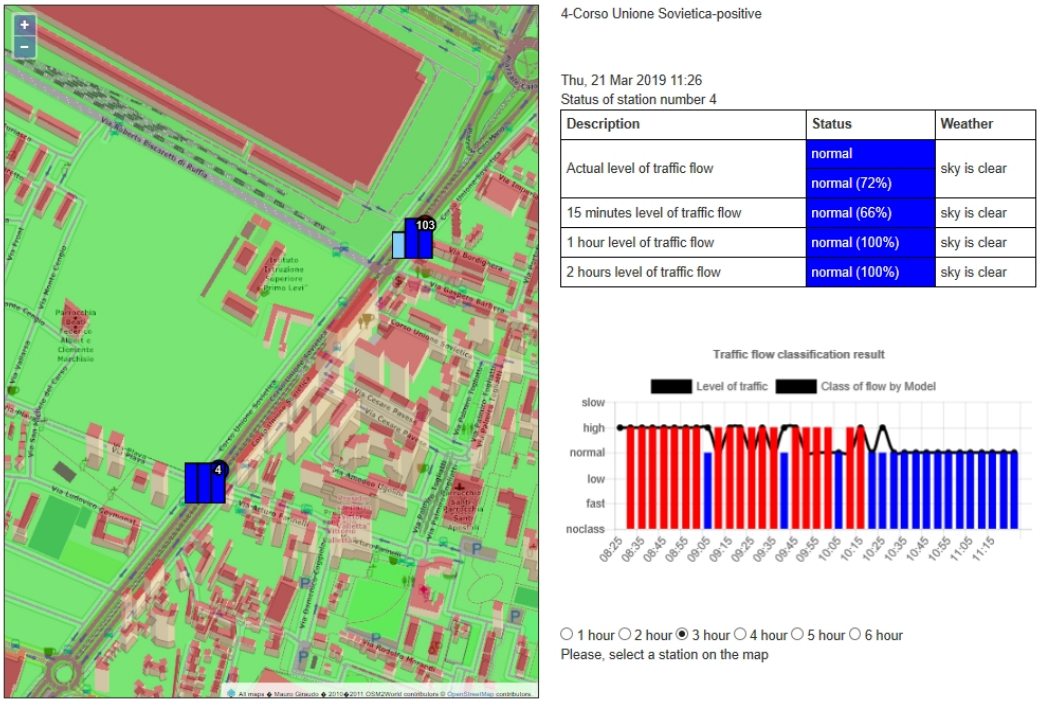

Figure 5.6: Example of system output visualization

Chapter 6

Results

In this chapter, we expose the results obtained from the application of the framework described in pervious chapter 5.

6.1 Results from clustering step

As presented in section 5.1 for the test bed of station 4, we applied the clustering method of K-means at the whole set of stations.

The elbow identification method return as good candidate for number of cluster the value of '5' for almost all stations, with only 3 cases where the identification wasn't clear and easy.

In any case, we operated the choice to adopt this numerosity of cluster for all the station, included the worst ones, to maintain coherence in the framework.

We report in table 6.1 the values of final centroids for a significant set of stations, the same used in section 4.1 for basic statistics, after the labelling of cluster based on the policies exposed in section 5.1:

- FAST: very low level of flow and very high speed - absence of traffic, with high speed;

- SLOW: very low level of flow and very low speed - absence of traffic, with slow speed, or blocked traffic;
- LOW: low level of flow and relative high speed - low level of traffic;
- NORMAL: medium level of flow and medium speed - normal level of traffic;
- HIGH: highest level of flow and relative low speed - high level of traffic.

Table 6.1: Final cluster centroids with label

Station Speed (Km/hr) Flow (Veh/hr)	FAST	NORMAL	LOW	SLOW	HIGH
4	82.2556	51.8505	59.1947	38.051	47.9622
	82.1648	510.0687	176.6407	167.5296	799.7557
20	82.2006	53.9076	58.2147	37.1051	45.2265
	72.8438	410.7657	156.3136	121.3266	722.8157
40	85.1909	57.8845	62.3274	41.5109	49.3695
	62.3341	516.6587	186.7820	147.9845	923.3314
73	89.9556	61.9505	69.9427	47.2061	57.9263
	42.1748	340.6037	122.4037	97.2356	579.7659
90	81.6796	52.8566	57.3347	39.1259	47.6531
	75.9438	459.3375	155.2236	141.6346	792.8547
106	72.2357	44.0515	47.7149	27.9061	37.2692
	102.6148	710.0687	496.9367	367.3286	1299.9327

6.2 Results from classification step

The aim of this step was to obtain the models of classification for every stations, and, as for the test bed of station 4 in the section 5.2, best results were achieved using Decision Tree method.

We report in table 6.2 the values of Weighted Average Precision, Correctly Classified Instances (in percentage), and Incorrectly Classified Instances, for the selected subset of stations.

As we can see, the Decision Tree method performed better than Naive Bayes, so the final decision we took was to adopt this classification model for all the stations selected.

Table 6.2: Classification statistics output - J48 and NB methods

Station Parameter	Value (J48)	Value (NB)
4		
Precision	0.999	0.964
Correctly Classified Instances	99.9132%	96.4184%
Incorrectly Classified Instances	0.0868%	3.5816%
20		
Precision	1.000	0.961
Correctly Classified Instances	99.9957%	95.9905%
Incorrectly Classified Instances	0.0043%	4.0095%
40		
Precision	1.000	0.952
Correctly Classified Instances	99.9977%	95.5237%
Incorrectly Classified Instances	0.0023%	4.4763%
73		
Precision	1.000	0.933
Correctly Classified Instances	99.9958%	93.2741%
Incorrectly Classified Instances	0.0042%	6.7259%
90		
Precision	1.000	0,972
Correctly Classified Instances	99.9947%	97.113%
Incorrectly Classified Instances	0.0053%	2.887%
106		
Precision	1.000	0,953
Correctly Classified Instances	99.9986%	95.2395%
Incorrectly Classified Instances	0.0014%	4.7605%

We also tested the robustness of the framework making some cross-classification: we tried to classify data from a station with models of another station.

In particular, we chose three stations: 4, 20 and 106. The first and the second had similar basic statistic, while the third is quite different (cfr. table 4.2).

Therefore, we selected a sample of 10000 records from those stations and we classified them with models of other stations. In table 6.3 we report the results in terms of Precision.

As we can see, considering a baseline of 0.20, we obtained good results for cross-classification from similar stations, and acceptable, even if not so good, from the different ones.

Table 6.3: Cross-Classification for station 4, 20, and 106

Station	Model	Precision
4	20	0.8953
	106	0.6918
20	4	0.8619
	106	0.6940
106	4	0.6842
	20	0.5685

6.3 Results from forecasting step

The next step of the framework was the forecasting of level of traffic starting from both historical and real-time data.

As described in section 5.3, we used a statistical approach to determine the most probable class at the same time of real-time data, and a three different range of time delay (15, 60, and 120 minutes).

We obtained good level of accuracy in our forecasts, in particular we report in table 6.4 the Precision of the forecasts for the subset of stations, for all the four time intervals, with the filter of weather condition applied to selection of historical data.

Table 6.4: Precision of forecasts

Station	Current time	15 minutes	60 minutes	120 minutes
4	0.669	0.662	0.659	0.648
20	0.637	0.632	0.627	0.622
40	0.659	0.652	0.649	0.638
73	0.648	0.651	0.648	0.641
90	0.663	0.661	0.649	0.641
106	0.654	0.649	0.642	0.638

After seeing the good but not optimal results of forecasting procedure, we tried to eliminate the filter of weather status when selecting data for statistical count of frequencies.

In table 6.5 we report the accuracy of forecasting aggregating the intervals, i.e. what was the precision of the correctness of the forecasts at the first two intervals (current time and 15 minutes), at three intervals, and at all four intervals.

Table 6.5: Precision of forecasts aggregate by intervals number

Station	2 intervals	3 intervals	4 intervals
4	0.572	0.563	0.527
20	0.537	0.532	0.527
40	0.557	0.556	0.546
73	0.548	0.543	0.539
90	0.563	0.561	0.549
106	0.564	0.548	0.542

In table 6.6 and in table 6.7 we report the same measures without the filter of weather status and forecast.

As we can see, the results had an improvement, reaching very good level in term of precision of forecasts. Even the aggregation of the intervals of forecasts gain in precision, and in stability of the prevision.

Table 6.6: Precision of forecasts without weather filter

Station	Current time	15 minutes	60 minutes	120 minutes
4	0.751	0.752	0.751	0.749
20	0.748	0.748	0.747	0.745
40	0.750	0.749	0.745	0.743
73	0.747	0.745	0.743	0.740
90	0.733	0.730	0.728	0.720
106	0.748	0.745	0.745	0.741

Table 6.7: Precision of forecasts aggregate by intervals number without weather filter

Station	2 intervals	3 intervals	4 intervals
4	0.750	0.749	0.745
20	0.747	0.746	0.744
40	0.749	0.745	0.741
73	0.745	0.741	0.738
90	0.732	0.726	0.719
106	0.745	0.743	0.739

Chapter 7

Conclusions and future works

Smart City is a paradigm rapidly evolving and ICT has to support it, not only by infrastructure and end-user service, but also with services addressed to city managers.

Inside the project 'Torino As a Platform' we studied a system to classify vehicular traffic flow within the boundaries of the city of Turin, and make predictions about its status in short, medium and long time, with the aim to support the City governance to implement correct actions to address traffic issues.

The system proposed is a three-level framework: the first level consists in a two-step algorithm, one for labelling a training set using K-means clustering method, and one for building a decision tree model for the traffic data classification; the second level makes forecasts for classes of traffic flow, starting from previous classification models; the third presents results from both the previous level in a simple way.

We applied this framework to a set of 46 stations of traffic flow detection, starting from historical data from year 2015 to 2017 to build models, and using real-time open data web service for current data.

We used those models to made forecasts at short, medium and long period (from 15 minutes to 2 hours).

An Information Visualisation web-based service shows real-time and fore-

cast data to city management.

In the following section we discuss the results in relation to the research questions presented in chapter 3:

1. How implement a classification framework that uses historical data of vehicular traffic flow;
2. Use of results of classification process for the prediction of future status of traffic flow;
3. Focus on the pratical case of a large city and on support to its management;

7.1 Discussion of results

Results exposed in chapter 6, and in particular mode those of section 6.3, need some analyses and discussion.

The clustering step returned very good results, with clusters well-defined that easily permitted the labelling in human readable format.

The clusters identified by this step of the framework can be transformed in the final classes by the simple algorithm proposed:

1. identify the two clusters with the lowest level of flow: one with the higest level of speed and the other with the lowest one;
2. assign the label "FAST" to the first and "SLOW" to the second;
3. order the remaing three cluster by crescent flow;
4. verify that previous step orders also by decrescent speed;
5. assign in order the labels "LOW", "NORMAL", and "HIGH".

In addition, the classifying step returned very good results, with a clear identification of method to be used in the Decision Tree, and models built that had very good performance in term of precision.

The models of classification returned good results also in application of a cross model classification, in particular mode when stations presented similar basic statistics (speed and flow minimum, maximum, and mean).

Those results provide a good answer to the first research question: we built the first level of framework for automatic labelling of an historcal data set and for classifying new data gathered from open data service in real-time.

Results obtained from the forecasting level, at first, were not so good, with precision that varied around the value of 0.65 for all four interval where we calculated the prevision.

In addition, the aggregation of intervals did not return good results, with values around 0.55.

Therefore, we decided to simplify the model of forecasting: in fact, we eliminated the filter on weather status and prevision.

This action led to a great improvement in all four intervals, and also in the aggregation check, with values that reached a precision of 0.75, stable and without great variance.

The same stability was reached also in the aggregation check: in fact, the model predicted at the same precision the level of traffic at every intervals.

We can find the reason of this improvement in the fact that the Web Service from where we retrieve the data (Open Weather Map) is not originated in Italy, and has forecast organized in blocks of 2 - 3 hours.

This type of data, probably, introduced a bias error in the frequencies calculation and reduced in this way the accuracy of the forecasts.

Results obtained from the forecasting level of the framework results in a positive answer to the second research question: this system based on the labelling and classification process determine a good way to make prediction on future status of traffic flow.

For the third research question we proposed the last level of the framework: the info-vis system.

Even if in a embrional status, it can help city managment officiers to check the status level of the traffic flow and give a visual information for the expected one in three different range of time. This can be helpful for the identification of traffic issues related to mid range time period and for analysing historycal trends.

7.2 Future works

Future work and further improvement will be the integration of the system with a database of actions taken in traffic issues treatment, with analysis of

impact in terms of troubleshooting in their specific application.

We will propose an evaluation of results of action by studying the deviation of real flow from the predicted one.

The aim of this integration is to build a knowledge base of possible actions that can be taken in traffic issues with a grade of awaited result based upon previous contexts and applications.

Another line of research could be the integration with a more accurate system of weather forecast, specialized in local (Piedmont, or better, Turin Area) weather monitoring.

At the end, we can suggest the application of the framework to different systems of traffic flow data retrieving system, like traffic cameras, or access gates.

Bibliography

L.G. Anthopoulos. Understanding smart cities: A tool for smart government or an industrial trick? *Public Administration and Information Technology*, 22, 2017.

Virgil Chichernea. The use of decision support systems (dss) in smart city planning and management. *Romanian Economic Business Review*, 8(2): 238–251, 2014.

S. P. Hoogendoorn and P. H. L. Bovy. State-of-the-art of vehicular traffic flow modelling. In *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, volume 215, pages 283–303, 2001.

Smart Cities, Preliminary Report. International Standards Organization (ISO), 2014. URL `http://www.iso.org/iso/smart_cities_report-jtc1.pd`.

George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

M.H. Lighthill and G.B. Whitham. On kinematic waves ii: a theory of traffic flow on long, crowded roads. In *Proceedings of the Royal Society of London series A*, volume 229, pages 317–345, 1955.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, page 281297, 1967.

M. Montazeri-Gh and A. Fotouhi. Traffic condition recognition using the k-means clustering method. *Scientia Iranica*, 18(4):930–937, 2011.

Alessio Pagani, Francesco Bruschi, and Vincenzo Rana. Knowledge discovery from car sharing data for traffic flows estimation. In *Smart City Symposium Prague (SCSP)*, 2017.

Pavla Pecherková and Ivan Nagy. Analysis of discrete data from traffic accidents. In *Smart City Symposium Prague (SCSP)*, 2017.

N. Petrovska and A. Stevanovic. Traffic congestion analysis visualisation tool. In *IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas*, pages 1489–1494, 2015.

Selvaraj Shanthi. Feature relevance analysis and classification of road traffic accident data through data mining techniques. In *Proceedings of the World Congress on Engineering and Computer Science*, volume I, San Francisco (USA), October 24-26 2012.

So Young Sohn and Sung Ho Lee. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in korea. *Safety Science*, 41(1):1–14, 2003.

Anthony Stathopoulos and Matthew G. Karlaftis. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C*, 11:121–135, 2003.

Thammasak Thianniwet, Satidchoke Phosaard, and Wasan Pattara-Atikom. Classification of road traffic congestion levels from gps data using a decision tree algorithm and sliding windows. In *Proceedings of the World Congress on Engineering*, volume I, London (U.K.), 2009.

Miroslav Vaniš and Krzysztof Urbaniec. Employing bayesian networks and conditional probability functions for determining dependences in road traffic accidents data. In *Smart City Symposium Prague (SCSP)*, 2017.

E. I. Vlahogianni. Some empirical relations between travel speed, traffic volume and traffic composition in urban arterials. *IATSS RESEARCH*, 31(1): 110–119, 2007.

E. I. Vlahogianni, N. Geroliminis, and A. Skabardonis. Empirical and analytical investigation of traffic flow regimes and transitions in signalized arterials. *ASCE Journal of Transportation Engineering*, 134(12):512–522, 2008.

Eleni I. Vlahogianni, Matthew G. Karlaftis, and John C. Golias. Short-term traffic forecasting: Where we are and where were going. *Transportation Research Part C*, 43:3–19, 2014.

Rong Yu, Guoxiang Wang, Jiyuan Zheng, and Haiyan Wang. Urban road traffic condition pattern recognition based on support vector machine. *Journal of Transportation Systems Engineering and Information Technology*, 13(1): 130–136, 2013.

Y. Zhang, N. Ye, R. Wang, and R. Malekian. A method for traffic congestion clustering judgment based on grey relational analysis. *ISPRS Int. J. Geo-Inf.*, 5(5), 2016.