**TAaP: Torino AS a Platform. An ICT framework to support governance of Smart City**

**This is the author's manuscript**

*Original Citation:*

<table>
<thead>
<tr>
<th>Availability:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>This version is available</td>
<td><a href="http://hdl.handle.net/2318/1680913">http://hdl.handle.net/2318/1680913</a> since 2022-02-01T13:35:56Z</td>
</tr>
</tbody>
</table>

*Publisher:*

IGI PUBLISHING

*Terms of use:*

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
ABSTRACT

Smart City is a paradigm rapidly evolving which brings with it new ICT issues to address. Torino As a Platform is the project that the City of Turin is developing to address those goals: collect 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. Inside the project we studied a system to classify vehicular traffic flow within the boundaries of the city, and make predictions about its status in short, medium and long period (15 minutes, 1 hour, 2 hours).

KEYWORDS

Smart City, Vehicular Traffic Flow, Classification, Data Driven

1 INTRODUCTION

As described in [1] Smart City is a paradigm rapidly evolving which brings with it new ICT issues to address. In [2] 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 fulfill the purposes it has set for itself. Public data are a pillar on which found its development, 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 address following objective: collect 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.[3]

The project realization bases its development on two main branches: on one side realize and deliver to communities an SDK with standard API for increase and facilitate collection and reuse of data and citizens’ participation. On the other side develop 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 we report the big picture of the project.

![Figure 1: TAaP full diagram](image)

Inside this second branch, we studied a system to analyze 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).

Aim of the final output of the system is to support the City governance to implement correct actions to address traffic issues.

The remainder of this paper is organized as follows. The Section 2 is a brief overview of previous work on vehicular traffic flow modelling and classification followed by the description of the case study of City of Turin and finally by the framework proposed. Section 3 describes the dataset used in this study and Section 4 describes the results obtained by our framework in term of classification, prediction and output visualization. Finally, Section 5 summarizes the findings of the work and open to new research question.

2 Vehicular traffic flow

2.1 Overview and previous works

Analysis of traffic flow is a long time life research theme, with studies focused on two main aspect: traffic modelling and traffic forecasting.
In 2001, Hoogendoorn and Bovy made a good state-of-the-art analysis of the previous works on traffic modelling [4], focusing their attention on the nearly fifty years of traffic flow theories and models.

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).

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

On the other side Vlahogianni et al. in 2014 presented their analysis of existing works in short-term traffic flow forecasting [5]. Their work based on a set of ten challenges stemming from the changing needs of ITS (Intelligent Transportation System) applications.

Most of all studies analyzed has largely used single point data from motorways and has employed univariate mathematical models to predict traffic volumes or travel times.

### 2.2 Turin case study

The case we study in this work presents some peculiarities that are opposed to previous works and that have guided the implementation choices.

First, the context in which we are operating: the traffic flow we want to analyze is that inside the city boundaries, in an urban scenario and not that of motorways or freeways. Every road has peculiarities, e.g. number of lanes, traffic lights, intersections, different speed limits or other limitations.

Second, our start point are the data collected from city traffic management, in particular from loops present in city roads, in the last 3 years.

Third, our objective is a model for classify traffic flow, e.g. from fast to high, and predict future class.

At the end, the result of our work is an InfoViz system that output the traffic status and forecasts to support city traffic manager to address issues.

In this context major issues we have faced are: i) manage the large amount of data, ii) build a correct training set for the classification algorithm, and iii) realize a simple InfoViz system to be presented at public manager.

### 2.3 Related works

Some scholars approached similar issues.

Yu et al. studied a system based on SVM (Support Vector Machine) that recognize traffic condition pattern in urban road traffic, testing their model on a simulated dataset of 50 cases [6].

Petrovska and Stevanovic presented an innovative visual tool with the aim of detecting and avoiding road traffic congestion, based on Google Maps data [7].

Montazeri-Gh and Fotouhi used k-means clustering algorithm for traffic condition recognition, collecting data from an ad-hoc hardware device installed on a vehicle [8].

Stathopoulos and Karlaftis 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 [9].

Thianniwet et al. proposed a classification system of road traffic congestion based on Decision Tree Algorithm and Sliding Windows, collecting data from a pc on a car and human rating did initial classification of congestion [10].

Other works focused on particular condition and data, e.g. traffic accident analysis, traffic congestion or car sharing data [11][12][13][14][15][16].

### 2.4 Our framework

In our work, we proposed a combined approach to analyze historical data of traffic flow from city of Turin: 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 labeled dataset, to assign class to the data.
We tested different combinations of methods (C4.5, decision tree, ADABoost, kNN, Naive Bayes, CART) to develop this classifier framework for traffic data.

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

As next step, we proposed a predictive model about traffic flow status in short, medium and long period (15 minutes, 1 hour, 2 hours), and finally visualizes all data in a visual web-based info-system, to support responsible offices to make decision about actions to be taken.

3 Dataset

We gathered historical data from 5T consortium, a public society that manage 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, in a 3 years period from January 2015 to December 2017.

The traffic data consist in: i) mean speed (Km/Hr), ii) flow (Veh/Hr), and iii) accuracy (percentage); are calculated in 5 minutes aggregation, and accompanied by geolocation of stations (longitude and latitude) and road information.

For the real time analysis, we gathered data from 5T OpenData web service, which publish the same type of information in XML format.

(http://opendata.5t.torino.it/get_fdt).

This 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 1 we report an example of historical data dimension.

<table>
<thead>
<tr>
<th>Station</th>
<th>Lon</th>
<th>Lat</th>
<th>Road Name</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7.62745</td>
<td>45.01608</td>
<td>Corso Unione</td>
<td>309019</td>
</tr>
<tr>
<td>93</td>
<td>7.64825</td>
<td>45.03448</td>
<td>Corso Unione Sovietica</td>
<td>243629</td>
</tr>
<tr>
<td>95</td>
<td>7.63681</td>
<td>45.02413</td>
<td>Corso Unione Sovietica</td>
<td>197759</td>
</tr>
</tbody>
</table>

We done some preliminary operations on data: excluded that had accuracy equals to 0, excluded stations which had gaps on data retrieving (e.g., some didn’t work for long period, or started working only in 2017), and stored it in a MySQL database.

Final dataset consisted in about 15 millions of records from 51 stations, which figure 2 shows the distribution along the city road.

4 RESULTS AND DISCUSSION

As presented above our framework consist in a two-step system, first for labeling of the training set, and the second the classification of data.

The final classification model is used for real-time task on new data, which are displayed in the InfoViz service, along with forecast.

For the tests and the implementation of the system we used WEKA software [17].
4.1 First step: clustering for label data

First operation we done in this step was the choice of correct number of cluster to be used.

We analyzed 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, and then selected 5 as number of cluster to use in this step.

In figure 3 and 4 we show the graph from two stations, as examples, with red arrow identifying the “elbow” corresponding to a good choice for the number of cluster.

In table 2 we report result of K-means clustering for the station 4, with final centroids of class. Table 3 reports the association of cluster obtained for this station with human readable labels that identify traffic typologies, based on speed-flow values.

Figure 5 shows the instances distribution by speed and flow, with cluster association by colors.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Speed</th>
<th>Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster# (nr. of instances)</td>
<td>Km/Hr</td>
<td>Veh/Hr</td>
</tr>
<tr>
<td>Full data (197065)</td>
<td>53.2929</td>
<td>394.5504</td>
</tr>
<tr>
<td>0</td>
<td>82.2556</td>
<td>82.1648</td>
</tr>
<tr>
<td>1 (12960)</td>
<td>51.8505</td>
<td>510.0687</td>
</tr>
<tr>
<td>2 (64871)</td>
<td>59.1947</td>
<td>176.6407</td>
</tr>
<tr>
<td>3 (55486)</td>
<td>38.051</td>
<td>167.5926</td>
</tr>
<tr>
<td>4 (27185)</td>
<td>47.9622</td>
<td>799.7557</td>
</tr>
<tr>
<td>(36563)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Cluster labeling for station 4

<table>
<thead>
<tr>
<th>Cluster#</th>
<th>Label (traffic typology)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 0</td>
<td>FAST</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>MEDIUM</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>NORMAL</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>SLOW</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>HIGH</td>
</tr>
</tbody>
</table>
We repeated the process for all stations with the result of labeling 51 training set, one for each station, to be used in the next step.

### 4.2 Second step: classification model

We used the labeled training set to test different classification algorithms, and we obtained best result with Decision Tree C4.5, in the J48 implementation of WEKA.

In table 4 we report statistics output from model building for station 4, in table 5 the detailed Accuracy by Class, and finally in table 6 the confusion matrix.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken to build model</td>
<td>3.4 seconds</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>197065</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
<td>196894 (99.9132 %)</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>171 (0.0868 %)</td>
</tr>
<tr>
<td>Kappa statistics</td>
<td>0.9989</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0005</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.0162</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>0.1769 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>4.1823 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>FAST</th>
<th>MEDIUM</th>
<th>NORMAL</th>
<th>SLOW</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
<td>0.999</td>
<td>0.998</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Precision</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>0.997</td>
</tr>
<tr>
<td>Recall</td>
<td>0.999</td>
<td>0.998</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>MCC</td>
<td>0.999</td>
<td>0.998</td>
<td>1.000</td>
<td>0.999</td>
<td>0.998</td>
</tr>
<tr>
<td>ROC Area</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PRC Area</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

As previous step, we repeated the process for all the 51 stations in the dataset, and obtained the classification models for every traffic data collecting point we chose to study.

We than labeled all the 3 years historical data traffic using appropriate station model, stored results in the same database, so we had a complete set of traffic class associated to stations and relative date and time.

### 4.3 Forecasting of traffic class

Using classification data obtained from the two-steps process above we built a statistical trend, analyzing for each station data by weekday and time.

With this operation, we had as output a table stored in the database that associates stations, weekdays, and hours with the statistic distribution of single traffic class in percentage, e.g. station 4, Monday, 05:25, FAST 0.85, MEDIUM 0.00, NORMAL 0.08, SLOW 0.07, and HIGH 0.00.

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.

We obtained, with this simple statistical approach during a test bed of 1 week, a precision of 0.93 for forecast at 15 minutes, 0.89 at 1 hour, and 0.81 at 2 hours.

### 4.4 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 MapSplit software [18] for the extraction of tiles from OpenStreetMap [19] data file including City of Turin at zoom level from 12 to 18, then we used OSM2World [20] to transform 2D tile images in 3D images.

The web-based service presents using OpenLayers library [21] a map in 3D of Turin with stations location where sensors collect traffic data, circle tags with colors representing real-time status of traffic (by class), and a 3 column colored histogram representing forecast at short, medium and long period.

A table summarizes same information in human readable format.

In figure 6 we show an example of the output visualization.
5 CONCLUSIONS

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 manager.

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

The system proposed is a two-steps framework, first for label the training set using k-means clustering, and second for build a decision tree model for the traffic data classification.

With models obtained we classified a 3 years set of 51 stations of data collection, and used this output to made forecast at short, medium and long period (from 15 minutes to 2 hours).

An Information Visualization web-based service shows real-time and forecast data to city management.

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.

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.

ACKNOWLEDGMENTS

Thank you at 5T S.r.L. consortium for historical data and City of Turin for support and interest, in the person of Deputy Major at Smart City and Innovation Paola Pisano.

REFERENCES

[11] Selvaraj, Shanthi, Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining
[12] So Young Sohn, 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), 2003, pages 1-14.


[18] https://github.com/PedaB/mapsplt

