An Unsupervised Behavioral Analysis of Highway Traffic Flow for Security Applications[*]

Fabrizio Balducci[1†], Gabriella Calvano[1], Donato Impedovo[1] and Giuseppe Pirlo[1]
[1] University of Bari Aldo Moro, 70125 Bari (Italy)
name.surname@uniba.it

Abstract

Every day millions of people drive along urban roads and highways featuring driving own behaviors. When considering the highway network, great efforts have been employed to monitoring such large number of users considering traffic code infringements, accidents detection or traffic congestion estimation. Such information are useful to reveal hidden drivers' habits: in this study an analysis on traffic data has been developed to highlight suspicious events of highway users exploiting a dataset of the Italian Traffic Police. The analysis employed unsupervised clustering techniques and a series of filters on possible routes useful to isolate suspicious car stops and meetings on service areas.

1 Introduction

In a world characterized by continuous big data production and recording, vehicular traffic data can be useful for real-time applications dealing with traffic congestion evaluation and secure routing (Pan J et al., 2017; Impedovo D et al., 2019a), for accident and traffic code infringements detection (Chang B.R. 2010) and, of course, traffic flow prediction (Impedovo D et al., 2019b). An emerging topic deals with security and safety checking where vehicular data analysis could reveal crimes or suspicious behaviors like vehicles that repeatedly stop together in the same time intervals into service and refueling stations.

The aim of this paper is to carry out specific research tasks on a dataset released by the Italian Traffic Police which contains numerical logs about vehicular detections in a specific fragment of the Italian highway network. Developing a tool able to detect from recorded data abnormal behaviors is inherently difficulty since such large volume of data requires a not supervised approach. This study provides an operational approach with some basic assumptions about the nature on the data and on the required results that lead to interesting results.

The work is organized as follows: in Section II related works are presented while Section III shows the data sources describing the nature of their information and introducing the tasks proposed by the Italian Police. Section IV relates to the methods, algorithms and tools exploited and in Section V are described the workflow processes developed and the prior assumptions for the analysis whose results are commented and exposed in Section VI. Finally, conclusions and future works are exposed in Section VII.

2 Related Works

In Beshah et al. (2010) it has been showed that road accidents do not occur randomly but are generated by complex relationships between multiple factors and to demonstrate that, driver behavior, road conditions and vehicle types are treated by well-known classifiers (J48, Naïve Bayes and KNN) on the dataset provided by the Addis Abeba traffic control department.

The work of Pakgohar et al. (2011) analyses a road accidents dataset provided by the Iranian police to discover that the role of the human factor in their incidence and severity is very relevant for the type of accident so the CART (Classification and Regression Trees) and Logistic Regression models were adopted to identify three accident classes (fatal, with injured and without injured).

In Jain et al. (2016) an analysis has been carried out on data from the Indian Government in order to identify areas most prone to road accidents together with the dominant factors that trigger them by using K-Means clustering classified with Bayes Nets, Multinomial Bayes and J48 reaching a predictive accuracy of 76,7%; another work about traffic is in Zamani et al. (2010) where IoT and data mining are exploited to create a traffic management system using road signs and traffic lights: data from sensors of five crossroads in Isfahan (Iran) provided information regarding the traffic volume and the type of vehicles creating several clusters with TOD (time-of-day) intervals to identify the traffic conditions.

A deep learning approach for traffic analysis comes from Lv et al. (2015) featuring a mean absolute error (MAE) of 64.0 on a 30-minutes traffic flow prediction comparing four classification techniques while the study of Di Mauro and Ferilli (2018) propose a sequence-to-sequence autoencoder to detect anomalous routes consisting of a Long Short Term Memory (LSTM) model that maps input paths to a vector of a fixed length while a LSTM decoder recovers valid routes with an accuracy of 95,6%. Moreover, Buono et al. (2018) introduce a visual approach to detect anomalous routes using a sequential pattern mining while Fumarola F. and Lanotte (2017) exploit unsupervised learning for traffic anomalies recognition while the work of Bernaschi et al. (2018) results interesting since they apply unsupervised learning techniques on the TRAP-2017 dataset but focusing only on a statistical survey and on the inconsistent records.

Dangel et al. (2014) exploited data collected from 45 routes in Dublin to extract behaviors and levels of self-similarity while Perez-Zuriaga et al. (2013) compare speeds from videos and GPS using questionnaires on ten road segments with two checkpoints on both directions: this data lead to drivers profiling and to a models that estimates the driving styles. The work of Brambilla et al. (2017) introduces motion tracking to perform filtering processes on features like speed estimation and GPS using Hidden Markov Models to compare different driving styles defined on spatio-temporal factors (2015).

3 Tasks and Data Sources for a Generative Approach

In the Event occurred at 25-26 October 2017 in Rome (Trap Traffic Police, 2017) there have been presented to researchers some problems regarding surveillance activities on highway traffic with topics

like the detection and tracking of road users (vehicles, bikes, trucks, etc.), behaviors and traffic scenarios understanding, operative applications of traffic surveillance and vehicles accident analysis.

The research tasks proposed were:

1. find sequential visit of service areas each facing the other at the same time (inconsistent)
2. isolate sequential visits of service areas in the same direction (suspicious)
3. detect inconsistent transits under the space-time point of view (proof of cloned plates)
4. combinations of the previous cases

Considering the proposed tasks, in this work are performed analyses about the classification of inconsistent or suspicious paths, the identification of paths that sequentially visit service areas in the same traffic flow direction and the detection of recurring car meetings in service areas.

The provided public dataset is TRAP-2017 (Leuzzi F. et al, 2018 and Trap event: Polizia di Stato, 2019) containing automatic acquisition of vehicular plates by official highway portals (SICVe-PM systems) placed on section of 350Km in the Italy territory. For such research studies only a small part of the highway detections has been provided by the Italian State Police: in particular, 365 CSV files from 1st January 2016 to 31 December 2016, containing the transits of 14,351,059 different vehicles of which 6,958,429 have at least two registered transits for a total of 111,089,717 textual entry records.

Each record consists of five attributes:

1. **plate**: univocal anonymized identifier of a real vehicular plate
2. **gate**: the gate number where the detection happened
3. **lane**: the travel lane of the detected vehicle
4. **country**: the Nation to which the detected plate belongs
5. **timestamp**: date and time of detection, with precision up to seconds

An official map representing the stretches of highway to which the detections belong has been provided resulting useful to make assumptions about the nature of the plate recordings while the distances (kilometers) between gates and the position of the service areas is also available.

The dataset is not provided with labels (ground-truth) about the effective anomalous detection so it results impossible to use the supervised approach: it is necessary to extract and calculate features that allow to make hypotheses on the nature on the nature of such vehicular detections.

The CSV files have been translated into tables of a relational database with the aim to determine the average speed of vehicles between two successive gates as well as to outline steps of larger paths by generating paths through the concatenation of passages under the highway gates with a temporal order (consecutive detections). Given the specific interest on vehicular stops at service stations, all paths generated for each vehicle through gate detections will be reduced to those containing at least a service station.

In this way, it results useful a pre-processing step that deals with anomalous values according to the map topology, and specifically to isolate the following cases:

- a plate recognized at gate A and C without being detected at gate B (detection fault or suspicious path)
- temporally near detections over very far gates (a system timing error or clues of cloned plates)

4 Methods and Tools

The main approaches required to perform the steps described in the previous section for big data analysis are: Regression and Classification (use pre-labeled data to predict values or labels for never seen before objects) and Clustering (divide data into similar subsets named clusters without previous labeling). Since in the TRAP-2017 dataset there is no a-prior information on data to analyses (i.e. the specific behavior of a vehicle is not known) and, more specifically, there is a lack of knowledge about the history of each vehicle circulating on the highway, a possible way to work is to use part of available data to build such knowledge in a complete unsupervised manner adopting some basic assumption.

Among the unsupervised techniques the K-Means has been chosen due to its power, speed and relative simplicity of execution not needing excessive computational resources: it consists in associating n elements (vehicles in this case) in K clusters (specified depending on the specific task performed) according to their neighbor (calculated with some metrics based on numerical distances). The vehicle set will result clustered in groups that are compact as possible within them (intra-class closeness), and almost distinct (inter-class distance) from each other. Each cluster is identified by its centroid, or midpoint since K-Means algorithm results suitable to group similar behaviors starting from a collection with no prior knowledge as in this work.

To perform machine learning advanced clustering, the *RapidMiner Studio* framework will be adopted: it is a visual workflow tool used to manage big data from pre-processing phases to the algorithms applied on heterogeneous data sources. It provides a set of already available facilities also allowing to integrate its visual part with code like Python scripting. The use of this powerful but also accessible tool with a friendly user interface and a fast productive time, brings advantage like a quick replication, reuse and customization of workflows and/or their components (blocks) and like the possibility of a soft introduction in small/medium business and industrial environments.

5 Generation of Vehicular Paths for Service Area Visits

From CSV and serialized SQL tables all data have been imported on RapidMiner generating the possible routes between highway gates made by each vehicle (supposing the consecutive gate connections): each record considered as a “candidate step” of a generated vehicular path has the form [plate - entrance gate - exit gate] in such generative approach while the ‘country’ attribute will be ignored in the following since missed in almost 35% of the records.

There have been developed three sub-processes for this task where the first one is ‘Generate road steps’ performed using peculiar RapidMiner blocks, starting from an ExampleSet where the code plate of all vehicles have been ordered:

- Multiply: doubles the ExampleSet by creating two identical copies
- Merge Attributes: takes the two ExampleSets and combines attributes according to their position and order; in this way can be joined the consecutive highway gate detections that have been crossed by a vehicle two by two, useful for the next step that generates the vehicular paths
- Filter: select only records in which vehicular plate maintains an ascending or descending detection order (i.e. instances of vehicles that proceed on the same highway route)

Next, according to the highway real topology (three branches, see the TRAP-2017 dataset documentation) in the provided map inconsistent routes can be removed. In the workflow in Figure 1 there is the second sub-process ‘Generate and Remove inconsistent roads” where the first block is an operator configured for each of the three highway branches in the provided map (parameter *n*).

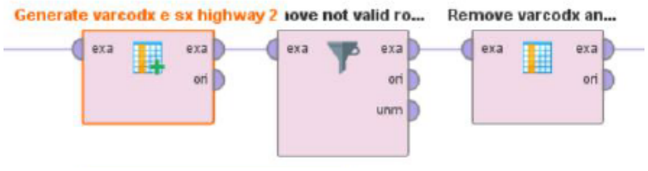

Figure 1: Workflow to generate vehicular paths and Remove inconsistent roads.

- Generate left & right gate highway (param. n): based on the value assumed by the parameter n, generates the new attribute “left gate” and “right gate” using the gate code from the map for the highway branch n.
- Remove not valid paths: remove the unlawful steps of vehicular path by checking only those where real consecutive gates in left or right travel direction (from the map) correspond to the produced records of the selected highway branch n
- Remove left & right gate: remove the ‘left/right gate’ vanilla attributes previously created since invalid paths have already been removed

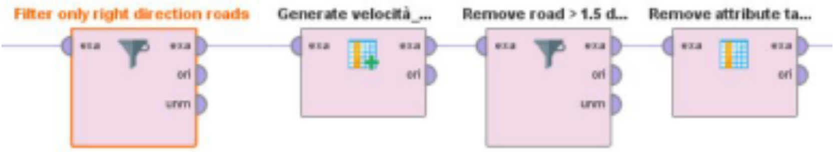

Figure 2: Sub-process “Select Gates”.

The sub-process in Figure 2 applies further filters for a valid vehicular paths on each highway branch.

- Filter only correct direction roads: selects path routes for correct consecutive directions (for example, if from the map the position of gate 5 comes first of the gate 8 then the opposite will be discarded)
- Generate average speed: generates the average speed of vehicles (that can be used as a clustering feature) with a Python code script that exploits the timestamps
- Remove paths > 1.5 days: paths whose total time-duration is greater than 1.5 days are removed because out of the scope of this investigation, assuming highway entrances on different days rather than very long stops
- Remove attribute plate_2: removes from the dataset an attribute related to the building of the vehicular steps during previous ExampleSets merging

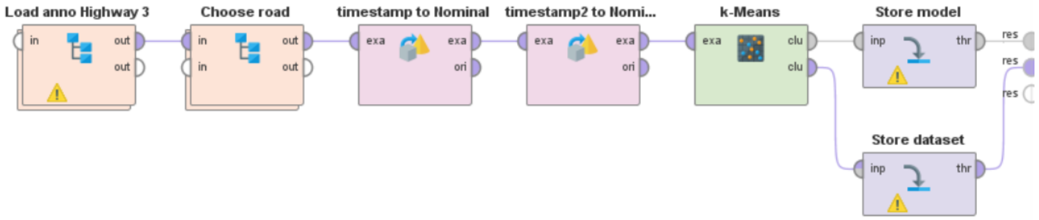

Figure 3: Workflow for the vehicular path unsupervised clustering.

The "vehicular path" clustering workflow in Figure 3 performs an unsupervised K-Means clustering on the paths generated by the previous processes. Figure 3 reports the whole workflow for a single highway branch (that must be executed for each of the three branches).

The first sub-process block (Load) is used to retrieve data from the repository concerning the highway part to which routes to be classified belong: there are three 'Load year Highway n' sub-process with twelve Retrieve operators, one for each month, so that they can create a global ExampleSet.

The latter is given as input to the 'Choose Road' sub-process in which there are nine filter operators, one for each route pieces (founded manually from the provided documentation highway map) containing almost a couple of service station: the two timestamps in fact, indicate the passage under the first (entering) and second (exit) gate respectively with a calculated average speed. Next, the timestamp values for each detection gate couple are transformed in numeric values to be employed in the K-Means clustering: the number of executions was set to 50 while the Euclidean Distance was chosen for similarity measures, having to compare the average velocity; the number of K after several attempts has been set to an optimal value of 3.

An aspect of the study involves a further analysis to search the generated vehicular paths while considering their amount of stopping time in services areas considering the stops that involved more than one vehicle (suspicious encounters).

At the end of the previous clustering process, the relevant clusters were re-clustered with the K-Means with the aim to highlight vehicular paths featuring an average velocity low enough to suggest the presence of almost one stop in them. As a result of the processes a set of nine repositories is obtained, one for each path containing almost a service area in which there are vehicles that stopped almost one time in them. Each record contains therefore two timestamps, which indicate the passage respectively under the first and second passage and the average speed.

Now that they have been obtained the individual stops of the various vehicles, it is necessary to found which of them stopped together in almost one service station. To do this, a new K-Means clustering can be used in a different way since the main idea is to find k circles, each of which corresponds to simultaneous or very close stops by two or more cars. In this way, the workflow in Figure 4 exploits the dataset containing detections for 365 days where a year is composed of 8,760 hours: the timestamps are converted into numeric values and a trial-error method is used to set the cluster number to be detected trying different combinations (8760 as the same, 4380 by dividing by 2, 2920 by dividing by 3 and so on ...). An optimal number of K was found to be 2920 with which the algorithm is able to identify simultaneous stops and even stops reasonably close to each other to be put together.

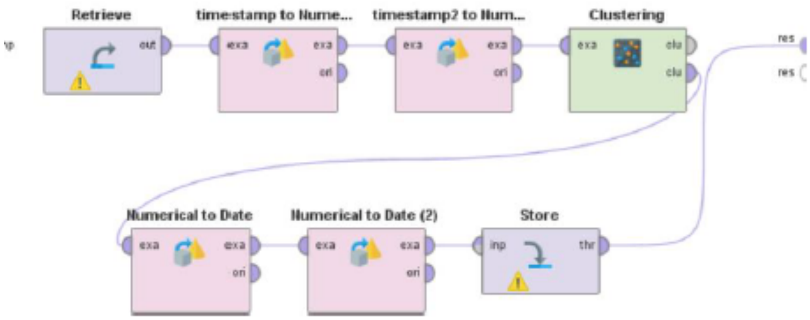

Figure 4: The clustering workflow of the vehicular stops to find vehicles that stopped together in almost one service station during their paths.

The step before the search for plates that have been together several times stopping in a service station consists in the union of the stops identified among all the generated paths. In the workflow of Figure 5 is depicted the workflow where repositories of the various paths are loaded together adding an attribute to differentiate them (called ‘service_station’), indicating the two gates within the service station is located; moreover, the routes outputs are merged into a single ExampleSet with the Append operator and, after have been sorted by the ‘plate id’ attribute.

At this point, a filter removes plates that appear only once, since this means that they have not stopped several times and therefore are not necessary; subsequently, now all the clusters containing only one record are removed since they are vehicles that have not stopped with others.

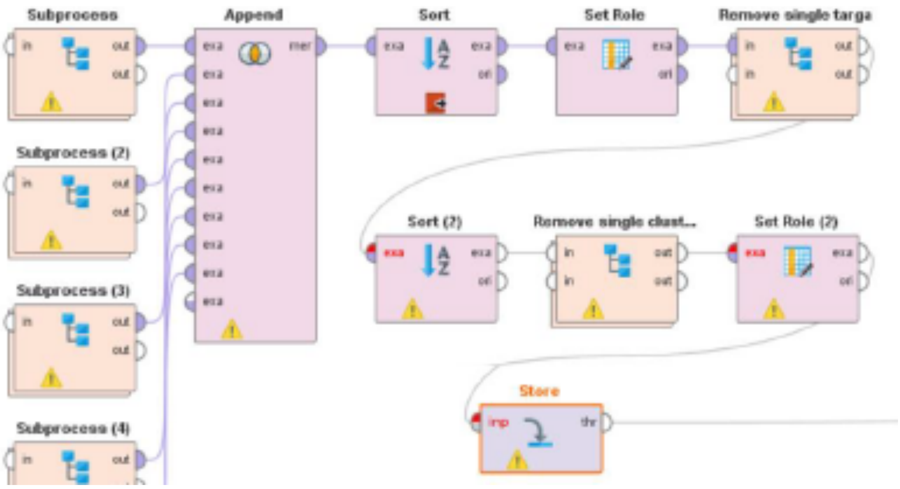

Figure 5: The workflow for the union of the generated vehicular paths that contains stops in service stations.

The final workflow depicted in Figure 6 consists in crossing all the detected stops for all the service stations during the entire year, with the aim to find all the times that vehicles stopped several times together. Starting from the previous dataset about all the generated records, a new one is created in which plates belonging to the same cluster are coupled two by two; the set is split and is given as input to a Loop operator to execute operations for each group (i.e. for each group of vehicles that have almost one stop together). In this way, finally, there will be obtained N repository, where N is the number of plates found more than once together stopped in service stations.

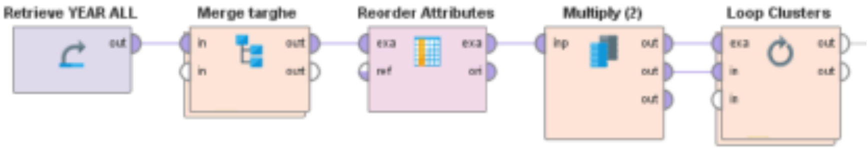

Figure 6: The final workflow “Multiple Stops”.

6 Experimental Results

Results for the classification of the vehicular paths generated exploiting the highway gate detections and that includes service areas are in Table I. The unsupervised classification consists in three clusters where the average velocity is the main feature: the ‘standard paths’ cluster (velocity between 105 and 122 Km/h) contains the 73.07% of the total while the ‘impossible’ paths (average velocity over 500 Km/h) are the 0.03%.

Path Classification	Velocity (avg)	Total Paths generated	Amount on the total
Standard	105-122 Km/h	1.797.838	73.07%
Impossible	>500 Km/h	813	0.03%
Too many stops	10-70 Km/h	661.646	26.9%

Table 1: Results of the “Vehicular paths” clustering

Moreover, emerges that paths containing too many stops (featuring a suspicious average velocity between 10 and 70 Km/h) are the 26.9% and so, together with the impossible routes cluster, can request further inspections from the Traffic Police.

Looking to the clustering process outputs, it has been found that:

- the largest cluster denotes an average velocity between 105 and 130 km/h (the standard for the Italian highway)
- the less numerous cluster features impossible speed values between 500 and 10.000 km/h; vehicular plates in this cluster require further investigation by the police and by the highway owners

In Table II it is depicted a further clustering process that takes into account the length of the generated vehicular paths, considering transits and stops through services stations: vehicles that had long journeys are 18.5% while the average ‘normal’ ones are 53.1%. Very short paths through the highway branches that could be suspicious are the 28.4%.

Route Classification	Velocity (avg)	Total Paths generated	Amount on the total
Normal	50-85 Km/h	283.043	53.1%
Long	0.5-15 Km/h	98.618	18.5%
Short	22-37 Km/h	151.358	28.4%

Table 2: Results of the “Vehicular stops” clustering.

From this second clustering stage that specializes the first one emerges that the total number of stops is 2.495 of which 2.267 (90.9%) are classified as ‘standard’ while the vehicular stops classified as ‘suspicious’ are only 228 (9.1%).

Moreover, at the end of the clustering pipeline can be noticed interesting considerations when taking into account the time waiting for all the stops in the service stations contained in a vehicular path, especially for a stop time of one hour or more: for example, there are two vehicular plates that stopped immediately following each other all the Thursday nights in the same service station (between gates 18 and 23): this fact appears suspect because the owner of plate '55302' may have left something hidden inside the service station, that the owner of plate '33811' could take soon after. Similar cases occur in the same service station where the two vehicles stationed simultaneously or when the two vehicles remain in the service areas for almost an hour at each meeting.

Further improvements can be obtained considering the stability of repetitive patterns using a similarity measure (D. Impedovo et al., 2012; Pirlo et al., 2013) and updating strategies while discovering new patterns (Pirlo et al., 2009; Impedovo et al., 2011).

7 Conclusions

This work presents a pipeline to approach in an unsupervised way the analysis of the TRAP-2017 vehicular dataset. The proposed approach builds a workflow process exploiting the RapidMiner framework able to build vehicular paths on highway branches exploiting consecutive gate detections estimating the average velocity, with the aim to highlight certain vehicular behaviors. In such way the described workflows can isolate sequential visits of service areas clustering vehicles that suspiciously stopped together several times as well as find inconsistent paths under the space-time point of view which are indicative of cloned license plates.

The use of such visual workflows that constitute a customizable automatic tool easy to use and integrate could provide great advantages to the Police forces that with human abilities are able to analyze only a small part of the daily traffic data recorded massively throughout the highway network. The provided results in fact isolate groups of suspect vehicles (and therefore worthy of further checks) certainly of much lower number than the total amount, in fact from results emerges that about 1/4 (26.9%) of the generated vehicular paths postpones suspicious stops and only 0.03% could affect vehicles with copied or altered plates.

Since the proposed approach is unsupervised through K-Means clustering, it emerges the need to validate the assumptions made during the processes; in future work it will be useful retrieving the real vehicles plates corresponding to the identifiers provided in the dataset in order to compare the unsupervised technique with the supervised one; additional data could come from services able to verify the suspected plate owner legal status. A further improvement could be made by revealing exactly which part of highway has been provided as well as the weather conditions and the highway information concerning the traffic, in order to better understand any delay or stop due to bad weather, road works or accidents, thus excluding bad and misunderstanding detections.

8 References

Bernaschi, M., Celestini, A., Guarino, S., Lombardi, F., Mastrostefano, E. (2018). *Unsupervised Classification of Routes and Plates from the Trap-2017 Dataset*. In Traffic Mining Applied to Police Activities, pp. 97–114.

Beshah, T., Hill, S., (2010). *Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia*. In AAAI Spring Symposium: Artificial Intelligence for Development.

Brambilla, M., Mascetti, P. & Mauri, A. (2017). *Comparison of different driving style analysis approaches based on trip segmentation over gps information*. In IEEE International Conference on Big Data, pp.3784–3791.

Buono, P., Legretto, A., Ferilli, S. & Angelastro,S. (2018). *A visual analytic approach to analyze highway vehicular traffic*. In 22nd International Conference Information Visualisation (IV), pp. 204–209.

Chang B.R., Tsai B.F., Young C.P. (2010). *Intelligent data fusion system for predicting vehicle collision warning using vision/GPS sensing*. In Expert Systems with Applications 37(3), pp. 2439-2450.

Dangel, U., McDonagh, P. & Murphy, L. (2014). *Micro analysis of urban vehicular data for enhanced information services for commuters*. In IEEE Vehicular Technology Conference, pp. 1–7.

Di Mauro, N., Ferilli, S. (2018). *Unsupervised lstms-based learning for anomaly detection in highway traffic data.* In Foundations of Intelligent Systems, Springer International Publishing.

Ellison, A.B., Greaves, S.P. & Bliemer, M.C. (2015). *Driver behaviour profiles for road safety analysis*. In Accident Analysis Prevention 76, pp. 118–132.

Fumarola F., Lanotte, P.F. (2017). *Exploiting recurrent neural networks for gate traffic prediction*. In Italian Conference for the Traffic Police Springer, pp. 145–15.

Impedovo, D.; Balducci, F.; Dentamaro, V.; Pirlo, G. (2019a). *Vehicular Traffic Congestion Classification by Visual Features and Deep Learning Approaches: A Comparison*. Sensors 2019, 19, 5213.

Impedovo, D.; Dentamaro, V.; Pirlo, G.; Sarcinella, L. (2019b). *TrafficWave: Generative Deep Learning Architecture for Vehicular Traffic Flow Prediction*. Appl. Sci. 2019, 9, 5504.

Impedovo, D. and Pirlo, G. (2011). *Updating Knowledge in Feedback-Based Multi-classifier Systems.* International Conference on Document Analysis and Recognition, Beijing, 2011, pp. 227-231.

Impedovo, D., Pirlo, G., Sarcinella, L., Stasolla, E., Trullo, C.A. (2012). *Analysis of stability in static signatures using cosine similarity*. Proceedings of the International Workshop on Frontiers in Handwriting Recognition, IWFHR, art. no. 6424397, pp. 231-235.

Jain, A., Ahuja, G., Mehrotra, D. (2016*). Data mining approach to analyse the road accidents in India*. In Reliability Infocom Technologies and Optimization (Trends and Future Directions), 5th International Conference on, pp.175–179.

Leuzzi F., Del Signore E., Ferranti R. (2018). *Towards a Pervasive and Predictive Traffic Police. In: Traffic Mining Applied to Police Activities. TRAP 2017*. In Advances in Intelligent Systems and Computing, vol 728. Springer, Cham. ISBN 978-3-319-75607-3.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F. (2015). *Traffic flow prediction with big data: A deep learning approach*. In IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865–873, IEEE.

Pakgohar, A., Tabrizi R.S., Khalili, M., Esmaeili, A. (2011). *The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach*. In Procedia Comput. Sci., vol.3, pp. 764–769.

Pan J., Popa I.S. and Borcea C. (2017). *DIVERT: A Distributed Vehicular Traffic Re-Routing System for Congestion Avoidance*. In IEEE Trans. on Mobile Computing, vol. 16, no. 1, pp. 58-72.

Perez-Zuriaga, A.M., Camacho-Torregrosa, F.J., Campoy-Ungria, J.M. & Garcia, A. (2013). *Application of global positioning system and questionnaires data for the study of driver behavior on two-lane rural roads*. In IET Intelligent Transport Systems 7(2), pp. 182–18.

Pirlo G. and Impedovo D. (2013). *Cosine similarity for analysis and verification of static signatures*. IET Biometrics, vol. 2, no. 4, pp. 151-158.

Pirlo, G., Trullo, C.A., Impedovo, D. (2009). *A Feedback-Based Multi-Classifier System*. 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, pp. 713-717.

Trap 2017: Traffic Police (2017). Retrieved from: https://www.poliziadistato.it/articolo/19945b448fbfdd32f820048425 (last access: 07/05/2019).

Trap event, Polizia di Stato (2019). Retrieved from: https://trap2017.poliziadistato.it (last access 07/05/2019).

Zamani, Z., Pourmand, M., Saraee, M.H. (2010). *Application of data mining in traffic management: case of city of Isfahan*. In Electronic Computer Technology (ICECT), Int. Conference on, pp. 102–106.