PAPER REVIEW:

ROAD ACCIDENT ANALYSIS WITH DATA MINING APPROACH (EVIDENCE FROM ROME)

LECTURERS: PHAUK SOKKHEY & CHAN SOPHAL

Presenter: KHUN Dararith

PAPER OVERVIEW

TITLE: Road Accident Analysis with Data Mining Approach: evidence from Rome

AUTHORS: Antonio Comi*, Antonio Polimeni, Chiara Balsamo.
Department of Enterprise Engineering, University of Rome Tor Vergata, 00118 Rome, Italy

SOURCE: https://www.sciencedirect.com/science/article/pii/S2352146522002265

ABSTRACT

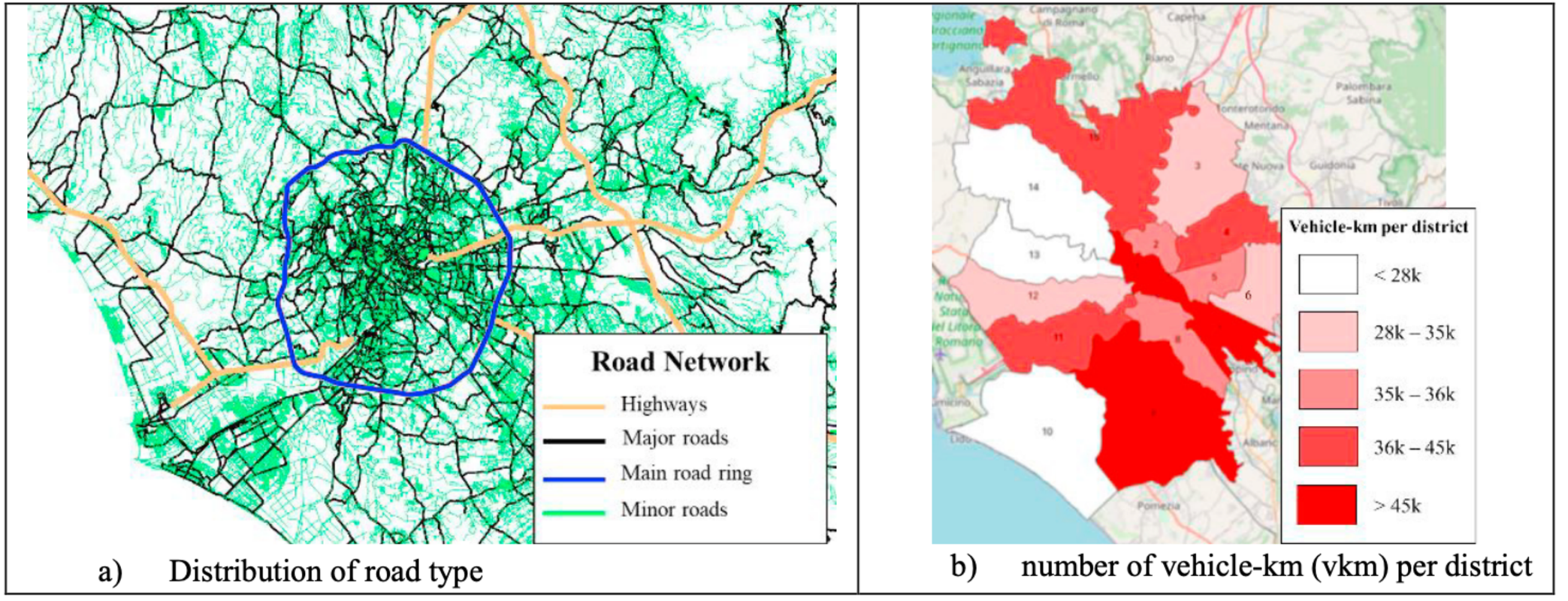
Road accident is one of the main causes of mortality worldwide. This paper was arranged to:

- Measure the accident to reduce or mitigate the accident impacts
- Identify the most effective measure
- Able to identify and classify the cause that can trigger an accident
- Study of data mining to clustering approaches to analyze accident data of 15 districts in ROME
- Analyses which tool is powerful for planning suitable measures to reduce accidents.

DATASET

The considered dataset (Open Data - Roma Capitale, 2020) consists of:

- 97,297 road accidents that occurred from 2016 to 2019 in Rome (Italy),
- Divided into 15 districts, which have an area of 1285 km2
- The resident population of nearly three million inhabitants



a) Distribution of road type

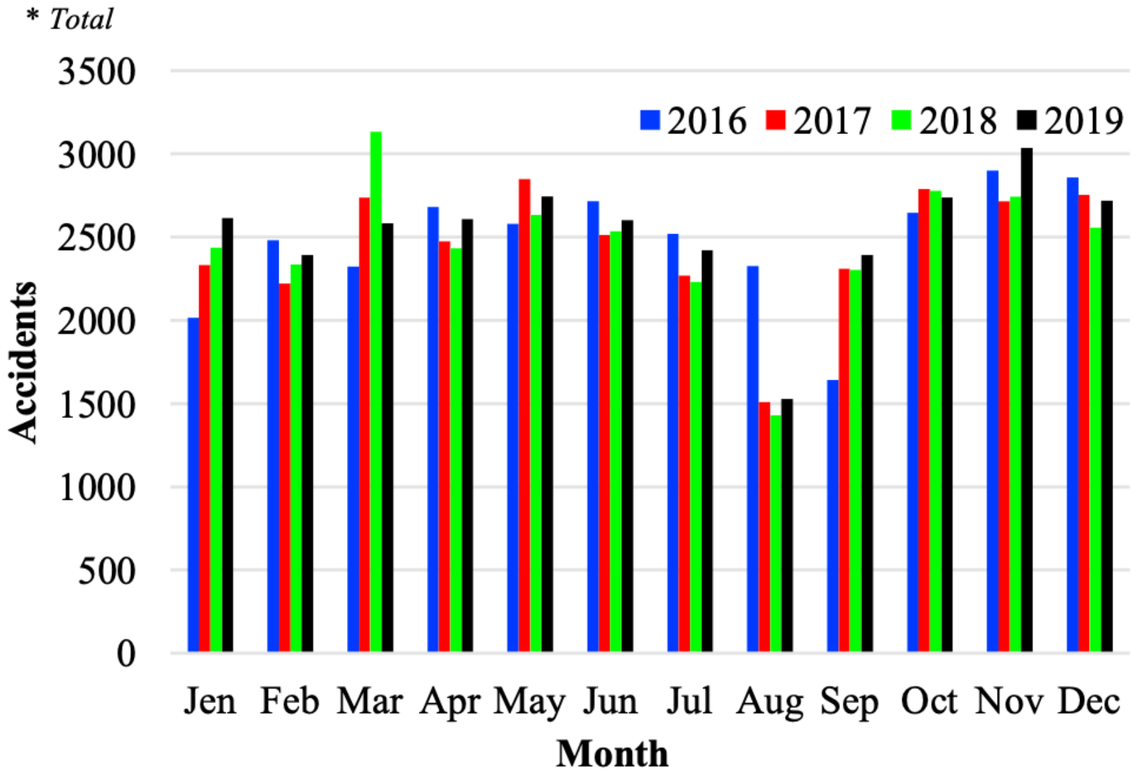
b) number of vehicle-km (vkm) per district

Fig. 1. Road network data.

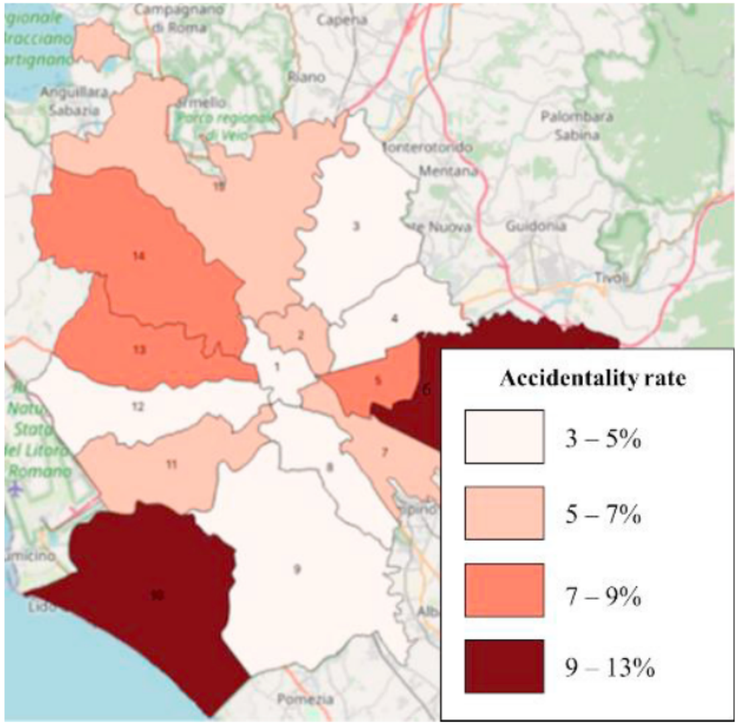
	Pedestrians & Cyclists	Two-wheeled vehicles	Four-wheeled vehicles	Public services vehicles	Heavy vehicles	TOT
Single vehicle accidents	91	2,918	11,703	319	561	**15,592**
Pedestrians & Cyclists	23	1,790	5,825	163	394	**8,195**
Two-wheeled vehicles		895	17,568	258	1,182	**19,903**
Four-wheeled vehicles			41,745	1,904	7,339	**50,988**
Public services vehicles				23	142	**165**
Heavy vehicles					347	**347**
TOTAL	**114**	**5,603**	**76,841**	**2,667**	**9,965**	**95,190**

Table 2. Types of road accidents.

Type of accident	2016 [%]	2017 [%]	2018 [%]	2019 [%]
Collision with obstacle	26.19	23.06	27.42	25.93
Rear-end collision	15.37	15.94	14.99	15.33
Side collision	22.54	23.65	22.64	23.62
Head on collision	25.72	26.86	24.85	24.99
Rollover	1.02	0.97	0.89	0.92
Pedestrian hit	6.38	6.66	6.46	6.65
Sudden braking	1.27	1.18	1.3	1.07
Vehicle fall	0.67	0.76	0.64	0.58
Run-off roadway	0.84	0.92	0.81	0.91



a) Number of accidents



b) Accidentality rate per district

Fig. 2. Accidents.

METHODOLOGY

This study employs several models to explore their performance in describing and predicting road accidents inside Rome.

- k-means clustering
- Kohonen network

which is useful to uncover clusters (i.e., groups) of data objects that are more similar to each other.

Predictive analysis:

- Decision trees,
- Association rules
- Artificial neural networks (ANNs)

CLUSTER	MONTH	TYPE OF ACCIDENT	CHARACTERISTICS OF ROAD SECTION	VEHICLE TYPE	INJURIES	DEATHS	DISTRICT
1	June	Pedestrian hit	Slope section	Unknown	1.015	0.025	7
2	March	Side collision	Slope section	Four-wheeled	0.421	0.003	11
3	June	Rear-end collision	Flat section	Four-wheeled	0.616	0.012	7
4	April	Side collision	Slope section	Four-wheeled	0.465	0.002	3
5	October	Rear-end collision	Straight section	Four-wheeled	0.459	0.002	11
6	October	Side collision	Slope section	Four-wheeled	0.475	0.003	5

Table 6. Kohonen clustering.

CLUSTER	MONTH	TYPE OF ACCIDENT	INJURIES	DEATHS	DISTRICT
1	January	Pedestrian hit	High	High	14-15
2	December	Run-off the roadway	Medium	Low	14-15
3	December	Vehicle fall	High	Low	5-6
4	December	Fall from an overpass	High	High	1-2
5	January	Side collision	Low	Low	1-2
6	January	Rear-end collision	Medium	Low	3-4

Table 8. C5.0 confusion matrix (test set).

Observed injuries	Predicted injuries: 0	1	2-3	4-5	>5
0	7113	101	0	0	0
1	1608	1171	108	1	0
2-3	267	345	87	3	0
4-5	13	42	19	0	0
>5	1	3	2	0	0

Table 9. CHAID confusion matrix (test set).

Observed injuries	Predicted injuries: 0	1	2-3	4-5	>5
0	7188	26	0	0	0
1	1651	1185	52	0	0
2-3	269	380	52	1	0
4-5	14	49	11	0	0
>5	1	4	1	0	0

Table 10. ANN confusion matrix (test set).

Observed injuries	Predicted injuries: 0	1	2-3	4-5	>5
0	7121	93	0	0	0
1	1604	1253	31	0	0
2-3	408	196	98	0	0
4-5	61	9	4	0	0
>5	5	1	0	0	0

Before training the introduced models, the dataset is divided into two sections: training/validation and test sets. Each model is trained and validated on the training set (i.e., years 2016, 2017, and 2018). Then, the models are compared in their prediction performance on the test set (i.e., 2019). C5.0 got 77%, CHAID 76% and ANN 77,8%

CONCLUSION

This paper had the purpose to find out which data mining techniques are more suitable to analyze road accidents. The selection of these techniques is based on a review of the state of the art.

As the result:

- k-means and Kohonen network is advantageous for descriptive analysis
- Decision trees and neural networks are useful for predictive analysis

Result of applying Data mining:

- Identify characteristics of the roads
- A new study of road safety
- Recognizing accident patterns without the “noise”

was suggested to use hybrid prediction approch

THANK YOU!