TIME-SERIES ANALYSIS OF UK TRAFFIC ACCIDENT DATA

Peter Ljubič, Ljupčo Todorovski, Nada Lavrač
Department of Intelligent Systems
Institute Jožef Stefan
Jamova 39, SI-1000 Ljubljana, Slovenia
Tel: +386 1 4773419; fax: +386 1 4251038
e-mail: peter.ljubic@ijs.si

John C Bullas
AA Foundation for Road Safety Research / CSS
Hampshire County Council
Winchester, SO23 7RX, United Kingdom
Tel+44 7946 317 467, Fax +44 1962 873761
e-mail: john.bullas@hants.gov.uk

ABSTRACT

This paper presents the results of the analysis of a large database containing UK personal injury traffic accident data. The first goal of the analysis was to provide a better understanding of the data through the use of simple statistics and visualization. The second goal was to find some trends and patterns in the dynamic change of the number of traffic accidents during time.

1 INTRODUCTION

Papers presenting the development of new ML and pattern recognition methods often report the results of tests on 'perfect' datasets. The test datasets are perfect in the sense that they are usually small (in terms of the number of examples and features), they are not noisy, and the developers know what kind of patterns they are searching for. When applying these methods to real-world datasets, one faces a whole spectrum of problems, e.g., what to do with missing or irregular values, and how to evaluate the results when there is no reference data available.

In this paper, we deal with a large real-world database about personal injury traffic accidents in the UK for the period of 21 years from 1979 to 1999. All the problems mentioned above are very present in the task of analysis of this database: the database is huge, there are a lot of noisy/missing data and probably the most important problem - there are not clearly defined expectations about the kind of patterns we are going to discover. Therefore, we decided here to present the results of the preliminary analysis of the dataset that can help us to get a better understanding of the data. First, we applied basic statistical analysis and visualization of the data. Second, we performed clustering of time-series to discover frequent patterns in the dynamic change of the number of traffic accidents through the time.

This paper is organized as follows. In Section 2, the structure and organization of the UK traffic database is presented. The data analysis methods used along with the results of their application to the UK traffic database are presented in Section 3. Evaluation of results done by the end-user is described in Section 4. Finally, Section 5 concludes the paper and proposes directions for further work.

2 PRESENTATION OF THE PROBLEM DOMAIN AND DATA UNDERSTANDING

The UK traffic database contains data about personal injury road accidents in the UK and their consequent casualties for the period of 21 years, from 1979 to 1999, obtained for the Sol-Eu-Net project (http://soleunet.ijs.si) [1].

The datasets is organized in three tables: Accident table, Vehicle table and Casualty table. Accident table contains one record for each accident. The 30 attributes describing an accident can be divided in three groups: date and time when the accident has occurred, description of the road where the accident has occurred, and conditions at which the accident has occurred (such as weather conditions, light and junction details). In Accident table there are more than 5 millions of records. Vehicle table contains one record for each vehicle involved in the accident from the Accident table. There can be one or many vehicles involved in a single accident. Vehicle table attributes are describing the type of the vehicle, maneuver and direction of the vehicle (from and to), vehicle location on the road, junction location at impact, sex and age of the driver, alcohol test results, damage on a vehicle, and the object that vehicle hit on and off carriageway. There are 24 attributes in the

Vehicle table which contains almost 9 millions of records. The third, Casualty table contains records about casualties for each of the vehicles in the Vehicle table. There can be one or more casualties per vehicle. The Vehicle table contains 16 attributes describing sex and age of casualty, type of casualty (e.g. pedestrian, cyclist, car occupant etc.), severity of casualty, if casualty type is pedestrian, what were his/her characteristics (location, movement, direction). This table contains almost 7 millions of records.

The domain expert specified the following data mining goals:

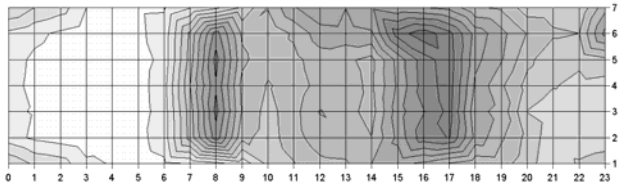
- Developing accidents occurrence models based on the following variables: Road Surface Condition, Skidding, Location, Street Lighting in order to inspect the influence of these variables on the occurrence of accidents;
- Trend analysis within a long time period;
- Finding some particular types of accidents that become more prevalent by using ML approaches, such as interesting groups discovery and clustering;
- Correlation analysis between accident characteristics and age of drives, speed of cars and so on.

This paper addresses mainly the problem of trend analysis.

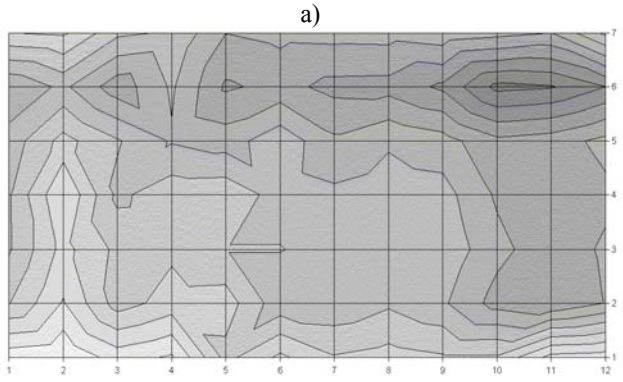
3 METHODS

3.1 Simple statistics and visualization

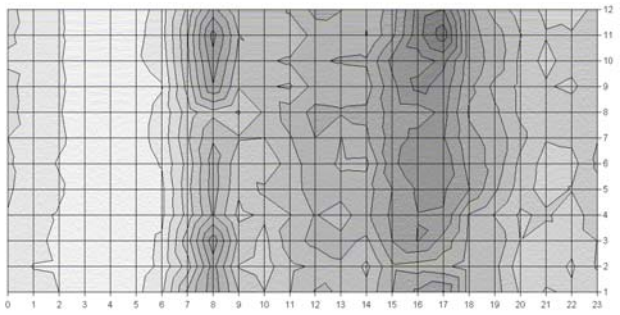
First, we applied some basic and simple statistical analysis in order to understand the data better. We calculated distributions of different attributes, such as the distribution of age of drivers, distribution of accidents over years etc. Here, we decided to present the distributions of the number of accidents over different time periods. Even for this simple analysis, we decided to make it on a representative sample of 60.000 accidents, since there are too many records for such an analysis in the original database. Results are shown in Figure 1.



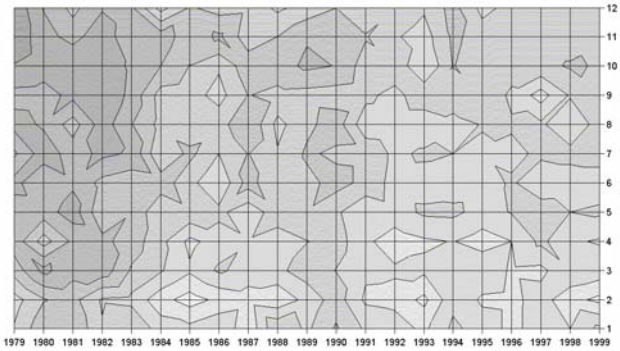
a)



b)



c)



d)

Figure 1: T*wo-dimensional distributions of the number of accidents over different time periods. a) Hour/day of week (1 = Sunday, 2 = Monday, ...). b) Month/day of week. c) Hour/month. d) Year/month. Darker shades of gray represent higher number of accidents.*

In Figure 1a) we can notice a high number of accidents (darker color) at 8am and between 16pm and 17pm during the working days, and a peak of the number of accidents at 16pm on Friday. Figure 2b) tells us that most of the accidents happen in October and November on Friday, and that there are many more accidents on Fridays and at the end of the year. The distribution in Figure 1c) resembles the one in Figure 1a), where 8am and 17pm are critical points, especially at the end of the year. Finally Figure 1d) illustrates the decreasing number of accidents through the years and again reveals the fact that there is a slight increase of the number of accidents at the end of each year.

3.2 Qualitative clustering of short-time series

Finding all these interesting patterns analysing different time periods lead us to the idea to further analyze the dynamic change of the number of accidents through years, months of year and hours of day. More precisely, we tried to identify spatial areas in the UK, where the patterns of change of the number of accidents are similar. Namely, the data about each accident contains also information about the police authority area where the accident occurred. In the UK, there are 51 police force authorities. Therefore, in

our second analysis task, we used clustering to discover groups of police authorities with similar temporal change of the number of accidents. The analysis is based on qualitative clustering of short time-series methodology presented in [2]. What follows is a brief presentation of the methodology.

Clustering is an unsupervised learning method [3]. Given data about a set of objects, a clustering algorithm creates groups of objects following two criteria. First, objects should be close (or similar) to the other objects from the same group (internal cohesion) and distant (or dissimilar) from objects in the other groups (external isolation). One of the most important aspects in clustering is how the distance between two objects is measured, i.e., the distance of two time series in our case. The distance measures used for clustering time-series are usually based on correlation [4]. However, the correlation based distance measure has two drawbacks. First, it is well known that the correlation coefficient is very poorly estimated when we have a small number of observations (i.e., short time series). Second, it is capable of capturing only the linear aspect of the dependence or relation between time series. Two time series that are non-linearly related to each other will be distant from each other, regardless of the similarity of their dynamic change through time. In order to overcome the drawbacks of the correlation based distance measures, we used an alternative distance measure that is based on a qualitative analysis and comparison of the shape of the time-series.

Consider two time series X and Y. We choose a pair of time points *i* and *j* and we observe the qualitative change of the values *X* and *Y*. Three possible values of qualitative change $q(X_i, X_j)$ can be distinguished (note that $i < j$):

- Increase, when $X_i > X_j$,
- No-change, when $X_i = X_j$,
- Decrease, when $X_i < X_j$.

We calculate the qualitative difference between *X* and *Y* by summing up the differences for all the pairs of time points:

$$D_q(X,Y) = \frac{1}{N(N-1)} \sum_{i<j} Diff(q(X_i, X_j), q(Y_i, Y_j)),$$

Where *Diff(q1, q2)* is a simple function that defines the difference between different qualitative changes, defined in Table 1.

$Diff(q_1, q_2)$	Increase	No-change	Decrease
Increase	0	0.5	1
No-change	0.5	0	0.5
Decrease	1	0.5	0

Table 1: *Definition of the Diff function.*

Before running the clustering algorithm, we had to prepare the time-series data. We had to count the number of accidents in each year in the observed period from 1979 to 1999. This was done for each police force authority, to get the time-series data structured as in Table 2. For each police force authority we obtained time series of length 21 that reflect the number of years in the observed period. We applied a similar data transformation in order to get: (a) time-series of length 12, reflecting the change of the number of accidents through months of year, (b) time-series of length 7, reflecting the change of the number of accidents through day of week and (c) time-series of length 24, reflecting the change of the number of accidents through hour of day.

Year / *Pfc*	*1979*	*1980*	*1981*	...	*1998*	*1999*
xxx	521	552	467	...	338	476
yyy	343	343	343	...	454	454

Table 2: Format of *data preparation for qualitative clustering of short-time series – the number of accidents per police force authority through years.*

4 RESULTS

By applying the clustering algorithm to the data from Table 2, we got groups (clusters) of police force authorities with most similar patterns of dynamic change of the number of accidents (patterns of dynamics). Police force authorities are represented with police force codes from 1 to 51. Results of the clustering are the following:

- The number of accidents through *years*. We obtained two clusters, which means that we have two types of dynamics. The first cluster contains 40 authorities with mainly decreasing trend in the number of accidents, whereas the second cluster contains 11 authorities with mainly increasing trend. Both clusters are presented in Figure 2: the first cluster is on the left-hand side and the second cluster is on the right-hand side of the figure. We also compared the results of clustering with the results of a straightforward application of linear regression to the time-series of each police authority. We can observe that the first cluster contains authorities with a lower (mainly negative) slope of the linear regression line, whereas the second cluster contains police authorities with a higher (mainly positive) slope.
- The number of accidents through *month of year*. The clustering algorithm discovered four clusters of authorities. In the first cluster (11 authorities) there are authorities with a peak of the number of accidents during the summer and low number of accidents at the beginning and the end of the year. The second cluster (36 authorities) contains authorities with the highest number of accidents at the end of the year. In the third cluster (2 authorities) there are two authorities with three peaks, one in February, the second in summer time, and the third in September. The fourth cluster (2 authorities) contains two authorities with peaks in February and at the end of the year. Typical

representatives of the four clusters are presented in Figure 3.

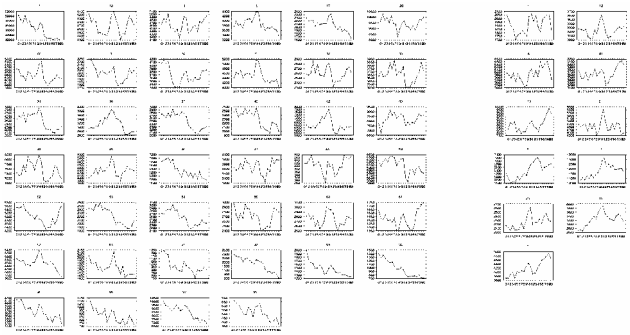


Figure 2: *Two clusters of police authorities with similar dynamics through years; left six columns represent the first cluster with a decreasing tendency, right two columns represent the second cluster with an increasing tendency.*

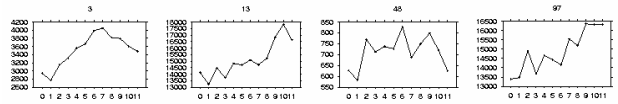


Figure 3: *Typical representatives of four clusters with similar dynamics through months of year; from left to right are clusters one to four.*

- The number of accidents through *day of week.* There were four clusters discovered. The first cluster containing 2 authorities with very low number of accidents during the week, but with very high peak at the end of the week. Trend of the authorities from the second cluster (29 authorities) has two peaks, lower one on Monday and higher one on Friday. The third cluster (11 authorities) contains authorities with high peak on Friday, while the rest of the week there is even distribution of the number of accidents. The fourth cluster (9 authorities) contains authorities with increasing trend of the number of accidents with peak on Friday.
- The number of accidents through *hour of day*. Since all police authorities have almost the same distribution of accidents through hours, the algorithm found one cluster only, so there is nothing to discuss.

5 EVALUATION

Clustering was done on police force authorities' dynamics through different time measures. Clustering through the years divides police force authorities with decreasing and increasing dynamics. Some were found to have increasing trends:

(i) Through the months,
(ii) With peak in summer,
(iii) Peak at the end of year, and
(iv) Increase through day of week.

The phenomenon that we would like to further analyze is why are the numbers of accident through years increasing for some areas? May this be due to the increase in Traffic or Population, and/or do improvements in Reporting procedures play a role?

The preliminary analysis indicates that tourist impact in summer time may explain mid-year peaks for certain areas with relatively lower traffic outside of the holiday season. Some authorities have a very small number of accidents during the daylight hours, except Friday and Saturday when there are a lot of accidents. Some authorities are very industrial, where there are quite a lot of accidents during the week and with peak at the end of week (Friday). Dynamics through hours is not shown since all fall into one cluster.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we used a clustering approach to the analysis of time series data about the dynamic change of the number of accidents through different time segments. We used hierarchical clustering to obtain groups of police authorities with similar patterns of trends through these time segments. A qualitative measure of distance between time series was used, which is suitable for very short time series, where the use of correlation-based distance is not appropriate.

The preliminary analysis of the obtained clusters shows the usability of the approach. The number of clusters is small and the discovered clusters make sense (e.g. authorities located at the coast, industrial authorities, etc.). An in-depth analysis of results is planned in further work. We will also investigate whether to refine the clustering methodology by introducing more fine-grained set of values of qualitative change, potentially leading to finer clusters.

It must be mentioned that, though desirable in assisting the understanding of this work, the representation of locations on maps and/or the naming of the police force areas falling within the categories described could not be presented due to the terms of the Data Mining Agreement.

References

[1] D. Mladenić. EU project: Data Mining and Decision Support for Business Competitiveness: a European Virtual Enterprise (Sol-Eu-Net). In *OES-SEO 2001: Open enterprise solutions: Systems, experiences and organizations*. Roma: LUISS, pp. 172–173.

[2] L. Todorovski, B. Cestnik, M. Kline, N. Lavrač, S. Džeroski. Qualitative Clustering of Short Time-Series: A Case Study of Enterprises Reputation Data. In: *Proc. IDDM Workshop on Integration Aspects of Data Mining and Decision Support*. Helsinki. 2002.

[3] J. A. Hartigan. *Cluster Algorithms.* Wiley, 1975.

[4] P. Ormerod, C. Mounfield. Localised Structure in the Temporal Evolution of Asset Prices. Proc. New Approaches in Financial Economic Conference. Santa Fe, NM. 2000.