THE ARCHIVES OF TRANSPORT
VOL. XXIV NO 1 2012
10.2478/v10174-012-0006-4

Road Traffic Accident Patterns: A Conceptual Grouping Approach to Evaluate Crash Clusters

Marzena Nowakowska*

Received July 2011

Abstract

The aim of the work is to highlight road traffic accident patterns in the context of interrelations between road characteristics and a traffic safety threat. The actual data concerning multi-vehicle accidents without pedestrians on non-urban roads in a chosen region of Poland was the subject of the research. The roadway and roadside data at the accident site have been combined with the crash data that define the roadway hazard, i.e. driver's behaviour, type and accident severity. The data were subject to multivariate segmentation by means of such conceptual grouping techniques as the K-means clustering algorithm and competitive artificial neural networks. The Ward's method was used as a supporting tool in establishing the final number of accident profiles. Six distinct accident patterns have been recognised, quantified and labelled, where the first, second and third one are typical of rural areas, the fourth and fifth – of built-up areas, and the last one – of intersections. The analysis indicates that apart from threat factors, the following road related features play an important role in road accident profiling tasks: area type and area development level, roadway surface condition, intersection indicator, shoulder type, and also to some extent: lighting conditions, shoulders' width, and horizontal curve radius.

1. Introduction

National roads in Poland serve as interregional connections between important administration, economic and tourist centres. They have the highest project parameters of all undivided, two-directional roads with two lanes, one in each direction, which enables vehicles to achieve the highest speeds. Because of the amount, the

* Kielce University of Technology, Faculty of Management and Computer Modelling, Al. 1000-lecia Państwa Polskiego 3, 25-314 Kielce, POLAND

structure, and the type of traffic, national roads constitute a higher level of threat than other most common road categories, that is provincial, district and communal [18]. Some aspects of threat on the national roads in the Świętokrzyskie province, in the south of Poland, are investigated in the study. Some aspects of threat on the national roads in the Świętokrzyskie province, Świętokrzyskie province, in the south of Poland, are investigated in the study.

Supervised modelling techniques are used in the vast majority of research works on the analysis of road factors role in the transportation system safety. They incorporate statistical as well as modern machine learning tools to determine the value of the target variable on the basis of a set of confounding factors. If the explained variable has a numeric character (e.g. annual accident frequency, crash rate on chosen road sections, number of vehicles involved in a certain accident type) quantitative models are built, like for example in [1, 8, 21]. If the explained variable has a descriptive character (e.g. accident severity) qualitative models are utilised, like for example in [6, 22, 35]. In both cases a certain level of aggregation is usually adopted for the variables that describe design parameters of the road, for example the curvature of the road section where the accident was registered.

The presented study focuses on the accident pattern recognition with respect to the connection of road features with a road traffic hazard. Contrary to other works, the research deals with roadway and roadside factors at the accident site. The threat is expressed by three fundamental accident characteristics: behaviour of at-fault driver, accident type and accident severity. A conceptual grouping approach is employed, owing to which, instead of defining judgmentally the segments in the dataset using one or two variables, the multivariate segmentation and therefore the management of complex information in a nutshell is possible. Unsupervised learning techniques are used.

The data clustering concepts are rarely present in road safety analyses. However, there are some works in which detecting accident profiles is the direct subject of the study or it is an indirect way for further, usually supervised, analysis.

In the work [33] the analysis is primarily descriptive and includes single frequency distributions as well as two- and three-way cross-tabulation of the crash type using other variables of interest. The pedestrian crash data, from six U.S. states, concerning human-related factors with severity and roadway characteristics such as: road system, crash location, and lighting conditions are investigated. The typology is based on over-representation and under-representation in contingency tables. Because of the sixteen discussed categories of the accident type and a large road network area, the crash situations can vary from state to state and any generalisation can be difficult.

According to the problem formulation, Baltes categorizes Florida pedestrian crashes in his work [2]. However, it is not quite clear whether the author investigates the data concerning road accident pedestrian casualties or the data concerning crashes involving pedestrians (in the latter case, the number of accidents can be smaller). In the case of a crash, it is not explained how a human feature (e.g. age) is

attributed to the crash in which more than one pedestrian were injured. Similarly to the previous work, the analysis is based on over-involvement or under-involvement in crashes detected in up to three cross-classifications. Non-road factors as well as road factors such as lighting conditions, number of lines, road system identifier, area type, and road surface conditions were examined. The findings have indicated the influence of human factors mainly, and then interrelations with area type, intersection and part of day.

The quasi-induced exposure technique was used to analyse the relationship between driver, roadway and environmental factors in crashes on low-volume roads in [32]. An index calculated for contingency tables allowed drawing some conclusions concerning the road factors. The interrelations between road environmental factors and crashes depend generally on whether one or two vehicles were involved in the accident. They can differ within groups under discussion. For example: (1) a single-vehicle crash propensity decreases as a lane width increases, (2) a speed limit does not have a significant effect on two-vehicle crashes.

From the methodological point of view the work by Dixon et al. is worth mentioning [7]. The solution to the problem of multivariate analysis of roadside conditions that posed a greater risk was presented. The a priori common profiles of fixed-object crash features were defined. Then contingency tables of crash severity distribution were analysed within each profile.

Pande and Abdel-Aty presented another method in discovering patterns in road accident data. They used market basket analysis to find direct [27] and indirect [28] associations among features that described accidents on roads in Florida. The combination of descriptive human, environmental and roadway characteristics defined the dataset under investigation. Association rules were discussed in both works. It was concluded in the first study that: (1) there is a significant correlation between lack of illumination and high severity crashes, (2) under rainy conditions straight road sections with a vertical curve are particularly crash prone. In the other work two patterns were discovered which, in addition, were not identified in the first study. However, one indirect association resulted from a simplified database structure, which did not allow registering both the straight grade and curve level for a single accident. The other association indicated intersection indicator as a remarkable probe attribute. It should be pointed out that it was impossible to discover any associations concerning road crossing, as the study in the first work was devoted only to non-intersection crashes.

The author herself also employed association analysis in [25]. The road accident data included fifteen mixed factors, similar to these in the works presented above. Two- and three element rules were interpreted, as more-element rules had too weak association measures. The methodology turned out to be a very useful tool in discovering similarities and differences between patterns for crashes caused by drivers and pedestrians, as well as in quantifying the pattern importance.

A cluster analysis was proposed in [37] to categorise over a hundred of road safety strategies implemented in Hong Kong into distinct groups according to their im-

plementation schedules. Seventeen clusters were identified using the Ward's method, and then they were analysed to assess the safety strategy effects. The work is an example of indirect cluster analysis applied to the road safety investigation. However, from the methodological point of view, there was no information on the structure of the dataset processed in the clustering algorithm. No solutions to establish the final number of clusters were presented. It is also not clear why the authors employed hierarchical and not non-hierarchical clustering method in the case of a common problem of multivariate segmentation.

Das et al. also used the cluster analysis as a supporting toll in modelling the accident severity [5]. They grouped urban/sub-urban state road corridors in Florida of variable lengths on the basis of road design parameters. Partitioning around medoids algorithm was applied. The study seems interesting in terms of the road factor accident patterns detection. However, the authors presented neither the list of the design variables taken to the cluster analysis nor the domains of the variables. It was not even explained whether the domains were continuous or categorised. A reader can only assume that the design factors referred not to the crash site but to an administrative road unit where an accident took place.

Hanowski et al. presented an approach, which identified infrastructure-related and non-infrastructure related problems on an urban road network [11]. Incident clusters were defined as incidents of similar characteristics occurring on a concentrated area. The grouping was done on the basis of sketches and drawings of the critical incidents at a chosen site. Although the authors underline the strength of the presented approach that allows one to look beyond the apparent factors of any single event, the method seems to be exhaustive and micro-scale specific.

Wong and Chung adopted rough set theory and statistical tests to derive rules for grouping single auto-vehicle accidents [36]. Over 20 features were used to describe the dataset. Half of them were quantitative attributes concerning roadway conditions and the others were human factors. Three clusters were determined, and only the licence type and the roadside marking were non-significant condition attributes. However, only one cluster appeared to be substantially different from the others as well as from the whole dataset. No obvious causes in the road or environment patterns were found, except for the road surface. It seems that human features overwhelmed road factors. It is also possible that the dataset, i.e. width (23 variables) and length (2311 observations), is too broad to obtain distinct patterns just in three clusters.

Rough sets were also used by Kim et al. to investigate the circumstances related to hit and run crashes in Hawaii [12]. 20 different descriptive attributes referring to drivers, vehicles, locations, and time defined the dataset of 1342 observations. Four clusters were obtained in the form of decision rules. The authors claim that they aimed not to get a perfect fit but to determine important attributes for a further supervised analysis in their study. Therefore no profiles were described. It turned out that all the attributes were important in generating the rules. The conclusion that the rough set analysis is robust and reliable can be disputable, as the logistic

regression conducted by the authors further on, indicated ten factors as significant in classifying the target variable, out of which as much as seven were human factors. A suggestion that the K-means algorithm was a possible better cluster analysis technique was made elsewhere in the work. The authors referred to their earlier paper [13], in which the non-hierarchical clustering method was used to analyse and visualise spatial patterns of pedestrian-involved accidents in Honolulu. The purpose of the work was to examine the K-means technique in terms of advantages and disadvantages in the investigation of road crashes, indicating that the method is a useful spatial analytic tool for safety research.

In order to obtain typical patterns of road-traffic accidents during driving practice in Sweden, the Hierarchical Ascendant Classification was applied in [3]. They investigated 1081 records, both single and multiple vehicle accident records, taking into account some human and environmental features, and road-relating factors such as area type, speed limit, accident site, and country region. The following four patterns were identified: 1) two patterns for rural areas with straight stretches that differ in speed limit, 2) two patterns for built-up areas, low-speed related – one with rear-end crashes and the other at road junctions. Some over-representation was noticed in fatal/severe crashes in rural accident patterns and in PDO crashes in built-up area patterns.

It seems that the potentiality of data clustering is still not fully appreciated and utilized. The cross-tabulation method in accident data grouping is slightly limited because no more than three dimensions at a time are usually analysed. Also defining a priori accident profiles can be biased by non-objectivism. Multidimensional pattern recognition techniques are an alternative not only in the investigation of vehicular accident pattern aetiology but also in diagnosing road safety problems for the local authority and the road administration, in particular when roadway and roadside factors are investigated.

The author has not found any segmentation of the crash data combined with the data concerning road characteristics at the accident site. The aim of this paper is to investigate road factors that characterise the site of the accident and their correlations with the road traffic threat. The research is undertaken to explore both methodologies and the road traffic safety issue findings, although being focused on a certain region of Poland. Therefore it may be interesting for researchers in other places. There are the following objectives of the study:

– to explore and compare the methods applied to conceptual clustering,
– to find accident patterns from a road-related elements perspective using conceptual grouping techniques,
– to detect which road-related elements define crash profiles and how.

2. Methodology

Some techniques of conceptual grouping [4, 10, 44] have been employed for the same dataset in order to compare and verify the results. Each method divides events under investigation into a given a priori number of groups in such a way that each observation belongs to only one cluster that defines an accident profile. The clusters define the road accidents profiles by discrete or continuously distributed features.

The first one is the K-means method, which belongs to non-hierarchical algorithms of cluster analysis [20, 43]. Dataset partitions are created so that all members of each subset could be similar according to the Euclidean metrics. The K-means method minimizes the sum of the within cluster variation (in consequence maximising the variation between classes) of the K partitioning of a multidimensional dataset [15, 17]:

$$WCV(K) = \sum_{k=1}^{K} \sum_{x_j \in C_k} d(x_j, \bar{x}_k)^2 \tag{1}$$

where: $d(x_j, x_k)$ is the distance between the x_j observation and the centroid $\bar{x}_k$ (the mean) of the C_k cluster.

The basic idea of the K-means algorithm is as follows. It starts with a random, initial partition and keeps reassigning the samples to clusters, based on the similarity between samples and clusters, until a convergence criterion is met, i.e. when there is no reassignment of any observation from one group to another that causes a decrease in the total squared error.

The next method is Kohonen SOM – Self-Organizing Map [14, 26, 40, 45]. This is an unsupervised artificial neural network. Such a network is a specific neurocomputing idea in which there is no teacher, i.e. an output signal is unknown. The network discovers its patterns, regularities, and classes in multidimensional datasets on its own.

The Kohonen map consists of input and output layers only. The width of the processed dataset defines the input nodes to the network. The output layer is usually arranged in a rectangular grid of nodes. In a training process, the input vectors that are closer to an output node (the winner) reinforce the weights to the given output, whereas input vectors further away turn on other output nodes and corresponding weights. The competitive training WTM (Winner Takes Most) algorithm is utilised according to the formula [14, 26]:

$$w_r(l+1) = \begin{cases} w_r(l) + \eta(l)\gamma(i,r)(x_j\text{-}w_r(l)) & \text{for} \quad r \in \Omega_i \\ 0 & \text{for} \quad r \notin \Omega_i \end{cases} \tag{2}$$

where: Ω_i is the neighbourhood of the winner neuron i, which can be interpreted as the set of output indexes closest to the winner element. The learning-rate factor $\eta(l)$ is a function that decreases monotonically with time – the l iteration step. The

function $\gamma(i, r)$ is called the neighbourhood function and it defines the distance of the $r-$th output neuron from the winner i. In this study the Euclidean metric is used to calculate the distances. In each stage the self-organization requires the indication of the winner i, i.e. the output neuron with the weight vector that differs from the input vector x_j least of all.

A certain pattern is associated with each output node. The Kohonen SOM is topological: similar clusters are closer to each other in the grid than more dissimilar ones.

Two versions of SOMs [15, 44, 45] were explored in the study:

– the incremental learning algorithm, in which the weights are updated after each presentation of a sample,
– the batch learning, in which all samples are available prior to computation and the weights are updated after the presentation of each epoch. It is considered to be faster and more stable.

There are no guidelines for choosing the final number of groups when using iterative partition-clustering procedures. A bargain of granularity is necessary to define the level of detail contained in a unit of data [19]. The more details there are, the lower the level of granularity. The fewer details there are, the higher level of granularity. Thus, in deciding about the proper number of clusters the following points have been considered:

– avoid too few clusters; otherwise a large variation in each cluster is allowed and summing up characteristics (generalizations) for a cluster do not provide much information,
– avoid too many clusters; otherwise essentially the same segments are attributed to many different clusters and any generalizations can be misinterpreted.

To find the optimal number of clusters the hierarchical Ward's clustering method is applied for a sample of data before the conceptual grouping is performed [9, 39, 42]. A peculiar aspect of this stage is that the optimal number of clusters is chosen with respect to a test statistic known as the cubic clustering criterion *CCC* [29]:

$$CCC = S \cdot \ln\left(\frac{1\text{-}\ E(R^2)}{1\text{-}R^2}\right) \tag{3}$$

The *CCC* criterion compares the observed proportion of variance R^2accounted for by the clusters (see the definition below) to the approximated expected variance $E(R^2)$ calculated under the assumption that the data of observations is selected randomly according to a uniform distribution for each variable. The S element is a multiplier that stabilizes the variance across different number of observation, variables and clusters. The final number of groups in the dataset segmentation procedures is stated as the number that corresponds to the smallest value of the *CCC* statistic greater than 3 in the plot of the *CCC* value against the number of clusters obtained from the Ward's method.

The *CCC* index calculated for the conceptual grouping results can also be treated as the measure of the separation quality. The separation is satisfactory when

the *CCC* index is greater than 3. In addition, the greater the *CCC* value is, the better the results are.

The quality of data segmentation can be assessed by other measures as well. Except for the mentioned above the within cluster variation *WCV* and the *CCC* criterion, the R^2 and *PSF* statistics [9, 15, 41] are also used in the study.

The R^2 index is derived from the idea of having a low internal cohesion W and a high external separation B. Thus its synthetic form is expressed by the following formula:

$$R^2 = \frac{B}{B + W} \tag{4}$$

The closer to unity the value of R^2 is, the better the group separation is. However, is should be stated that for $R^2 = 1$ there are as many groups as observations.

The pseudo-F *PSF* criterion accompanies R^2 and measures the quality of N samples separation at the c level:

$$PSF = \frac{B}{c - 1} \cdot \frac{N - c}{W} \tag{5}$$

The greater the value of the *PSF* statistic is, the more the average values of the clusters' vectors differ one from another.

To evaluate the role of each variable V_i in the dataset partition the *Importance* index derived from the decision trees techniques was utilised [31]:

$$Importance(V_i) = \frac{AI(V_i)}{\max\limits_{j}\{AI(V_j)\}} \tag{6}$$

where $AI(V_i)$ is the absolute importance value of the attribute V_i calculated for the decision tree in which the cluster identifier is a classified decision. For the most important variable the *Importance* index is equal to unity whereas for the least important attribute it is equal to zero.

There is no unequivocal criterion for evaluating the results of conceptual grouping but a wide range of criteria. Their use should strike a balance between simplicity and information content.

3. Data Preparation

The research is carried out on the non-pedestrian multi-vehicle accident data and the roadway data in the Świętokrzyskie province in Poland from the collection time period 2004-2007. The region is in charge of one road administration and that is considered as homogenous in terms of geography and climate. The investigation concerns only accidents on undivided, two-lane, two-directional roads that do not run via towns with civil rights. The accident data come from the Police Road Accident Database, whereas the road data come from a variety of sources:

the Computer Road Dataset Bank, the paper road documentation and the on-site visits of important places. Data cleaning, including crosschecking, was processed prior to the investigation. This preparation to the data investigation has been of the particular concern of the author [22, 23]. Own elaborated procedures and computer programs were employed to support the process. Only non-missing data observations concerning accidents with drivers at fault are considered in the study.

The road number and the kilometrage of each road accident site with accuracy of 100 m, are registered in the police database. It enabled characterizing the accident using road design parameters and road neighbourhood features on the accident occurrence site. The site is a road section defined by the area surrounding the accident site within a 100-metre radius. The side of the roadway on which a certain road element (e.g. pavement) is located is ignored, as in the police database there is no information about the direction of the vehicle at fault. The information is only used to state whether the element is present or not. Own computer programs have been elaborated, which enabled joining the data concerning the accident with the data concerning the road on the accident site.

There are some issues that should be taken into account while processing conceptual grouping. Clustering algorithms are sensitive to strong correlation. In addition, the variables with greater variances have greater influence on the final results than the variables with small variances. Such a situation can arise when analysed features do not have the same measures, or their domains are distant from one another, or they have different types (numerical and nominal). What is more, the K-means algorithm is very sensitive to noise and outlier data points. Therefore, to prepare the data for the analysis, the following preliminary processing has been conducted:

1. outlier observations were removed,
2. rare values of categorical variables were aggregated taking into account their merits,
3. qualitative features were coded to obtain their numerical equivalents by introducing dummy zero-unity variables,
4. all quantitative variables were min-max normalised, thank to which the domains of all analysed attributes are uniform,
5. the correlations were checked with the results as follows: the bivariate correlations do not exceed 0.8, thus not being too high [34] and the conditioning index does not exceed 6, indicating no strong collinearity [30, 34].

The obtained dataset *CrshRd_MltVhcl* (*Crash and Roadway factors for Multi-Vehicle accidents)* consists of 653 records. The dataset is characterised by the following variables:

- *PvPrs*; the pavement presence indicator with the values:
  P0 – no pavements,
  P1 – pavement on one side of a roadway,
  P2 – pavements on both sides of a roadway.

- *ArTp*; the area type indicator with the values: *NBt* – non built-up (rural) area,

  Bt – built-up area,

- *LgCnd*; the lighting conditions with the values: *NgDrk* – night darkness (no lighting at night), *PrLg* – poor lighting (dawn/dusk, artificial lighting of a road at night),
  Dlght – daylight,

- *ShTp*; the shoulders' type with the values:

  S0 – no shoulders, *S1* – any shoulder on one side of a roadway,
  SP – protected (semi-hard; usually strengthened with gravel) shoulders on both sides of a roadway, *SD* – different shoulder types on both sides of a roadway,
  SH – paved (hard; usually of asphalt material) shoulders on both sides of a roadway,
  SG – ground (soft) shoulders on both sides of a roadway.

- *BsStp*; the bus stops indicator with the values:

  N – no bus stops,
  Y – at least one bus stop,

- *Intrsc*; the road intersection area indicator with the values:

  N – road segment between intersections (no intersection and no intersection area),
  Y – intersection area,

- *RdSrf*; the roadway surface conditions with the values:

  SnIc – snow-covered or ice-covered, *Wt* – wet, *Dr* – dry,

- *HrCrv*; the radius of a horizontal curve [m]; the 10000 m radius represents a straight road section,
- *AcsNmb*; the total number of private or public road accesses on both road sides,
- *RdWdt*; the roadway width [m],
- *DwGrd*; the absolute value of a downwards grade [%],
- *VrCrv*; the radius of a vertical curve [m]; the 10000 m radius represents a level road section, i.e. the road section with a rectilinear profile and the constant slope equal to zero,
- *ShWdt*; the sum of shoulders' width,
- *DrBh*; the at-fault driver's behaviour defined by the values:

  Vrt – the variety of behaviours; the category that is an aggregation of very rare recorded behaviours. It contains mainly: driver's tiredness or falling asleep, driver's inattention, and the value "others" taken originally from the police reports,
  DrWrRdSd – driving wrong side of a roadway,

NAdSp – failure to adjust speed to traffic conditions,
NGvWy – not giving right of way,
In(U)Tr – incorrect turning back or turning,
InOvBp – incorrect overtaking or bypassing,
FlCl – following too closely,

- *AcTp*; the accident type with the values describing the collision between at least two vehicles:

  SdImp – side impact,
  HdCr – head-on crash,
  RrCr – rear-end crash,
  Oth – other accident types; the category that is an aggregation of very rare recorded crash types like for example: vehicle rollover, or an accident with a passenger,

- *AcSvr*; the accident severity expressed by the status of a road crash according to the level of a human casualty harm as follows [38]:

  Mnr – minor (slight) accident, without the seriously injured or killed and at least one of the casualties in the road accident was slightly injured, i.e. there was a body harm or a health disorder lasting no more than seven days according to the doctor's diagnosis,
  Srs – serious accident, without the killed and at least one of the casualties was seriously injured, i.e. there was a cripple for life, a serious mental or bodily illness, permanent or long lasting work incapacitation, and the like,
  Ftl – fatal accident, with at least one of the casualties killed on the spot or died within 30 days after the accident.

4. Identification of Multi-Vehicle Crash Patterns

The Ward's method, applied several times for various arrangements of the data examples presentation, has delivered different numbers of final clusters. This enables comparing and checking the repeatability of the results obtained using all the methods of conceptual grouping. The final number of clusters varies between six and ten, therefore the data segmentation has been conducted five times for the K-means algorithm as well as for the SOMs in the incremental learning and in the batch learning versions. Thus the total number of the obtained partitions equals fifteen.

Table 1 presents the average, minimum, and maximum values of the partition quality measures. It can be noticed that the results of all the methods have similar assessment. Though, the *PSF* and *CCC* values for the K-means method are lower, which can indicate that, on average, the separation is slightly worse.

Table 1

Statistics of quality measures of the *CrshRd_MltVhcl* dataset segmentation

The statistic	Within STD	R^2	$\frac{\lvert R_2 \rvert}{(1-R^2)}$	PSF	CCC
The K-means method					
Average	0.312	0.275	0.381	35.636	64.299
Minimum	0.304	0.235	0.307	32.868	56.900
Maximum	0.320	0.315	0.460	39.801	74.020
Incremental SOM					
Average	0.312	0.278	0.387	36.341	79.398
Minimum	0.309	0.246	0.326	32.197	74.908
Maximum	0.316	0.311	0.451	42.145	85.552
Batch SOM					
Average	0.316	0.279	0.389	36.467	80.102
Minimum	0.310	0.236	0.309	32.246	69.891
Maximum	0.325	0.311	0.451	40.044	85.767

The average, minimum, and maximum values of the *Importance* index for all the variables under investigation are summarised in Table 2. The table contains also the number of partitions in which the feature appears (*Importance* >0). Figure 1 presents the graphical illustration of a global average *Importance* (*GAI*), which is the *Importance* value for each feature averaged for all the fifteen partitions. The percentage of the partitions in which the feature is to be found is also illustrated.

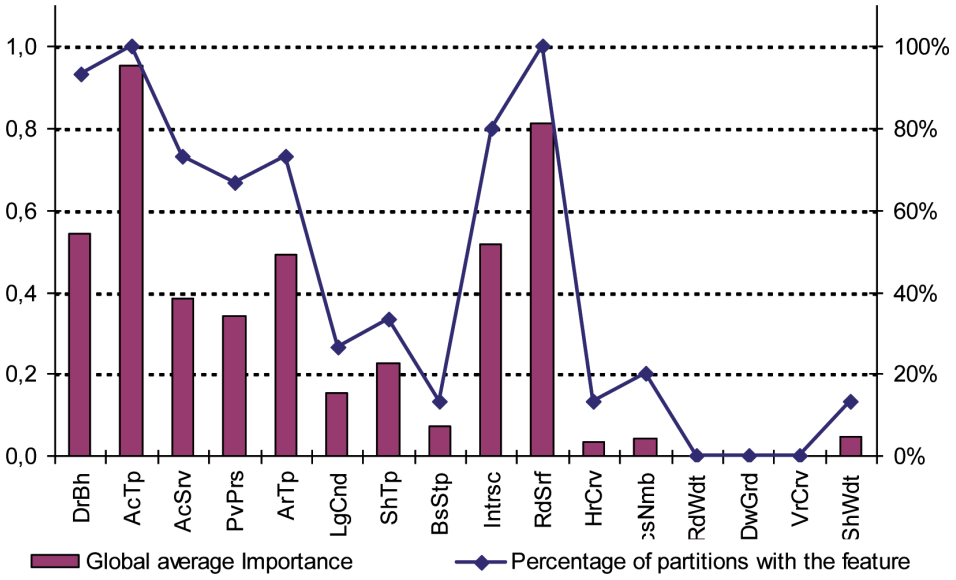

Fig. 1. The evaluation of factors in the process of accident patterns recognition

Table 2

Statistics of the *Importance* index for factors taken to the *CrshRd_MltVhcl* dataset segmentation

Specification	DrBh	AcTp	AcSv	PvPr	ArTp	LgCnd	ShTp	BsStp	Intrsc	RdSrf	HrCrv	AcsNmb	RdWdt	DwGrd	VrCrv	ShWdt
K-means method																
Average	**0.842**	**0.970**	**0.722**	**0.489**	**0.040**	**0.457**	**0.321**	**0**	**0.090**	**0.858**	**0.042**	**0.086**	**0**	**0**	**0**	**0.148**
Minimum	0.584	0.933	0.349	0.214	0	0	0	0	0	0.513	0	0	0	0	0	0
Maximum	1	1	1	0.719	0.199	0.677	0.857	0	0.252	1	0.210	0.224	0	0	0	0.587
Number of partitions with the feature	5	5	5	5	1	4	2	0	2	5	1	2	0	0	0	2
Incremental SOM																
Average	**0.359**	**1**	**0.223**	**0.096**	**0.665**	**0**	**0**	**0.221**	**0.681**	**0.734**	**0**	**0**	**0**	**0**	**0**	**0**
Minimum	0.188	1	0	0	0.566	0	0	0	0.649	0.511	0	0	0	0	0	0
Maximum	0.836	1	0.423	0.479	0.730	0	0	0.591	0.698	0.991	0	0	0	0	0	0
Number of partitions with the feature	5	5	3	1	5	0	0	2	5	5	0	0	0	0	0	0
Batch SOM																
Average	**0.423**	**0.891**	**0.204**	**0.443**	**0.766**	**0**	**0.358**	**0**	**0.783**	**0.853**	**0.060**	**0.047**	**0**	**0**	**0**	**0**
Minimum	0	0.671	0	0	0.575	0	0	0	0.641	0.621	0	0	0	0	0	0
Maximum	0.646	1	0.398	0.605	1	0	0.785	0	0.990	1	0.298	0.236	0	0	0	0
Number of partitions with the feature	4	5	3	4	5	0	3	0	5	5	1	1	0	0	0	0
All methods together																
Global average Importance	0.542	0.954	0.383	0.343	0.490	0.152	0.226	0.074	0.518	0.815	0.034	0.044	0.000	0.000	0.000	0.049
Total number of partitions with the feature	14	15	11	10	11	4	5	2	12	15	2	3	0	0	0	2
Percentage of partitions with the feature	93%	100%	73%	67%	73%	27%	33%	13%	80%	100%	13%	20%	0%	0%	0%	13%

The partitions in all the methods indicate the same hierarchy of the threat features in the accident patterns' identification, both in the frequency of the occurrence aspect and in the variable importance context. Accident type *AcTp* is a primary factor not only in the threat group but also in the whole set of the analysed variables. It appears in all the fifteen partitions and its *Importance* measure has always the highest value (from 0.671 to 1). The second in the hierarchy is at-fault driver's behaviour *DrBh* – found in fourteen partitions with the average *Importance* measure varying between 0.359 for the batch SOM and 0.842 for the K-means method. Accident severity *AcSvr* has less influence on discovering the accident patterns than accident type and at-fault driver's behaviour.

The results indicate four road factors that play the most discriminative role in the dataset segmentation – the features that are present in more than half of the obtained partitions. These are: road surface *RdSrf* (present in all partitions with the global average *Importance* value equal to 0.815), intersection indicator *Intrsc* (twelve occurrences, $GAI = 0.518$), area type *ArTp* (eleven occurrences, $GAI = 0.490$), and pavement presence indicator *PvPrs* (ten occurrences, $GAI = 0.343$). The first, second and third attribute are, on average, more important than accident severity.

Comparatively low and altogether not unequivocal position of accident severity in the ranking of the factors that determine road accident patterns is surprising. In the case of the K-means method the feature stands before all road characteristics except for surface condition. However, in neurocomputing the accident severity yields precedence also to other road factors. It seems the information carried by this threat variable can be taken over not only by accident type and driver's behaviour, but also by some road elements, which determine the dataset segmentation so strongly that accident severity plays a supporting role in the characterisation of patterns.

5. Defining the Profiles

While choosing a partition for the presentation of profiles from fifteen cases, not only quality measures were taken into account, but also the clarity of the segmentation results and the agreement of factors defining the profiles with the most significant factors listed in Table 2. A batch learning Kohonen SOM of the dimension 1×5 has finally been selected. Table 3 presents characteristics of the elements of the map. According to the statistics in Table 1 and 3, in which sample size and the number of clusters are also included, the segmentation is well supported. The graphical illustration of the map clusters is presented in Figure 2.

There is a strong discrimination by area type. Therefore the columns in Table 3 are arranged from the left according to this feature first and then to accident severity. The last column describes the area type combination cluster.

An identifier and a label have been assigned to each element of the map. The profiles are discussed below.

Table 3

Statistics for the batch learning Kohonen SOM road accident dataset partition

Basic statistics		**Whole dataset**	**Element MlVh-1**	**Element MlVh-2**	**Element MlVh-3**	**Element MlVh-4**	**Element MlVh-5**	**Element MlVh-6**
Number of observations		653	111	105	107	112	103	115
Cluster radius			2.76	2.73	2.67	2.82	3	2.8
Distance to the nearest cluster seed			1.29	1.45	1.29	1.34	1.34	1.77
Within cluster variation			0.31	0.32	0.3	0.36	0.33	0.32
Factor specification		**Factor mean value**						
Pavement presence indicator	PvPrs-P0	0.82	0.96	0.92	0.95	0.37	0.93	0.83
	PvPrs-P1	0.11	0.02	0.05	0.04	0.46	0.03	0.07
	PvPrs-P2	0.06	0.02	0.03	0.01	0.18	0.04	0.10
Area type	ArTp-Nbt	0.60	1	1	1	0	0	0.59
	ArTp-Bt	0.40	0	0	0	1	1	0.41
Lighting conditions	LgCnd-NgDrk	0.18	0.22	0.30	0.21	0.08	0.18	0.08
	LgCnd-PrLg	0.13	0.06	0.14	0.13	0.11	0.17	0.16
	LgCnd-Dlght	0.70	0.72	0.56	0.65	0.81	0.64	0.77
Shoulder type	ShTp-S0	0.13	0.04	0.18	0.07	0.24	0.10	0.17
	ShTp-S1	0.07	0.04	0.04	0.07	0.15	0.01	0.10
	ShTp-SP	0.09	0.12	0.06	0.02	0.25	0.07	0.03
	ShTp-SD	0.05	0.04	0.10	0.07	0.06	0.03	0.03
	ShTp-SH	0.20	0.46	0.16	0.06	0.10	0.10	0.32
	ShTp-SG	0.45	0.32	0.47	0.73	0.20	0.70	0.35
Bus stop indicator	BsStp-N	0.83	0.86	0.94	0.89	0.64	0.83	0.81
	BsStp-Y	0.17	0.14	0.06	0.11	0.36	0.17	0.19

continuation Table 3

Intersection indicator	Intrsc-N	0.74	0.81	0.97	0.96	0.71	0.95	0.08
	Intrsc-Y	0.26	0.19	0.03	0.04	0.29	0.05	0.92
Roadway surface conditions	RdSrf-SnIc	0.08	0.04	0.26	0	0.04	0.13	0.01
	RdSrf-Wt	0.27	0.09	0.73	0	0.24	0.38	0.18
	RdSrf-Dr	0.66	0.87	0.01	1	0.72	0.50	0.81
Radius of a horizontal curve	HrCrv	7427.47	8362.03	7438.83	8209.52	7328.01	6773.79	6469.76
Number of road accesses	AcsNmb	3.37	1.95	1.55	2.58	6.88	5.34	1.97
Roadway width	RdWdt	7.28	7.25	7.3	7.16	7.3	7.02	7.63
Downwards grade	DwGrd	1.07	0.97	1.05	0.97	1.45	0.89	1.06
Radius of a vertical curve	VrCrv	9746.4	9970.27	9691.43	9773.83	9564.29	9757.28	9722.61
Shoulders' width	ShWdt	2.96	3.47	2.86	3.19	2.41	3.13	2.74
At-fault driver's behaviour	Vrt	0.09	0.05	0.09	0.24	0.06	0.08	0.04
	DrWrRdSd	0.06	0.02	0.08	0.17	0	0.08	0.01
	NAdSp	0.26	0.19	0.63	0.14	0.25	0.36	0.01
	NGvWy	0.26	0.2	0.04	0.12	0.20	0.12	0.84
	In(U)Tr	0.05	0.04	0.03	0.07	0.06	0.09	0.03
	InOvBp	0.19	0.28	0.13	0.22	0.21	0.20	0.07
	FlCl	0.09	0.23	0.01	0.03	0.22	0.08	0
Accident type	SdImp	0.28	0.18	0.1	0.13	0.17	0.18	0.9
	HdCr	0.34	0.07	0.68	0.64	0.19	0.49	0.05
	RrCr	0.28	0.53	0.15	0.19	0.57	0.23	0
	Oth	0.09	0.22	0.08	0.05	0.07	0.10	0.05
Accident severity	Mnr	0.55	0.85	0.45	0.19	0.88	0.40	0.52
	Srs	0.28	0.12	0.29	0.42	0.11	0.41	0.38
	Ftl	0.16	0.04	0.27	0.39	0.02	0.19	0.1

Element *MlVh-1*: minor accidents in good traffic environmental conditions; $n = 111$

The group represents the pattern with minor accidents (85% observations). Rear-end crashes constitute over a half of the accidents. The most frequent at-fault driver's behaviour is incorrect overtaking or bypassing (28%), then following too closely (23%). These two categories are over-represented in the cluster.

The road conditions are comparatively good: non built-up areas with a small average number of road accesses. Almost 20% accidents were registered on intersections, but it is under-represented in the group. Road design parameters indicate rather straight, both horizontal and vertical, road sections. The sum of shoulders' widths is strongly over-represented (3.5 m), and so is solid shoulders' type (paved – 46%, protected – 12%).

Dry roadway surface in 87% of cases and daylight in 72% of cases complete the picture of good traffic environmental conditions.

The described above roadway and roadside characteristics are conducive to speeding. However, this accident cause is the fourth in succession and is also under-represented in such circumstances. Even if a driver makes a spectacular error (like incorrect overtaking) it will not result in a head-on crash and usually in a death. Road surroundings can be helpful in finding a safer escape, and as a result they can contribute to weakening accident consequences. On the other hand, the severity of rear-end crashes is comparatively low because even at a high speed, the change of velocity at the time of collision generates a smaller kinetic energy than in the case of any other impact [16].

Element *MlVh-2*: mainly severe accidents in adverse traffic environmental conditions; $n = 105$

In the majority of cases (55%) there are serious and fatal accidents in the second element of the map. The most typical behaviour is recognized as excessive speed (failure to adjust speed to traffic conditions on non built-up area), which usually results in a head-on crash. This most dangerous multi-vehicle accident type is over-represented by two times of the frequency in the whole dataset.

Non built-up road sections between junctions and scarce road accesses are specific to this profile. In comparison to other non built-up area accident patterns, the average roadway width is greater but the sum of shoulders' widths and the horizontal curve radius are smaller. The representation for shoulders' type is ground or no shoulders at all.

Daylight is under-represented in most of the clusters – 44% of accidents were recorded at poor lighting conditions (night, dusk, dawn). This is a profile for not a dry roadway surface.

The road width, the lack of cross traffic and rural surroundings can make a driver to be too much self-confident. A closer look into contingency tables revealed that as much as 65% of head-on crashes were caused by speeding. A driver speeds up without assessing (if ever) the threat of the adverse traffic environment conditions sufficiently (possible poor visibility, weak or no shoulders, wet, snowy or icy

surface causing reduced tyre adherence), which become unforgiving road factors accompanying a severe or fatal collision with another vehicle.

Element *MlVh-3*: rural road tragic accidents; $n = 107$

The most tragic accident profile is defined by the third map element. Serious and fatal accidents here are strongly over-represented (81%). The cluster describes most frequently found head-on crashes in non built-up areas. There is no strong dominant driver's behaviour – a variety of behaviours, driving wrong side of a roadway, and incorrect overtaking or bypassing are over-represented.

In all the non built-up area accident patterns, the third cluster has the smallest average roadway width (7.16 m). Comparatively wide ground shoulders on both sides of the roadway are typical of this profile. Almost all the accidents (96%) took place on the stretches between intersections.

Lighting condition has the same distribution as in the whole dataset. However all the accidents were registered as the ones occurring on a dry surface road.

Close to 71% of head-on crashes in the group were caused by the over-represented behaviours, which are quite heterogeneous. Dry road surface combined with almost no transverse traffic as well as with no local vehicular and pedestrian traffic, can make a driver to undertake a risky manoeuvre of overtaking (or by-passing). Driving wrong side of a roadway can be an effect of poor recognition of the route. The variety category includes driver's inattention or distraction, and the value from the police reports defined as "other". All these behaviours lead to loosing control over the vehicle. It seems that insufficient carriageable area as well as maybe insufficient horizontal and vertical road markings are the main road factors contributing to the accident profile of this group.

Element *MlVh-4*: minor accidents in developed area; $n = 112$

This is a group of minor accidents. The over-represented accident type is rear-end crash. Four behaviours constitute the accident cause at the similar level of frequency, i.e. failure to adjust speed to traffic conditions, not giving the right of way, incorrect overtaking or bypassing, and following too closely, though there is a strong over-representation of the last behaviour.

The group represents the accident profile of the area with a comparatively dense development and well-ordered road surroundings: the average number of access points is close to 7, in 36% of cases there is a bus stop (strong over-representation) and in 64% of cases there is a pavement on at least one side of the roadway (again strong over-representation) at the accident site. A fairly large share of transverse road traffic (almost 30% of accidents registered in an intersection area) characterizes the cluster, however it is very slightly over-represented. The discussed element has the smallest percentage of ground shoulders (strong under-representation) of all the six clusters and the greatest percentage of protected shoulders (strong over-representation). Admittedly, the lack of shoulders is the most frequent in the cluster but the accompanying lack of pedestrians' pavements does not exceed 10% of all observations in the group.

The quality of road lighting is good or very good: dark at night is registered in 8% of the accidents. Over 70% observations refer to a dry road surface.

The cross-classification of cause-effect relationships in this cluster indicated the following combinations of behaviours and accident types: (1) following too closely and failure to adjust speed to traffic conditions with read-end crash, (2) incorrect overtaking or bypassing and failure to adjust speed to traffic conditions with head-on crash. In each case the interpretation of excessive speed is different. The former is connected with erroneous distance estimation to a preceding vehicle and a collision itself is connected with the higher speed of a following at-fault driver vehicle than a hit vehicle. The latter deals with driving onto the opposite traffic road lane either by misjudgement of the possibility of making the overtaking/bypassing manoeuvre successfully or by loosing control over the vehicle because of excessive speed. Notwithstanding the case, even if the change in a colliding vehicle velocity is the result of the sum of appropriate velocities at the crash time, it is not as big as to result in a serious or fatal casualty.

The traffic environmental conditions of the profile are like those for the urban one (including both along and through local traffic), which are typified by lower speeds as well as by more driver's attention and thus a lesser severity of traffic accidents.

Element *MlVh-5*: mainly severe accidents in built-up areas; $n = 103$

The group describes severe crashes first (strongly over-represented – 41%). Fatal crashes are slightly over-represented. The most typical accidents here are head-on crashes. The profile defines accidents caused mainly by failure to adjust speed to traffic conditions (36%) and then by incorrect overtaking or bypassing.

The cluster represents the patterns for the area with buildings scarcely spread along the road. It is characterised rather by a road than a street transverse profile. The average road accesses' number does not exceed 5, there are almost no pavements, and a bus stop is rarely present at the crash site. The vast majority of accidents (70%) were registered on a road section with ground shoulders (strong over-representation). Crashes here took place almost exclusively on road sections between junctions, where the roadway width is the smallest of all the clusters.

There is a slight under-representation of daylight; yet it is found in 64% of cases. The accidents occurred in a greater proportion than expected by chance on wet, snowy or icy surface roadways (strong over-representation in each case).

In this pattern, both failure to adjust speed to traffic conditions and incorrect overtaking or bypassing are associated with head-on crash. In that type of collision, where the road surface was not dry, 68% of cases were recognised as caused by excessive speed, whereas where the road surface was dry, 71% of cases were recognised as caused by overtaking or bypassing manoeuvres.

In addition to road surface conditions, it seems that a road transverse profile in connection with insufficient road space, which make it difficult for a driver to dodge the collision, also play an important and unforgiving role in increasing threat on a road.

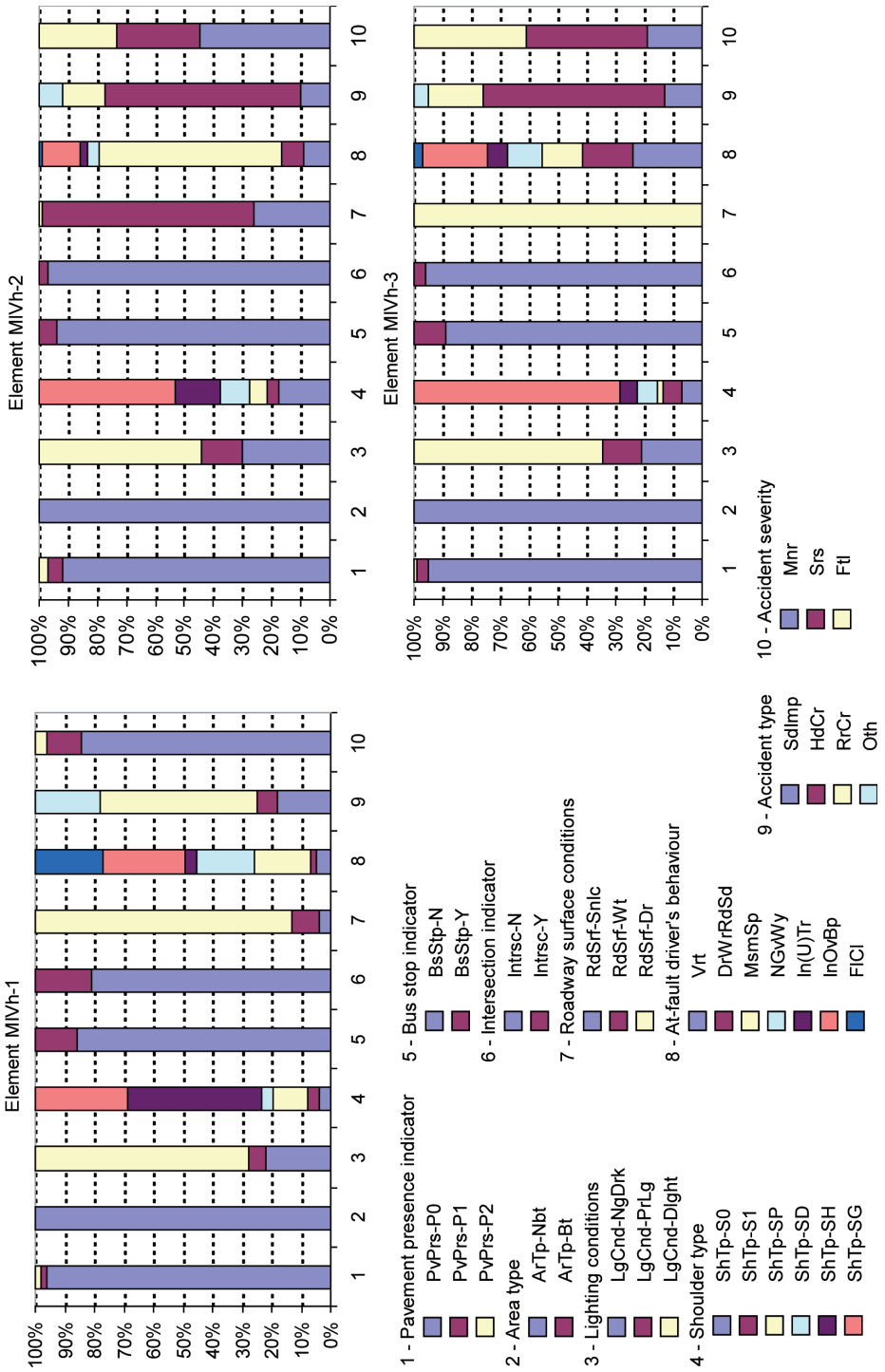

Fig. 2. The differences in the profiles of accident patterns in the batch learning Kohonen map

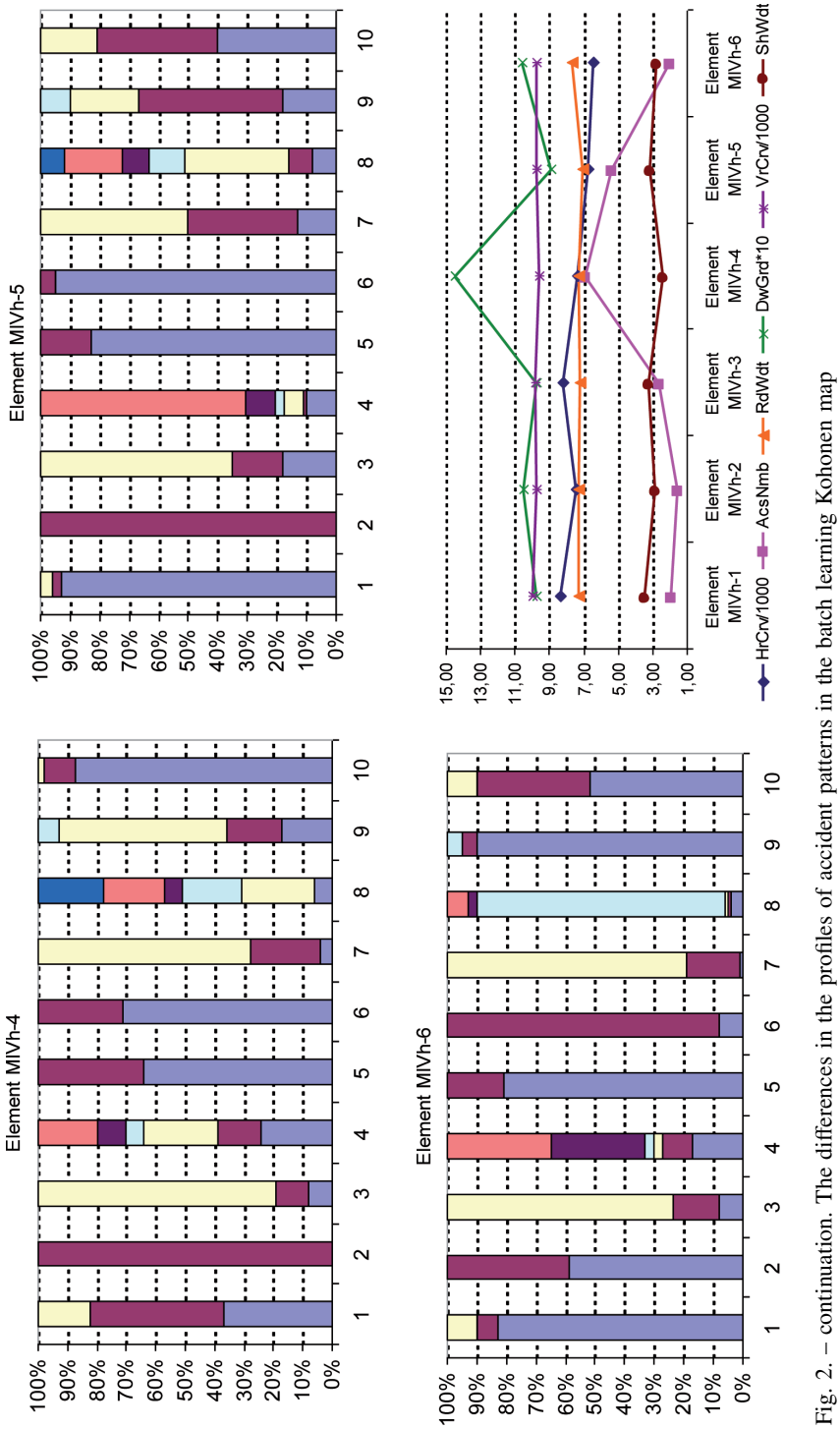

Fig. 2. – continuation. The differences in the profiles of accident patterns in the batch learning Kohonen map

Element *MlVh-6*: intersection area accidents; $n = 115$

The percentages of minor accident status and combined serious and fatal statuses are more or less the same (52% and 48% respectively); though fatal crashes are slightly under-represented. This is a typical pattern of cause-effect relationships, where at-fault driver's behaviour is not giving the right of way (84%) and the result is side impact (90%).

The sixth cluster describes an accident profile almost exclusively in the intersection area (92%) both on rural as well as on built-up roads. In the vast majority of cases (83%) there is no pavement either on one or on both sides of the roadway. The average value of horizontal curve radius is the lowest one of all the map elements, which can more often indicate the occurrence of bends. Paved shoulders are over-represented, however stronger shoulders (paved and protected) and ground ones occur with the same frequency (35%).

The environmental conditions are quite good: 8% accidents (under-representation) took place at night without artificial lighting and less than 20% (under-representation) not on a dry roadway. Contingency tables revealed that the simultaneous lack of lighting and lack of dry surface concerned only 3% of cases.

An insight into the patterns indicates that on non-urban roads with a road transverse profile, regardless of area type, a side impact accident at an intersection can have both a minor and a more severe status. According to traffic environmental characteristics, it can be stated that intersection crashes are the consequence of driver inattention and imprudence while approaching the junction. If, in addition, a higher speed is involved (possibly imposed by a road transverse profile) the serious accident status is reported.

6. Summary

The information concerning a road crash is registered according to the road accident database structure. It includes pure accident characteristics (threat factors in the paper) such as: at-fault person behaviour, crash type, and crash severity. There are also complementary items: road user features (like gender, age, intoxication), colliding vehicle attributes, and "obvious" accident site surrounding factors (like area type, bend or grade presence indicator, weather information). Therefore, the combination of the above variables is usually investigated in almost all disaggregated road traffic safety studies. In such a heterogeneous structure, factors that characterise a road user involved in the accident are indicated as primary significant causes of variety safety problems. It is not surprising because the human factor has been identified as the most significant element in the road traffic threat generation of a human-vehicle-road system. However, roadway and roadside elements can create a surplus value in the creation of accident circumstances. That is why the aim of this study was to focus on the road related potential determinants of the accident risk, particularly the ones connected with geometry and road environment.

Multi-vehicle crash data for a chosen region of Poland were categorised according to a variety of qualitative and quantitative factors. Before the research task was completed, the data concerning undivided national road characteristics at the site of the accident occurrence was collected first and then combined with the threat factors. In order to identify road accident patterns, both the crash location features and the threat factors were subject to conceptual grouping. The K-means algorithm as well as the incremental- and batch learning Kohonen neural networks were utilised in the process of the dataset segmentation, which has resulted in the recognition of the accident patterns labelled as follows: (1) minor accidents in good traffic environmental conditions, (2) mainly severe accidents in adverse traffic environmental conditions, (3) rural road tragic accidents, (4) minor accidents in developed area, (5) mainly severe accidents in built-up areas, (6) intersection area accidents.

The conclusions can be formulated in a twofold aspect: the methodology and the obtained results.

From the methodological point of view, it can be stated that decision about the final number of clusters is not straightforward, as it requires a series of trials and diagnosing the clarity of resulting partitions. All the algorithms used for conceptual grouping indicated similar factors participating in the road accident patterns recognition, though there were small differences in the factors' importance. The K-means method delivered slightly worse separation quality than Kohonen maps.

One may judge by the results of the analysis that the following road related features play an important role in the road accident profiling tasks: area type and area development level, roadway surface condition, intersection indicator, shoulders type, and also, to some extent, lighting conditions, roadway width, shoulders' widths, and horizontal curve radius.

To sum up, it is worth pointing out that crash typing can be a valuable tool in diagnosing circumstances of the hazard on non-urban roads and therefore in reducing the road accident deaths and serious injuries. The outcome is specific to a certain category of roads and a certain country region. If this specificity is similar for a region, the results may be helpful for a local road administration in the road safety management and in the remedial measures implementation.

Acknowledgement

The author is very much grateful to anonymous referees for their valuable suggestions that helped to improve the paper quality.

References

1. Anderson I.B., Bauer K.M., Harwood W.H., Fitzpatric K.: Relationship to Safety of Geometric Design Consistency Measures for Rural Two-Lane Highways. In Transportation Research Record 1784, Paper No. 02-2302, TRB, Washington, D.C, 2002, 108-114.

2. Baltes M. R.: Descriptive Analysis of Crashes Involving Pedestrians in Florida, 1990–1994. In Transportation Research Record: Journal of the Transportation Research Board, No. 1636, Washington D.C., 1998, 138-145.
3. Berg H-Y., Gregersen N. P., Laflamme L.: Typical patterns in road-traffic accidents during driver training. An explorative Swedish national study. Accident Analysis and Prevention, No. 36, 2004, 603-608.
4. Cichosz P.: Systemy uczące się. Wydawnictwa Naukowo-Techniczne, Warszawa, 2000. (in Polish).
5. Das A., Abdel-Aty M., Pande A., Santos J. B.: Severity analysis of crashes on multilane arterials using conditional inference forests. CD-ROM – the TRB 88th Annual Meeting of the Transportation Research Board, Washington D.C., 2009.
6. Delen D., Sharda R., Bessonov M.: Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention, 38, 2006, 434-444.
7. Dixon K. K., Liebler M., Hunter M.: Urban roadside safety – cluster crash evaluation. CD-ROM – the TRB 88th Annual Meeting of the Transportation Research Board, Washington D.C., 2009.
8. Garber N., Ehrhart A.: Effect of Speed, Flow, and Geometric Characteristics on Crash Frequency for Two-Lane Highways. In Transportation Research Record: Journal of the Transportation Research Board, No 1717, TRB, National Research Council, Washington, D.C., 2000, 76-83.
9. Guidici P.: Applied Data Mining. Statistical Methods for Business and Industry. John Wiley & Sons Ltd., Chichester, 2003.
10. Hand D., Mannila H., Smyth P.: Eksploracja danych. Wydawnictwa Naukowo-Techniczne, Warszawa, 2005, (in Polish).
11. Hanowski R. J., Medina A. L., Wierwille W. W., Lee S. E.: Incident Clustering Diagnostic Approach for Assessing Usability of Intersections and Other Road Sites. In Transportation Research Record: Journal of the Transportation Research Board, No. 1897, Washington D.C., 2004, 173-179.
12. Kim K., Yamashita E.: Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawai. Journal of Advanced Transportation 41, (1), 2007, 69-89.
13. Kim K., Pant P., Yamashita E. Y.: Hit and Run Crashes: Using Rough Set Analysis with Logistic Regression to Capture Critical Attributes and Determinants. CD-ROM – the TRB 87th Annual Meeting of the Transportation Research Board, Washington D.C., 2008.
14. Korbicz J., Obuchowicz A., Uciński D. Sztuczne Sieci Neuronowe. Podstawy i Zastosowania. Akademicka Oficyna Wydawniczaj PLJ, Warszawa, 1994, (in Polish).
15. Krzyśko M., Wołyński W., Górecki T., Skorzybut M.: Systemy uczące się, rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości. Wydawnictwa Naukowo-Techniczne, Warszawa, 2008, (in Polish).
16. Lachowski J., Nowakowska M.: Wpływ prędkości na skutki zderzeń pojazdów z obiektami. Drogownictwo nr 6, Czerwiec, 2005, 163-165, (in Polish).
17. Larose D. T.: Odkrywanie wiedzy z danych. Wydawnictwo Naukowe PWN, Warszawa, 2006, (in Polish).
18. Major H., Nowakowska M.: Charakterystyka zagrożeń brd na odcinkach zamiejskich dróg niższych klas technicznych. Konferencja naukowo-techniczna „Wpływ środków organizacji na bezpieczeństwo ruchu drogowego", Kielce, 12-13 maja 2005, (in Polish).
19. Marakas G.M.: Modern data warehousing, mining, and visualization. Prentice Hall, New Jersey, 2003.
20. Marek T.: Analiza skupień w badaniach empirycznych. PWN, Warszawa, 1985, (in Polish).
21. Milton J., Mannering F.: The relationship among highway geometric, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, Kluwer Academic Publishers, 1998, 395-413.

22. Milton J. C., Shankar V. N., Mannering F. L.: Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident Analysis and Prevention, No. 40, 2008, 260-266.
23. Nowakowska M., Major H.: Jakość danych w analizach bezpieczeństwa ruchu drogowego. Międzynarodowe Seminarium GAMBIT 2004, Wydawca: Fundacja Rozwoju Inżynierii Lądowej, Gdańsk, 13-14 maja 2004, 327-333.
24. Nowakowska M.: Poprawność i spójność wewnętrzna danych o zdarzeniach drogowych. VI Międzynarodowe Seminarium Bezpieczeństwa Ruchu Drogowego GAMBIT 2006. Miejsce programu GAMBIT w III Planie BRD Unii Europejskiej. Wydawca: Fundacja Rozwoju Inżynierii Lądowej, Gdańsk, 17-19 maja 2006, 109-120, (in Polish).
25. Nowakowska M.: Finding threat patterns in the interaction between road transportation and pedestrian traffic using market basket analysis. Monografie Zespołu Systemów Eksploatacji PROBLEMS OF MAINTENANCE OF SUSTAINABLE TECHNOLOGICAL SYSTEMS, Polskie Naukowo-Techniczne Towarzystwo Eksploatacyjne, Tom II, Warszawa, 2010, 140-162.
26. Osowski S.: Sieci neuronowe do przetwarzanie informacji. Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2000, (in Polish).
27. Pande A., Abdel-Aty M.: Market basket analysis: A novel way to find patterns in crash data from large jurisdiction. CD-ROM – the TRB 86th Annual Meeting of the Transportation Research Board, Washington D.C., 2007.
28. Pande A., Abdel-Aty M.: Discovering indirect associations in crash data using probe attributes. CD-ROM – the TRB 87th Annual Meeting of the Transportation Research Board, Washington D.C., 2008.
29. SAS Institute Inc.: SAS®Technical Report A-108, Cubic Clustering Criterion. Cary, NC: SAS Institute Inc., 1983, 56 pp.
30. SAS/STAT User's Guide. Version 8. SAS Publishing , Cary N.C., 1999.
31. SAS OnLine documentation 9.1. SAS Institute Inc., Cary, NC, USA, 2003.
32. Stamatiadis N., Jones S., Aultman-Hall L.: Causal Factors for Accidents on Southeastern Low-Volume Rural Roads. In Transportation Research Record: Journal of the Transportation Research Board, No. 1652, Washington D.C., 1999, 111-117.
33. Stutts J. C., Hunter W. W., Pein W.E.: Pedestrian Crash Types: 1990s Update. In Transportation Research Record: Journal of the Transportation Research Board, No. 1538, Washington D.C., 1996, 68-74.
34. Tabachnick B.G., Fidell L.S.: Using Multivariate Statistics. HarperCollinsCollegePublishers, New York, 1996.
35. Wang X., Abdel-aty M.: Analysis of left-turn crash injury severity by conflicting patterns using partial proportional odds models. Accident Analysis and Prevention, No. 40, 2008, 1674-1682.
36. Wong J-T., Chung Y-S.: Analyzing heterogeneous accident data from the perspective of accident occurrence. Accident Analysis and Prevention, No. 40, 2008, 357-367.
37. Wong S. C., Leung B. S. Y., Loo B. P. Y., Hung W. T., Lo H. K.: A qualitative assessment methodology for road safety policy strategies. Accident Analysis and Prevention, No. 36, 2004, 281-293.
38. Zarządzenie nr 635 Komendanta Głównego Policji z dnia 30 czerwca 2006 r. w sprawie metod i form prowadzenia przez Policję statystyki zdarzeń drogowych. Warszawa, 2006, (in Polish).
39. http://bus.utk.edu/stat/stat579/Hierarchical%20Clustering%20Methods.pdf; Schmidhammer J. L.: Agglomerative Hierarchical Clustering Methods. University of Tennessee, Department of Statistics, USA, Accessed July 7, 2010.
40. http://lord.uz.zgora.pl:7777/skep/docs/F29571/Gramacki_Ploug08.pdf; Gramacki J., Gramacki A.: Wybrane metody redukcji wymiarowości danych oraz ich wizualizacji. Uniwersytet Zielonogórski, Instytut Informatyki i Elektroniki, Poland, Accessed September 15, 2010.

41. http://www.palgrave-journals.com/jibs/journal/v37/n4/fig_tab/8400206t2.html; from the article: Lim L. K. S., Acito F., Rusetski A.: Development of archetypes of international marketing strategy. Journal of International Business Studies, July 1, 2006, Accessed July 7, 2010.
42. http://www.nargund.com/gsu/mgs8040/resource/dm/ClusterPaper.doc; Nargundkar S., Olzer T.J.: An Application of Cluster Analysis in the Financial Services Industry. May & Speh, "Strategic Decision Services", Atlanta, GA, USA, Accessed July 7, 2010.
43. http://www.statsoft.pl/czytelnia/marketing/przykladyzaawans.html; Sagan A.: Przykłady zaawansowanych technik analitycznych w badaniach marketingowych. Akademia Ekonomiczna w Krakowie, Kraków, Poland, Accessed September 15, 2010.
44. ftp://ftp.sas.com/pub/neural/FAQ.html#A_Kohonen; How many kinds of Kohonen network exists? And what is k-means? SAS FAQ pages, USA, Accessed July 7, 2010.
45. http://www.cis.hut.fi/somtoolbox/theory/somalgorithm.shtml; Kohonen T.: The Self-Organizing Map (SOM). Finland, Accessed July 7, 2010.