# Analysis of discrete data from traffic accidents

Pavla Pecherková and Ivan Nagy

*Abstract*— **This paper deals with the data analysis of traffic accidents. Traffic accidents can be caused by different reasons, e.g., by watchfulness of a driver, failure of a vehicle, bad structural arrangements, etc. The aim of this paper is to investigate seriousness of incidents in dependence on different circumstances of an accident. Description of these circumstances leads to the use of a high number of different variables (about 50 variables), which are mostly discrete. The majority of statistical methods dealing with discrete variables use a frequency table. This is not suitable for traffic data because of a huge dimension. In this paper, several methods are proposed for solution to the problem with high-dimensional traffic data.**

*Index Terms*— **data analysis, characteristic of the data sample, multivariate analysis, traffic accident, traffic model, traffic variable**

## I. Introduction

THE increasing traffic brings about unfortunately also growing number of traffic accidents [1, 2, 3]. This is a problem that does not suit smart cities and should be solved.

Some causes of the accidents are evident and they are concurrently removed. In spite of it, accident frequency remains and still constitutes a serious problem both in cities as well as outside them. The causes of these accidents, especially in cities, are not so clear. We can divide them into two groups:

1) Accidental incidents caused mostly by insufficient watchfulness of the driver, some watchfulness of the driver or, and it used to be rather seldom nowadays, by failure of the vehicle.
2) Some systematic cause that follows from bad structural arrangements of communication or evocation of a bad traffic situation (high speed etc.).

The first group is entirely under influence of randomness and it can be influenced only by public education. The second one comprises elements of determinism. It depends on some other measured variables or rather by their specific combinations. These relations are mostly very complex and for a simple reasoning totally hidden. In these cases it is necessary to use some exact statistical method and to analyse a data sample measured on traffic area to be investigated. Modern computers are sufficiently powerful to be able to work not only with values of variables but also, for smaller systems, with all their combinations. Nevertheless, the high dimension generated by some statistical algorithms is still a problem which causes that their practical solution is very hard problem. In the literature it is known as “curse of dimensionality”.

The traffic situation in cities, which are of main interest in this paper, is rather specific:

- The majority of accidents are only light ones caused by bad concentration of drivers.
- The traffic networks in cities are closely monitored so dangerous arrangements are usually soon corrected.
- From the reasons of improved possibility the structural arrangements of communications or regulations are often changed. Some of such changes can, as a side effect, lead to a decrease of safety.

The basic question before the analysis itself is if the measured data contain an information about the investigating phenomenon or they do not. In other words, if the data at disposal do not belong to the first group of data, mentioned.

## II. Choice of variables

This paper would like to lead a discussion on a general level - what are the problems with analysis of accidental data and which statistical methods can be used. However, to be more specific, we will use some typical data sample for demonstration. Nevertheless, the conclusions are general enough [4, 5].

The data sample used is a collection of 3894 data measurements measured at the beginning of the year 2012 in Prague, Czech Republic. Each data measurement is a vector of 59 items, storing various characteristics accompanying the specific accident.

Our aim is to investigate seriousness of incidents with the goal to suggest how to avoid them and reduce especially the hard ones.

The main characteristics of the data sample are

- The variables measured are practically all discrete.
- The number of variables is too high.
- The number of combinations of the variable values that

This work was supported in part by the project GAČR GA15–03564S.

Pavla Pecherková is with Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 110 00 Prague, Czech Republic and Department of Signal Processing, Institute of Information Theory and Automation of the ASCR, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic (e-mail: pecherkova@fd.cvut.cz).

Ivan Nagy is with Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 110 00 Prague, Czech Republic and Department. Of Signal Processing, Institute of Information Theory and Automation of the ASCR, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic (e-mail:nagy@fd.cvut.cz).

**978-1-5386-3825-5/17/$31.00 ©2017 European Union**

represent the system states is horribly too high.

Majority of statistical methods dealing with discrete variables use a frequency table whose dimension (number of entries) is given as a product of numbers of values of individual variables. Thus, e.g. for 50 variables with 10 values in average produce table of dimension $10^{50}$. If we take one millisecond for inspection of one table entry we would need $3.2 \times 10^{39}$ years for inspection of the whole table. The time of existence of the universe is estimated to be $2 \times 10^{12}$ years.

## III. Data sample and its model

From the above discussion it follows that this situation cannot be solved as it is and it is necessary first to reduce the huge dimension. The first natural step is to reduce the number of variables and possibly also the number of values of individual variables. The variables used are selected with respect to the main modelled variable (model output) which in our case is the seriousness of an accident. So, we must ask, which of the variables at disposal can best explain the output behavior; to explain why one accident is light and another serious.

### *A. The Modelled Variable*

For the choice of accident seriousness (AS) we use a standard procedure. Its continuous version (expressed as economic costs) is given as

$$AS = \frac{(433a + 4867.7b + 19440c + d)}{1000} \tag{1}$$

where

a is number of minor injuries,
b is a number of serious injuries,
c is number of deaths,
d is material damage.

The variable AS is usually discretized using five intervals

$$0 - .1 - .433 - 4.87 - 19.44 - \infty$$

obtaining five degrees of accident seriousness (DAS):

1) material damage up to 100 000 CZK (3 701 EUR), without injury,
2) material damage between 100 000 and 483 000 CZK (17 876 EUR), without injury,
3) minor injuries,
4) serious injuries,
5) deaths.

However, for our data sample the mentioned intervals lead to the following distribution of DAS

| DAS | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Number | 3546 | 0 | 345 | 3 | 0 |

It is clear, that for our data sample such discretization is utterly unsuitable. The values 2 and 5 do not occur at the sample at all and the value 4 is very poorly excited. The reason has already been mentioned above - in cities the accidents are mostly only light. So, it was necessary to produce a new division which would reveal the differences in light accidents. We used the following intervals

$$0 - .001 - 1 - \infty$$

which produced the following distribution

| DAS | 1 | 2 | 3 |
|---|---|---|---|
| Number | 2959 | 896 | 39 |
| probability | 0.76 | 0.23 | 0.01 |

Even here the last value 3 is represented very poorly.

### *B. The Explanatory Variables*

Now, the second step is to find variables which explain these values representing a grade of seriousness of accidents as well as possible. However, their number should be as small as possible and the same holds for their coding - assigning numbers to their individual states. Notice, that these two requirements are contradictory.

After a discussion with experts, we choose twelve variables with the originally designed values (states). However, already the first count of the dimension of the frequency table showed, that even such project is too big. Namely the dimension would be $2.2 \times 10^{10}$ which means, that if the inspection of one cell of this table takes one millisecond, we would need 42 years for the whole inspection. That is why, we had to reduce still more. So, the final setup made for our investigation of traffic accidents from measured data is as follows:

- The modelled variable is the code of accident seriousness with three values: 1 = very light accident, 2 = light accident, 3 = the rest of accidents (as discussed above).
- The explanatory variables are
  - 1 = alcohol (with 3 values)
  - 2 = conditions on the road surface (with 3 values)
  - 3 = state of communication (with 3 values)
  - 4 = weather (with 3 values)
  - 5 = visibility (with 2 values)
  - 6 = sight conditions (with 3 values)

### *C. Suitability of data*

Now the task is to verify if the explanatory variables really explain the explained variable (the degree of incident seriousness). In an ideal case, there should be a bound between explained and explanatory variables as strong as possible and, at the same time, the explanatory variables should be independent. This task should be accomplished by using some independence test. The mostly used ones are chi-square and Pearson (Spearman) tests. However, for our analysis dealing with discrete variables, even here problems are encountered.

Pearson test is a test of correlation coefficient. This test aims at continuous variables with the assumption of their normal distribution. However, its direct variant is Spearman test, which tests also correlation coefficient, however, for ranks of the entering data. The consequence is that the normality of data is not required and it is directly proclaimed to be suitable for discrete data. The result of testing is that this test is entirely unsuitable for our purposes. The reason is there is still a relatively wide gap between the border of independence/dependence (from the viewpoint of correlation) and the border, where we can speak about a bound between variables proper for modelling of dependency. This test rejects independence practically for all variables even for those that are evidently not suitable for explanation of the modelled variable (degree of seriousness of accidents). Thus, these tests must be excluded from our methods.

Chi-square test is based on a slightly different principle. Here, two-dimensional empirical probability function is compared to the equivalent but independent one. The core of independence is that the population probability function (table) is formed only from one normalized histogram. This histogram is closely related to prediction of the modelled variable from those from the condition (which are explanatory variables). This notion is directly connected with modelling which is of our interest.

Gamma coefficient [9] follows form of a newly discovered method which is practically not used and which seems to be absolutely ideal for our purposes of modelling with discrete data. It depends on comparison of prediction for marginal and conditional distribution. Let us have two random variables X and Y with conditional $f(x|y)$ and marginal $f(x)$ distributions. Let $x$ and $y$ be data sets sampled from them. Then we can construct the frequency table $T$. The rows of this table conditional distributions (without normalization) $f(x|y)$ and the vector representing the sum of $T$ over columns corresponds (up to normalization) to the marginal distribution $f(x)$. From these distributions and the given data sample we are able to determine number of prediction errors when predicting from marginal distribution (prediction of $x$ only, without knowledge of $y$) and those using the full conditional distribution (prediction with the knowledge of $y$). A ratio of those numbers produces the gamma coefficient. This coefficient testifies not about independence but about association of variables which is perfectly what we need.

For our data sample the association of all explanatory variables has been acknowledged.

### D. *Discrete model*

The discrete model of measured variables y (explained variable) and x (explanatory variables) is fully determined by the empirical joint probability function

$$f(y,x) \tag{2}$$

which is in nothing more that normalized frequency table specifying frequencies of occurrence for all possible combinations of values of the variables involved

## IV. Analysis of variable

After constructing the model we can come to the analysis itself. Recollect that our goal is to investigate the degree of accident seriousness with respect to selected measured variables, specifically, which combinations of values of these variables lead to serious accidents. The primary method is based on a full discrete model in a classified form and it leads to comparison of histograms. However, as already mentioned, this method suffers from the so called curse of dimensionality. Some ways, leading from this problems, are also mentioned. They are logistic regression and Bayes networks.

### A. *Equations*

For our method we transfer model (1) into the conditional form, which is

$$f(x|y) \tag{3}$$

for $y=1,2,\ldots,n_y$, where $n_y$ is the number of values of $y$. This operation consists just in normalizing the "rows" of the table to the sum equal to one.

Remark

*The multidimensional table $f(y,x)$ can always be ordered into two-dimensional one by coding the combinations of values of the variables in $x$ according to this simple example (for $x_1 \in \{1,2\}$ and $x_2 \in \{1,2,3\} \rightarrow x$).*

| | | | | | | |
|---|---|---|---|---|---|---|
| $x_1$ | 1 | 2 | 1 | 2 | 1 | 2 |
| $x_2$ | 1 | 2 | 3 | 1 | 2 | 3 |
| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |

The conditioning divides the table into $n_y$ vectors whose graphical expressions are histograms. Each histogram corresponds to one value of $y$, i.e. to one state of the explained variable. Now, the histograms can be compared. If selected histograms do not differ, then the corresponding two states of the explained variable are excited by the same combinations of values of the explaining variables. This case is not interesting from the viewpoint of the analysis. On the other hand, if the histograms differ in some entries then the corresponding combinations cause as a result either one or the other value of the explaining variable. Plainly speaking: if the two states of $y$ are light accident and serious accident, we can specify which combinations of values of explanatory variables lead do serious accidents.

The method is simple and very effective. The only, and we must admit that substantial, problem is that we work with the whole frequency table which, as already discussed, is of a huge dimension. So, the immediate interest is how to reduce this table to some acceptable size.

There exist two very promising methods, which however, have to be accommodated to our specific problem. They are as mentioned comparison of histograms providing some measure for similarity/dissimilarity of histograms and Bays nets that on the basis of conditional independence reduce the problem dimension.

### B. Results with the example data

For the model with 6 explanatory variables and their reduced number of values the histograms have 486 columns. In the following picture we are showing the first 20 of them. However, we can notice, that e.g. the fourth column is of interest because in the third histogram it differs from the previous two. The first two histograms both represent relatively light accidents while the third one is built of data from serious ones. So, the fourth column says that the mode denoted by 4 accompanies rather serious accidents. Now, we have to de-code the coded explanatory variable to the original ones. And they are 1 1 1 1 2 1, which means: all variables up to the last but one (which is visibility) are in a good state, only visibility is bad. So the result here is: very good driving condition and bad visibility leads rather to serious accidents.

### C. Comparison of histograms

The immediate help in making the analysis problem feasible is to use results from the statistical area of comparison of histograms.

There are two prominent papers concerning the problem. The first one provides an overview over the methods for comparison of histograms [6] and which is a follow-up to the first paper and where a practical method of comparison based on so called normalized significance of the difference [7]. The method provides automatic comparison of two histograms with the resulting proclamation of their identity or discrepancy.

### D. Bayes nets

Bayes nets use graphical representation of cause flows induced by conditional probability functions derived from a joint probability function using chain rule and conditional independence [8]. A very brief indication of their function is as follows: Let us have three random variables $X$, Y and Z. Their joint probability function is

$$f(X,Y,Z) = f(X|Y,Z)f(Y|Z)f(Z)$$

However, if the flow of information flows as follows Z\to Y\to X then X and Z are independent on condition that Y is known. For the joint probability function then holds

$$f(X,Y,Z) = f(X|Y)f(Y|Z)f(Z)$$

where the factorization has smaller dimension than the original distribution. Extend of reduction is significant for more variables, indeed.

The result is: Through a careful expert formulation of the problem to be solved, the Bayes nets can radically reduce an infeasible problem.

## V. Conclusion

The paper summarizes several methods which can be used for analysis of discrete data samples. These data arise from dealing with traffic accidents. Here, it is necessary to take into account relatively large amount of variables which leads to a high-dimensional problem. A novel method leading to comparison of histograms generating by conditioning of the independent variables by the dependent one is suggested. However, the proposed method will need further investigation with respect to dimensionality reduction – here Bayes networks look promising and automatic comparison of histograms – which can be bases on the theory of histograms association.

## References

[1] P.J. Ossenbruggen, J. Pendharkar, et al, “Roadway safety in rural and small urbanized areas,” *Accidents Analysis and Prevention*, vol. 33, no. 4, pp. 485-498, 2001

[2] Mussone, Lorenzo, Andrea Ferrari, and Marcello Oneta. "An analysis of urban collisions using an artificial intelligence model." *Accident Analysis & Prevention*, vol. 31, no. 6, pp. 705-718, 1999

[3] Sohn, S. and S. Hyungwon, "Pattern recognition for a road traffic accident severity in Korea." *Ergonomics*, vol. 44, no. 1, pp. 101-117., 2001

[4] Sohn, S. and S. Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea," *Safety Science*, vol. 41, no. 1, pp. 1-14, 2002

[5] K. Ng, W. Hung, W. Wong, "An algorithm for assessing the risk of traffic accidents," *Journal of Safety Research*, vol. 33, pp. 387-410, 2002

[6] Porter, Frank. "Testing Consistency of Two Histograms.", arXiv:0804.0380, 2007

[7] Bityukov S. I., Krasnikov N.V., Taperechkina V.A. and Smirnova V.V., „Statistically dual distributions in statistical inference“, in proceedings of *Statistical problems in Particle Physics, Astrophysics and Cosmology* (PhyStat’05), September 12-15, 2005, Oxford, UK, Imperial College Press, 2006, pp. 102-105

[8] Moore, Andrew W., and Denis Zuev. "Internet traffic classification using bayesian analysis techniques." *ACM SIGMETRICS Performance Evaluation Review*. Vol. 33., No. 1. ACM, 2005.

[9] Wilson, T. P. 1969. “A proportional-reduction-in-error interpretation for Kendall's tau-b”. Social Forces 47: 340–42.