Transactions on Transport Sciences | Peer-Reviewed Open Access Journal
DOI: 10.5507/tots.2020.008

ToTS Volume 11, Issue 2: pg77–pg83
Palacky University in Olomouc

Visualizing crash data patterns

Peter Wagner, Ragna Hoffmann, Marek Junghans, Andreas Leich, Hagen Saul
Institute of Transportation Systems, DLR, Rutherfordstrasse 2, 12489 Berlin, Germany

ABSTRACT: *This paper demonstrates an approach that makes it easy to find patterns in traffic crash databases, and to specify their statistical significance. The detected patterns might help to prevent traffic crashes from happening, since they may be used to tailor campaigns to the community at hand. Unfortunately, the approach described here comes at a cost: it identifies a considerable amount of patterns, not all of them are being useful. The second disadvantage is that is needs a certain size of the data-base: here it has been applied to a data-base of the city of Berlin that contains about 1.6 Million (M) crashes from the years 2001 to 2016, of which about 0.9M had been used in the analysis.*

KEYWORDS: *crash data patterns; crash analysis; contingency tables; Pearson residual; Cramers V; mosaic plot;*

1. INTRODUCTION

Analyses of crash data-bases may search for patterns that can be exploited for prevention of future crashes. There are well-known approaches for this that range from simple tables (e.g. reports as the ones published by statistical authorities) and contingency tables (see Tunaru (1999) or Kateřina et al (2019)), to sophisticated models for crash likelihood (Mannering (2018)) that try to summarize these data into models whose parameters are estimated from the data.

Especially the official reports of crash data by the various statistical authorities are often large contingency tables, where it is difficult to see at one glance what might be important, and what not. Here, a means to visualize these data-bases is investigated. It draws on methods from data science analysis such as in James et al (2013), and has therefore a preference toward data-bases with large number of crashes. The contingency tables used here will not work well for data-bases with a small number of crashes.

2. METHODS AND DATA USED

2.1 Methods

The method to look for patterns to be used here is a mixture of several well-known approaches. It draws from what is used in data science analysis by computing a kind of correlation between all those variables that may seem interesting or helpful to the analyst. Of course, this selection of variables is to a certain degree subjective, or, to put it more neutral, problem-specific. The method to compute this correlation draws on contingency tables and their related Pearson residuals (see Agresti (2007) or Kateřina et al (2019)). To compare different correlations they need to be normalized which is done by using Cramér's V (Cramer (1946)). After ordering the correlations according to their V-value, the most interesting ones are easily picked out. They can then analyzed into greater detail by looking at the matrix of the Pearson residuals and display them in a mosaic plot. Mosaic plots are described into greater detail by way of an example in Figure 3.

2.2 Data

The crash data-base of the city of Berlin from the years 2001 – 2016 has been used here. To avoid dealing with small numbers and therefore yield statistical meaningful results, the data have been filtered as follows. The raw data contain for each crash as many records as there are people involved, where the first record for each crash belong to the participant that has been identified as the main culprit of the crash. This has been aggregated in one record per crash, where each record contained (which is a subset of all the variables in the original data-set):

- Year (year), hour (hour), day of the week (weekDay).
- Number of fatalities (nFatal), number of heavily injured (nHeavy), and the number of lightly injured people (nLight).

- Traffic type of first participant (v1Type), of the second participant (v2Type), the type of crash (crashType), and the collision diagram (colDia). The crash types are defined as follows: 1: crash while driving alone, 2: crash while turning, 3: crash when crossing, 4: crash caused by crossing pedestrian, 5: crash with vehicles parking, 6: crash in longitudinal traffic (both head-on as well as rear-end collisions), 7: anything else.
- Age (age) and sex (sex) of the first involved, and whether there was alcohol intoxication reported (BAC). Note, that BAC is a binary variable only.
- Temperature (temp), humidity (humidity), and average daily traffic (ADT) of the year 2009 (adt2009) close to the place of the crash.

The data-base contained the place of the crash as well, but this was only used to assign the three external variables temperature, humidity, and adt2009. The ADT-values are from Berlin's official traffic model, and they are there from several years (2005, 2009, and 2014). Since the different years are highly correlated, we have chosen to use just the middle year (2009).

In addition to this, the data have been filtered as follows:

- Only crashes with two participants have been picked (about 92% of the data).
- Only the top 12 collision diagrams (about 66% of the data) have been used, see below and Figure 1.
- Only the top seven traffic types have been used (passenger car (Car): 76.98%; lorries (Truck): 7.25%; bicycles (Bike): 3.63%; miscellaneous vehicles (Misc): 2.65%; pedestrians (Peds): 1.35%; regular bus (PT-Bus): 1.14%; motorbike (MoBike): 1.11%)

This reduces the original data-base of 1,569,621 to 913,556 crashes that fulfil the criteria above. The final step involved the discretization of the (almost) continuous variables like temperature, humidity, adt2009, and age, which are typically aggregated into 10 categories of roughly equal size, with the exception of age. In age, a break at 18 years has been manually added to have a classification of participants younger and older than 18, the year where people in Germany can legally hold a driver's license.

In the following, Figure 1 to Figure 5 contain visualizations that display a few impressions of the data. First of all, in Figure 1 there is the distribution of the shares of the twelve most frequent collision diagrams. As already mentioned, these twelve collision diagrams cover almost 2/3 of all the crashes, so it is useful to concentrate on these.

Since the data-base is so large, at least a small test of the stability of the result displayed in Figure 1 can be performed. This has been done by sampling many times randomly from the data-base. For each sample, the share of each collision diagram can be computed, and then, the average share and other statistics can be computed. It turns out, that the shares for these top twelve collision diagrams are fairly stable, i.e. in all samples the ranking is basically unchanged and even the shares do not fluctuate a lot. So, Figure 2 demonstrates that the result displayed in Figure 1 is fairly robust.

It is also interesting to look whether some of the distributions change over the years, and in fact, they do. To analyze them, we make use of mosaic plots with an assumed model of independence (see Zeileis et al. (2007)).

This method is explained with the example of Figure 3 that displays the change of the daily crash pattern over the 16 years of data. For each hour of the day *i,* and for each year *j*, there is a particular number of crashes reported, that makes the contingency table n_{ij}. A mosaic plot displays each box so, that its size is proportional to the n_{ij}. In addition, each box is colored according to its deviation (the Pearson residual r_{ij}) to an expected count e_{ij} and the related standard deviation $\sigma_{ij}^{(e)}$ of the expected count:

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sigma_{ij}^{(e)}} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{1}$$

If the expected counts are drawn from a Poisson distribution, then the standard deviation is just the square-root of the expected counts e_{ij} and the second part of the equation is valid as well.

The simplest possible model, and the one that will be used in all cases in this work, is to assume that the counts are independent of the combination of the time of the day and the year. In this case, the e_{ij} can be computed directly from the measured counts n_{ij} as follows:

$$e_{ij} = \frac{N_{i,\cdot}\,N_{\cdot,j}}{N} \tag{2}$$

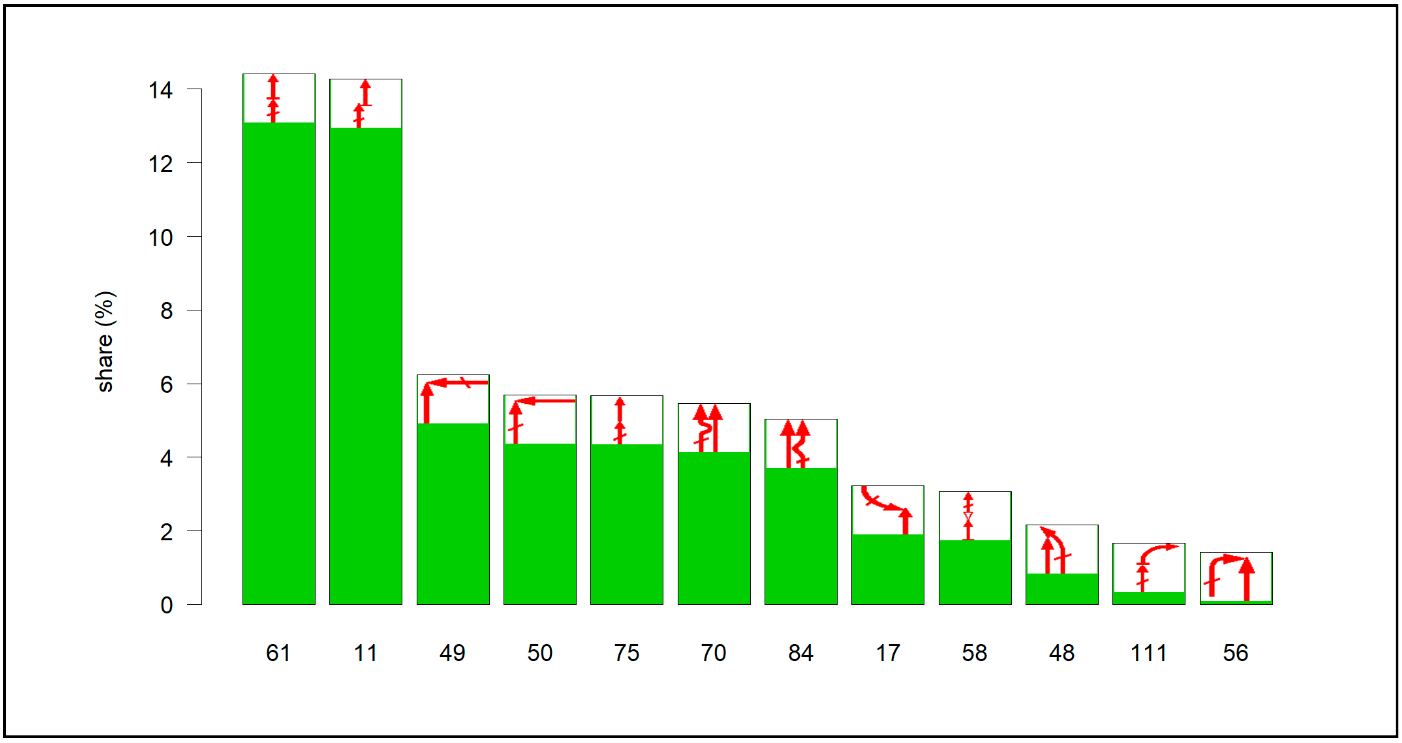

Figure 1: Share of the crashes of the twelve most frequent collision diagrams, together with a graphical representation of them and their number on the x-axis. The slanted bar in one of the arrows indicates the vehicle that caused the crash, a horizontal bar at the end of the arrow (in 61, 11, 58, and 111) a standing vehicle, the wiggles in 70 and 84 a side-swipe crash, and the unfilled arrowhead (in 58) a backward driving vehicle.

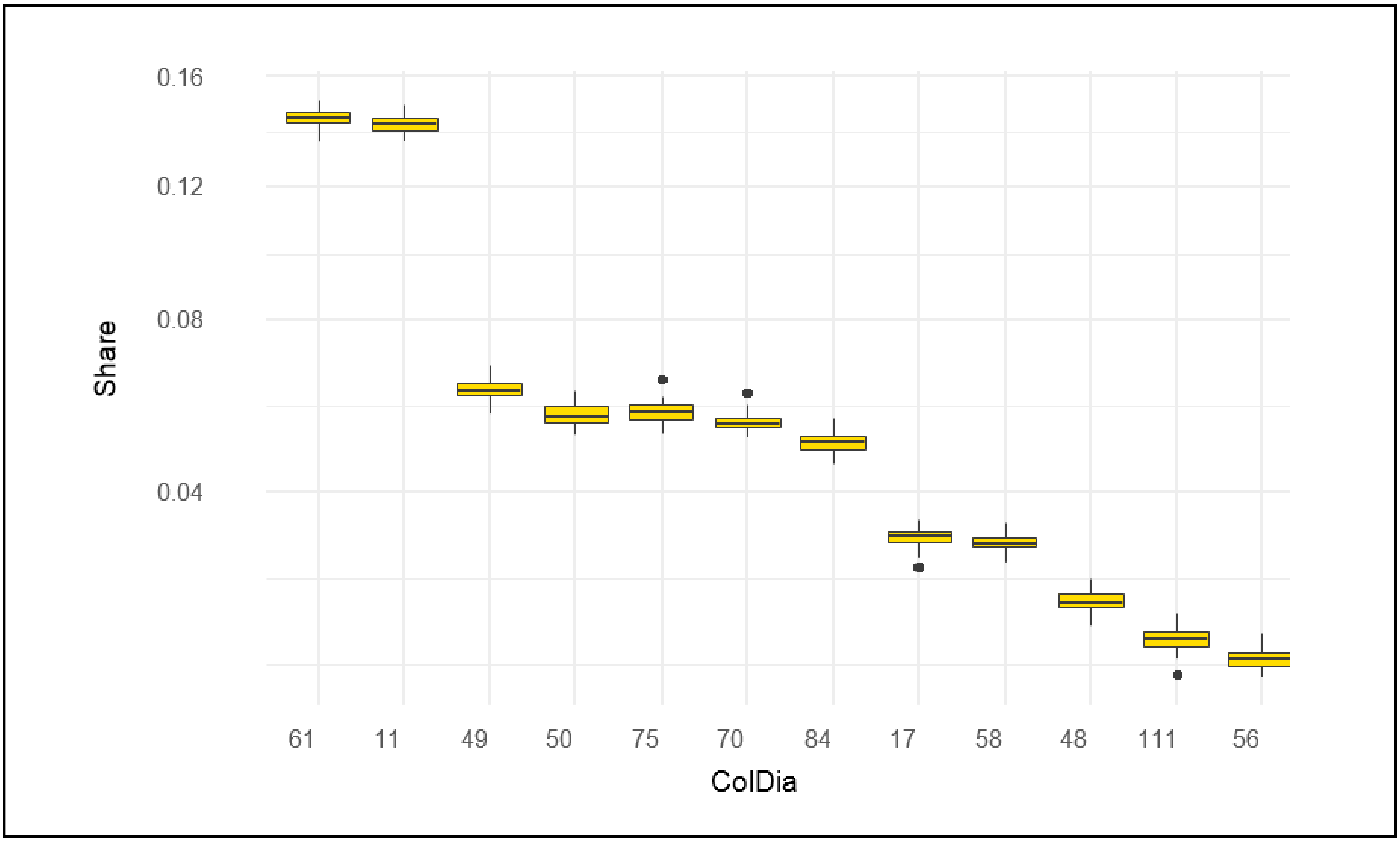

Figure 2: By sampling the data-base 50 times (and picking 10,000 crashes in each sampling), a distribution of the share of each collision diagram can be obtained. Note, that the y-axis has been scaled by a square-root to make small entries better visible.

Here, $N_{i,}\cdot$, $N\cdot_{,j}$ are the row and column sums of the matrix, and $N = \Sigma_i N_i$ is the number of observations in the database. A mosaic-plot now displays in addition to the size of the boxes also the Pearson residuals r_{ij} by coloring the boxes: a grey box indicates that it follows the expected counts; a red box indicates that the observed number of crashes in that category is lower than expected, while blue boxes indicate that it is higher than expected.

For the two variables time of day and year, there are a lot of grey boxes visible in Figure 3, indicating that these two features are in fact independent of each other. This is not true for all of them, but it seems that the daily pattern of crash occurrences does not change much over the 16 years covered by the data-base.

However, by looking at other pairs of variables, changes do occur. In Figure 4, the collision diagrams

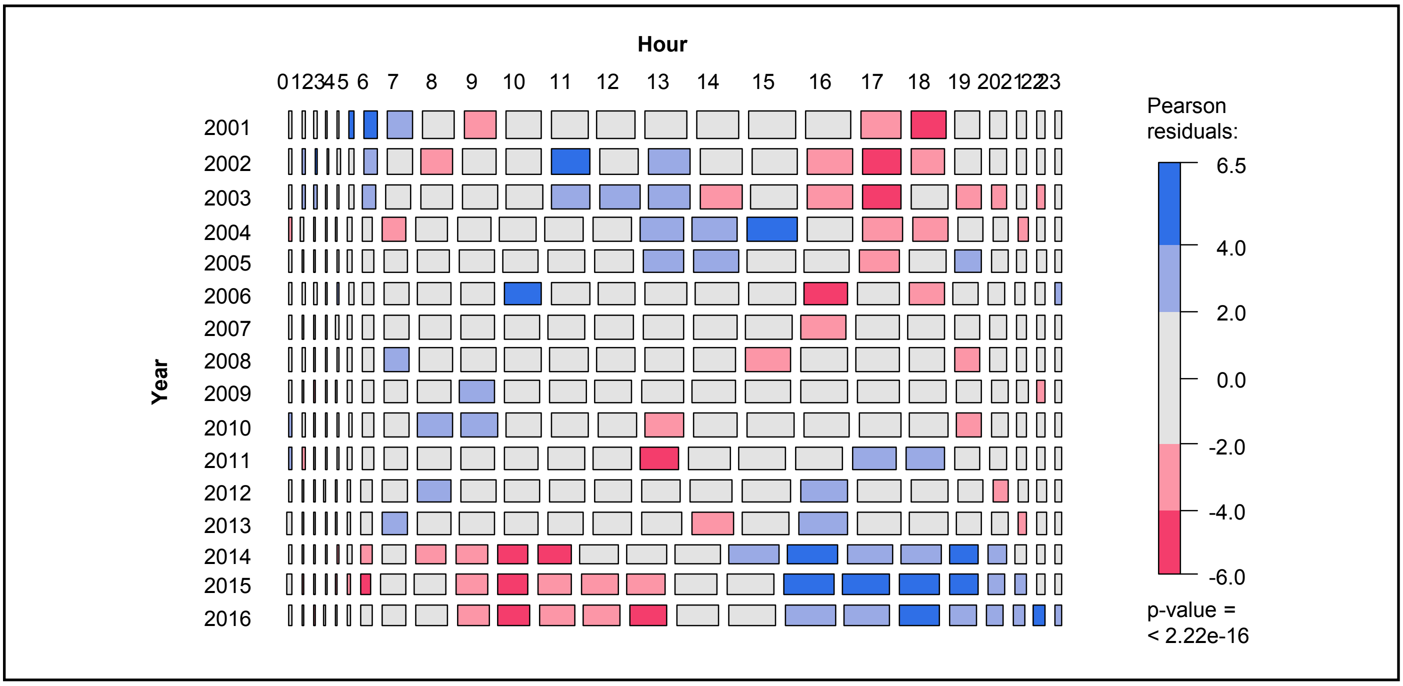

Figure 3: Mosaic plot of the number of crashes as function of the time of the day and year. The daily pattern can be seen easily, and it seems, that it does not change much over the years.

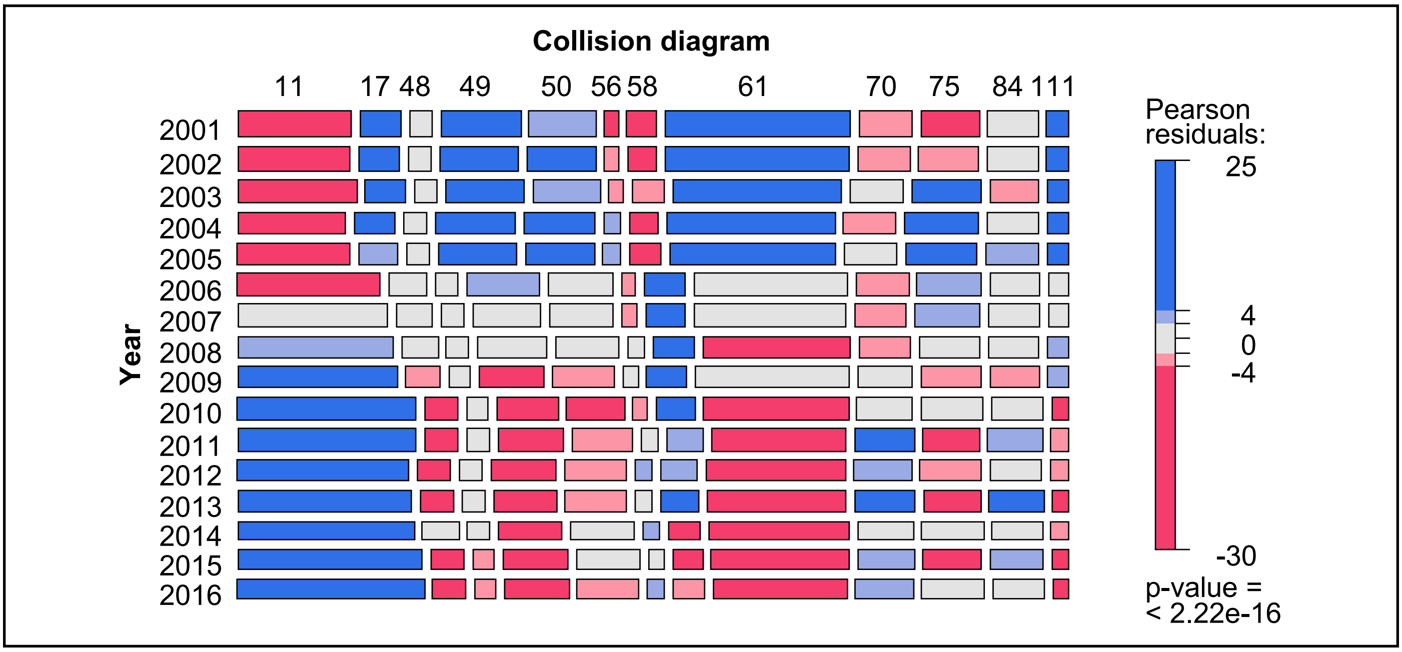

Figure 4: Mosaic plot of the number of crashes as function of the collision diagram and year.

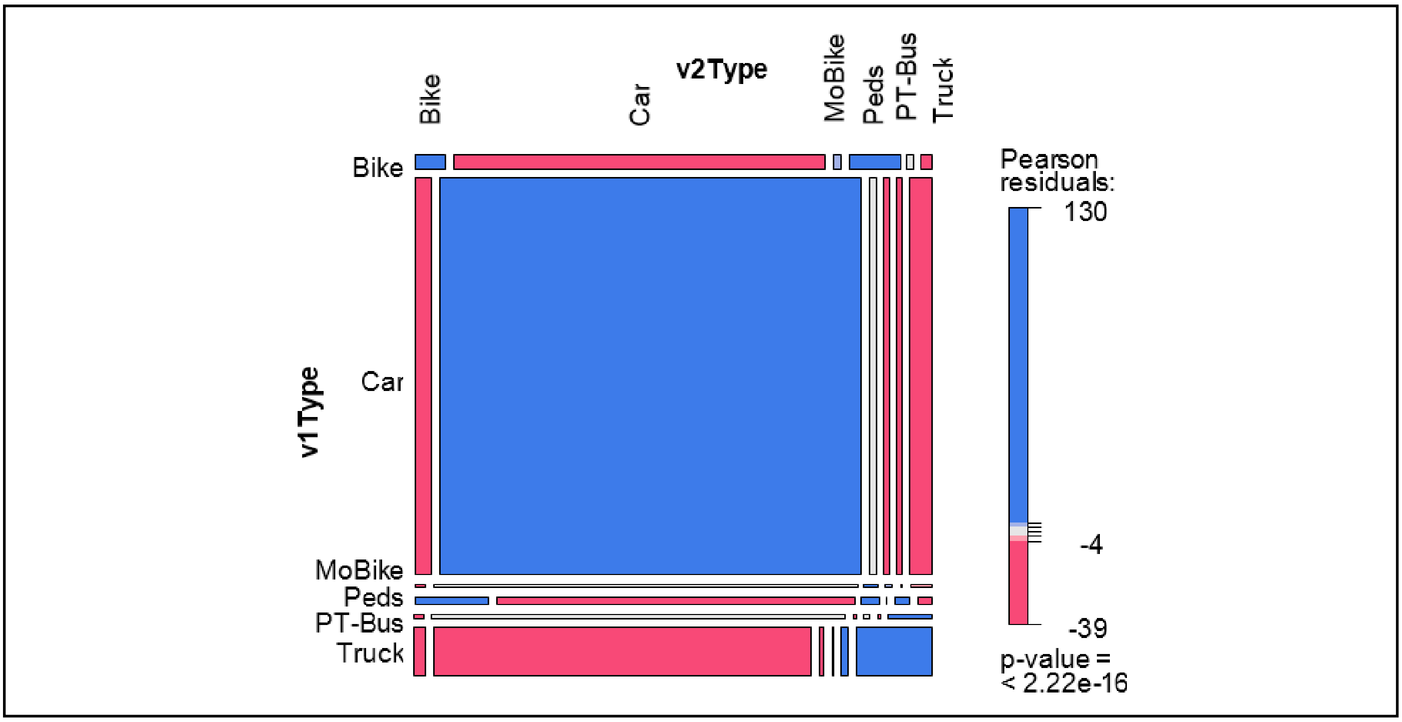

Figure 5: Mosaic plot of traffic types of the first and the second participant of the crashes.

are followed over the years, and here, considerable changes can be seen. Especially the shares of the two most frequent collision diagrams change over the years: diagram #11 increases in share while #61 decreases in share. There is currently no explanation available why this happens.

As a final example, the matrix of traffic participants is displayed in Figure 5. The majority of the crashes happen between two cars, in addition to this, a car-car crash is also more likely than expected from the naïve assumption of independence. Note, that a crash between a bike and a car where the bike driver is responsible also seem to happen less frequently than expected by the assumption of independence.

3. RESULTS

So far, we have mostly analyzed a few picked variables and how they change over the years. In the following, a more systematic path will be followed. To do so, a mosaic plot was computed for all the possible combinations of the k = 16 variables above. This yields k(k-1)/2 = 120 matrices of Pearson residuals. To compare these with each other with the goal of finding a ranking between important and non-important correlations, an index is needed. There are a few that can be used, here it has been decided to use Cramér's (Cramer (1946)). It is defined as follows:

$$(3)\ V = \sqrt{\frac{\chi^2/N}{\min\{c-1, r-1\}}} \quad \text{where } \chi^2 = \sum_{ij}\left(\frac{n_{ij}-e_{ij}}{\sigma_{ij}^{(e)}}\right)^2$$

where c is the number of columns and r is the number of rows of the contingency table. This corrects for the different shapes of the matrices and makes V a number in [0,1]: the larger V, the stronger are the two variables correlated, where correlation is defined in a very general sense.

The results are displayed in Table 1 for the first 20 strongest correlations.

Some results in Table 1 are to be expected. For instance, the crash type is derived from the collision diagram, it is a kind of aggregation, and therefore, the two should have a strong correlation, i.e. a large value of V. The next two are not too surprising as well, most likely it stems from the prevalence of cars among the crashes.

In Figure 6, just three mosaic plots of the entries 4, 6, and 11 of Table 1 are displayed.

A few interesting results can be picked from these plots:

- Crash participants who are younger than 18, and between 24 and 29, have a considerably larger chance to get involved in a bike crash. Also, the below 18 pedestrians are at a larger risk than expected from the assumption of independence.

- With cars, the youngest and the oldest drivers have an increased risk, as well as almost all truck drivers
- Alcohol abuse when driving has a strong temporal pattern, with a peak of the share near 3 o'clock in the morning.
- The crash-type also has an interesting pattern as a function of age. In addition to this, it is different for the two sexes included in the data-base (not shown here).

Again, the robustness of this approach has been tested by chopping the data-base randomly into ten sub-parts, and repeat for each of these subparts the procedure above (computation of the matrix and of Cramér's V). This, then, leads for each of the sub-parts to an individual ranking. These different ranking can be compared by assigning the average rank with the global rank obtained from the application of this method to the whole data-base. It shows, that the ranks are fairly stable, with a standard-deviation of each rank typically smaller than 3.

Table 1: Results of the correlation analysis

Rank	Column	Row	Cramer's V
1	crashType	colDia	0.754
2	crashType	v1Type	0.400
3	colDia	v1Type	0.256
4	age	v1Type	0.228
5	nLight	v2Type	0.208
6	hour	BAC	0.207
7	sex	v1Type	0.207
8	colDia	v2Type	0.206
9	crashType	v2Type	0.173
10	temp	Humidity	0.169
11	crashType	age	0.161
12	crashType	nLight	0.148
13	crashType	adt2009	0.145
14	nLight	v1Type	0.135
15	nHeavy	v1Type	0.128
16	crashType	nHeavy	0.123
17	nLight	colDia	0.111
18	colDia	adt2009	0.111
19	BAC	v1Type	0.098
20	v1Type	v2Type	0.097

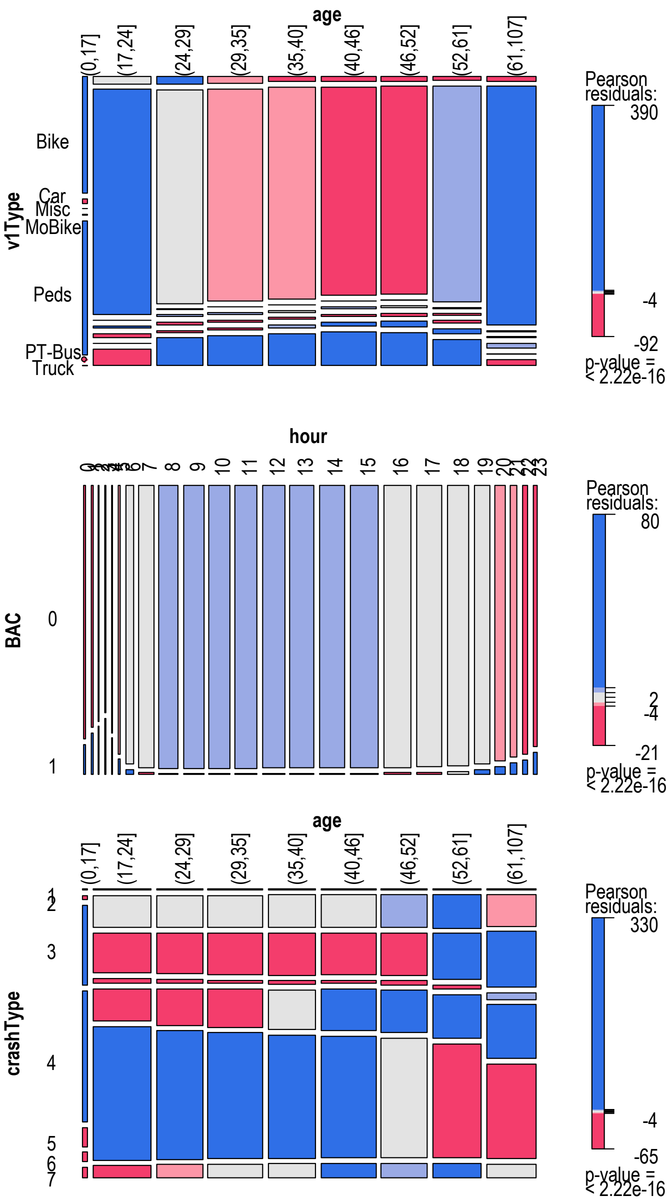

Figure 6: Mosaic plots of entries 4, 6, and 11 in Table 1.

4. CONCLUSIONS

Crash data contain highly significant patterns. Tools like mosaic plots are useful to visualize and enable us to find those patterns. In fact, it seems that they find too many patterns (Figure 6 shows only 3 out of the 120 possible patterns, and of course one may question the original choice of parameters), so the question for future work is how to work with these results, and how to best obtain information from them that can be used in the practitioners' daily work.

Another avenue of future research is to include other data into this analysis, so that the risk, e.g. in form of vehicle miles travelled can be taken into account. One step in this direction would be to have as a model something that is proportional to the exposure of the different groups. Demand data-bases such as the German MiD (MiD 2019) can help in this task, but our initial approach was not successful because the data-base for the city of Berlin had not enough trips recorded that could be used e.g. to compare the crash pattern with the demand pattern (as function of day and hour, e.g.).

Also, higher dimensional generalizations of this approach do exist and may be interesting to explore. These do, however, become difficult to work with because of the curse of dimensionality, which will rapidly lead to too small numbers in the boxes.

Finally, when we look at some of the mosaic plots, it might not straightforward to describe even the relationship between the two variables that make such a plot by a linear model. One particular difficult example is in Figure 6 the relationship between age and crash type, where a very complicated pattern could be seen. If the analyses put forward here are of any interest, they may demonstrate that more complicated models than just linear ones might be needed to model crash probabilities.

Acknowledgements: We would like to gratefully acknowledge the Police of the City of Berlin for making these data available to us. In addition, it is a real pleasure to thank our former colleague Leonhard Lücken for performing the analysis that added the weather data and the demand data to the original data-set.

REFERENCES

Agresti, A. (2007). An Introduction to Categorical Data Analysis, 2nd ed. New York: John Wiley & Sons.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

James G., Witten, D.,Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning, Springer Texts in Statistics.

Kateřina, B., Eva, M., Robert, Z., Pavlína, M., Martina, K., & Roman, M. (2019). Factors contributing on mobile phone use while driving: In-depth accident analysis. *Transactions on Transport Sciences*, *10*(1), 41-49. doi: 10.5507/tots.2019.008.

Mannering, Fred. (2018). Cross sectional modeling, in "Safe Mobility – Challenges, Methodology, and Solutions", edited by Dominique Lord and Simon Washington, pp 257–277. Emerald Publishing Limited.

Mobility in Germany (in German) (2019). Retrieved from http://www.mobilitaet-in-deutschland.de/index.html (last accessed 30 Dec 2019)

R Core Team (2019). R: A Language and Environment for Statistical Computing

Tunaru, R. (1999). Statistical modelling of road accident data via graphical models and hierarchical Bayesian models. PhD thesis, Middlesex University

Zeileis, A., Meyer, D., and Hornik K. (2007). Residual-based Shadings for Visualizing (Conditional) Independence. Journal of Computational and Graphical Statistics, 16(3), 507-525.