Journal Title
XX(X):1–9
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A smart data approach to traffic safety

A. Leich[1], R. Nippold[1], P. Wagner*[1, 2]

Abstract

This work demonstrates, how a large data-base of traffic crashes can be used to analyze ensemble data. It fused data from the German Unfallatlas (German Crash Database - GCDB) with Open Streetmap data (both publicly available), and a database from the German Federal State Northrhine-Westfalia (NW) named NWSIB that provides additional information about each intersection, most importantly an estimate of the ADT-values at each intersection. The results have to be taken with care, since the quality of the ADT's in the data-base is hard to control, and because this approach may have assignment errors. The results partially reproduce known findings; however, they allow in principle for a more detailed investigation of the relationship between crash-numbers and ADT-values than is possible with generalized linear models (glm). In line with the call, all the data, as well as the scripts that analyse the data are publicly available – this text is entirely written in the Rmd format[1], and most computations have been done in R[2] and QGIS[3].

Keywords

Traffic safety; intersections; crash-rate;

Note: the latest text-version to this is on overleaf. The paper there contains small amendments to this Rmd, but the analysis and the Figures have not changed.

Introduction

Many means to organize intersections between roads are available, each of them having different operational characteristics when it comes to questions like safety, efficiency,

[1] Institute of Transport Systems, German Aerospace Center, Rutherfordstrasse 2, 12489 Berlin, Germany
[2] Institute of Land- and Sea Transport Systems, Technical University of Berlin, Salzufer 17-19, 10587 Berlin, Germany

Corresponding author:
Peter Wagner
Email: peter.wagner@dlr.de

or required space. Here, the concentration is on safety, and out of the many different intersection controls, this paper picks the following four (this selection is mainly based on availability):

- Without any organization, which in Germany comes down to right-before-left (will be named X in the following)
- With priorization, where one road has priority over the other (named Prio)
- Organized as a roundabout (named RA)
- Controlled by a traffic signal (named TS).

Common wisdom notes that with regard to traffic safety, a roundabout is the safest of these four – which is one of the reasons for their usefulness, they mark a good compromise between safety and efficiency as long as the view is on cars only[4–7].

The data used

This research uses three main data-sources, all of them publicly available.

- The German Accident Database (Unfallatlas – GCDB)[8],
- The Northrhine Westfalia Road Traffic Database (NWSIB)[9],
- The OpenStreetmap (OSM) database[10]

The GCDB

For each crash in Germany with injured people (the record starts in 2016, the latest year for which data are available is 2021), this database has the detailed location of the crash (in latitude/ longitude), some information about the time of the crash, the crash type, and the modes involved in it. The mode list contains bike, car, pedestrian, motorbike, truck, and an unidentified rest.

Note, that the crash data are optimized for privacy: some information has been left out by purpose, e.g. the detailed time of the crash, as well as the actual number of injured or killed people in each crash. The traffic mode is limited to 0 or 1, i.e. even a crash with two cars is noted only with a 1, not with a 2. For this investigation these omissions do not seem critical, this may change for other investigations. Altogether 46,953 crashes could be assigned to the study area that result from the NW road data-base. The crash-data are in the following denoted by the letter N.

Openstreetmap (OSM)

From the OSM database, all the intersections in the German federal state Northrhine-Westfalia (NW for short) have been extracted, resulting in about 281,729 intersections. They have been classified mostly from OSM itself, with additional input from another project where the roundabouts have been identified by separate means, see[11].

NWSIB

Not all of the OSM intersections had entries in the NWSIB data-base, and not all intersections in NWSIB had usable ADT values (denoted in the following by $Q_{\text{car}}, Q_{\text{bike}}$) or simply Q, so it comes down to a total of 59,635 intersections that had a more or less complete set of data. This holds for the car counts, the number of intersections that had ADT values for the bikes are a little bit smaller (59,044) and we think that the bike-data are less reliable.

Analysing the data

The different data-bases have been matched by creating a unique identifier from the OSM, and then assigning the information in the NWSIB and the crash database to the intersections from OSM. The road format of NWSIB is not compatible with OSM, so the assignment was done by matching all crashes and all ADT-values within a radius of 75m around the OSM intersections. Of the ADT-values, the maximum was taken, since it was not clear how the ADT-values have been assigned to the intersections in the NWSIB database. Taking the max avoids double-counting from vehicles that enter and leave an intersection.

It has been observed, that some intersections had fairly large ADT-values, up-to 170,000 cars/day. While not impossible, it is very unlikely to have intersections with such a large demand in the data-set. Therefore, it was decided to eliminate the 1% of the data with the largest ADT-values with the exception of the RA, where it is thought that the Q-values are clean. In addition, the analysis below further restricts the modelling to ADT-values smaller than 65,000 veh/day since the data become very sparse for large ADT.

For the analysis of the data, two approaches are being used: the first one is the standard approach in traffic safety research which works with generalized linear models (glm), and the second one is a data-driven approach.

The second approach, which is especially useful when the data are plentiful, clusters the data in certain classes of Q and computes separate statistics for each class, like the mean value. This in essence allows for a much more general relationship between the number of crashes and the exposition Q, which is not restricted to the type of models compatible with a glm.

Approach based on generalized linear models (glm)

When modeling safety with a glm, it is assumed that the number of crashes can be described by a model of the type[12–15]:

$$\mu = \beta_0 Q^{\beta_1} \exp\left(\sum_{i\geq 2} \beta_i x_i\right) \tag{1}$$

Here, μ is the mean-value of the number of crashes N (which is strictly speaking a rate, since it is crashes/time-interval, where one year is often used for the time-interval), the

Q is the exposition where often ADT is used, the number of cars per day, and the x_i are various factors assumed to influence the crash-rate, such as the intersection organization, the position of the intersection (urban/rural) etc. The β_i describe how strong each factor influences the crash-rate, and they are estimated from the available data. Note, that with respect to the exposition, a very specialized function is used, i.e. a power-law (Q^β or $Q_1^{\beta_1} Q_2^{\beta_2}$ for two crossing streams with demand Q_1, Q_2). Also, the use of the exp()-function is debatable, especially when such a model is used to extrapolate and forecast. However, its use is difficult to avoid, it is a consequence of the fact that crash numbers are positive and the linear model is not for μ itself, but for the logarithm of the mean-value:

$$\log \mu = \beta_0 + \beta_1 \log Q + \sum_{i \geq 2} \beta_i x_i \tag{2}$$

To complete this description, these mean values are the mean-values of a Poisson (P) or Negative Binomial (NB) distribution, in the following a NB has been used, which is also an observed fact in many, but not in all investigations. The difference between P and NB can be seen when looking at the relationship between mean μ and variance σ^2, which for the Poisson distribution is $\sigma^2 = \mu$, while for the NB it has over-dispersion:

$$\sigma^2 = \mu + \mu^2/\theta$$

where the parameter θ describes the deviation from the Poisson distribution: for $\theta \to \infty$, a Poisson distribution is retained. Here, the fit results yield $\theta = 1.13 \pm 0.02$, indicating a rather strong deviation from P.

In addition, the NB approach yields the smaller AIC (139,000 versus 148,000), so it justifiable to assume that the data are NB-distributed.

The actual R command to fit these data is given in the following line:

```
m <- glm.nb(N ˜ isCtrl + isCtrl:I(log(Q)), data=mm)
```

This results in the following set of parameters:

Table 1. The result of the model specified above to fit the crash-numbers versus ADT, for the four intersection controls.

	Estimate	Std. Error	z value	Pr(¿—z—)
(Intercept)	-8.291	0.133	-62.366	0
isCtrlRA	2.671	0.462	5.788	0
isCtrlTS	2.426	0.233	10.427	0
isCtrlX	1.052	0.220	4.792	0
isCtrlPrio:I(log(qCarMax))	0.824	0.015	55.264	0
isCtrlRA:I(log(qCarMax))	0.582	0.049	11.884	0
isCtrlTS:I(log(qCarMax))	0.660	0.020	33.008	0
isCtrlX:I(log(qCarMax))	0.676	0.020	34.230	0

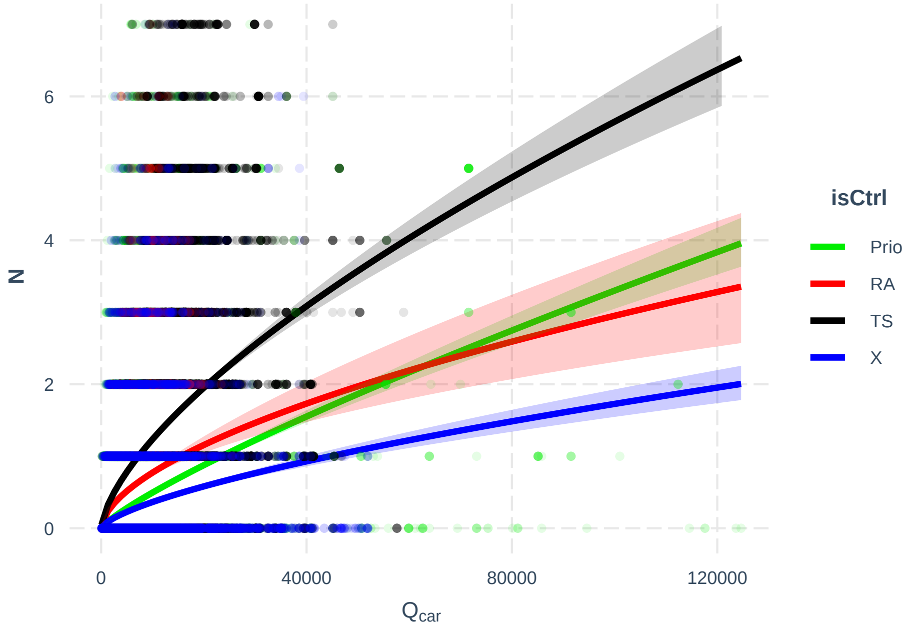

Figure 1. The result of the glm fit. Shown are some of the data, and the four curves that result from the fit for the four intersection organizations.

A graphical display of this is shown in Figure 1, where the crash numbers N are displayed as a function of the demand Q_{car} for the four intersection controls. All nine fitting parameters (two for each of the four curves $\beta_{0i}Q^{\beta_{1i}}$, and one for θ) are highly significant, eight have probabilities (p-values) smaller than $p < 2 \cdot 10^{-5}$, with the worst fit is for the parameter β_{0i} of the RA, which has $p = 0.014$. All the exponents are smaller than one $\beta_{\text{Prio,RA,TS,X}} = (0.82, 0.69, 0.66, 0.68)$, and they are very similar to each other.

A data-driven approach

A data-driven approach does not make many assumptions like the model in Equation (1). It simply aggregates the data into bins of similar ADT-values, where the bin-width has been chosen so, that each bin contains roughly the same number of data-points. Clearly, other choices are possible and valid as well, doing this quantile-based approach has the advantage to produce similar statistics in each bin, at the expanse of the bin-width. Within each bin, a number of statistical metrics could be computed, for simplicity, the mean value is chosen here. The result is shown in Figure 2. While the general form of the relationship between N and Q is roughly the same (as it should), a number of interesting differences could be seen.

First of all, this approach makes it much clearer where there are actually real data, and where extrapolation is applied; this is even more visible in Figure 3, where this approach is compared directly with the glm-ansatz.

As before, there is only a small difference in safety between a priority-controled intersection, and a round-about. The unregulated intersections are the safest, and the intersections controlled by traffic signals display the lowest level of safety. We think that these results will have to be modified once a more thorough quality control of the input data is established.

What is also interesting is that at least the curve for the unregulated intersections (and partially the one for the signalized intersections as well) have a very interesting behavior that is not compatible with the power-law assumption of Equation (1): for large demand the number of crashes seem to saturate. This is not in line with most results; the power-law approach yields ever-increasing crash-numbers with demand, even for small values of the exponent β. However, we have seen such a behavior in other investigations[16]. It might be due to the fact, that a large demand may slow vehicles down, and this would at least cut the number of crashes with injured persons. However, the very details of such a mechanism depend on the layout and design of each intersection, so additional investigations are needed to clarify this.

However, note that the quality of the data presented here is not overwhelming, so these results are not that reliable. However, it demonstrates, that some care has to be taken with forcing models too strongly onto data, one may in fact be suspicious that the approach with Equation (1) which is ubiquitous in traffic safety research may lead to the overlooking of some features of real data.

It is possible to dig a bit deeper into these data by directly comparing the data-driven approach with the glm-results. The result is shown in Figure 3, where the data-driven relationships have been decorated with the confidence intervals (level 0.95) of the mean-values. These have been computed by a boot-strapping approach, and the results have been put into a co-ordinate frame where the x- and the y-axis have a logarithmic scaling: this zooms in on the small values in x and y, and it transforms the glm-fits into straight lines. As mentioned already, this also makes it quite easy to see where the glm does an extrapolation.

This direct comparison seems a bit clearer about the tendency to a saturation of the crash-numbers for large demand. In addition, especially for the uncontrolled intersection organization, the increase for the small demands differs from the power-law significantly. This might be the point where the privacy restrictions of the GCDB is hindering more insights: one may speculate, that this comes from a larger amount of single vehicle crashes, and this number may increase linearly with Q, and therefore with a larger exponent than $\beta_{\mathrm{X}} = 0.86$.

Also, it is nice to see that roundabouts and traffic signals are typically placed only on intersections with a larger demand, in line with German regulations.

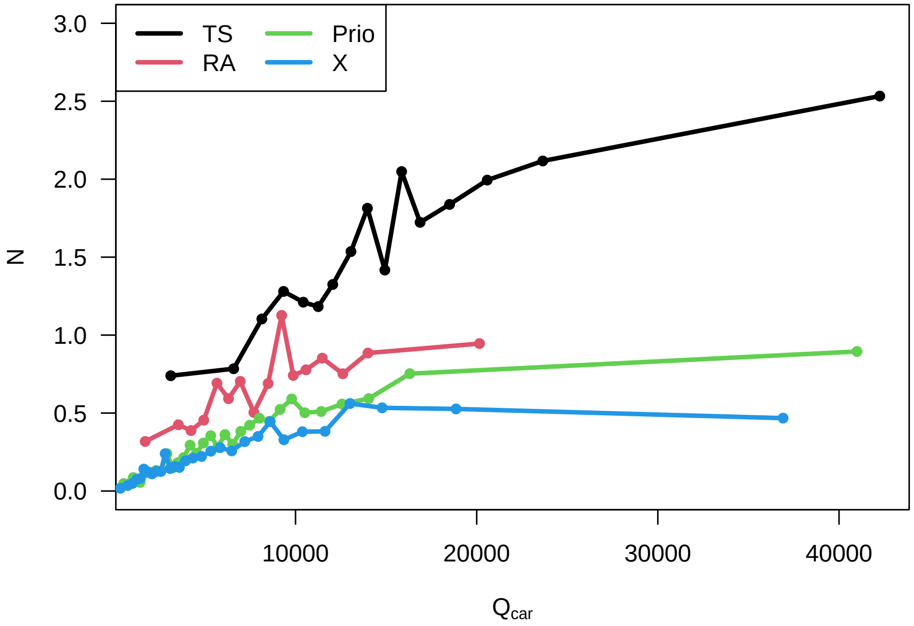

Figure 2. The result of the data-driven approach. Again, there are four lines for the four intersection controls.

Conclusions

The short approach described here demonstrates the possibilities, but also the weaknesses of an approach that tries to use large data-bases of road safety data. One of the challenges is clearly to have some kind of quality control of the data, which is lacking here. Nevertheless, the approach can partly reproduce some features that are known already, however it also demonstrates that with more data, and with an analysis approach that is not too strongly constrained by assumed relationships between the crash-numbers and the exposition, more and better information could be unearthed. This may lead to a better understanding of traffic safety and the factors that may decrease it, and ultimately to safer roads as well.

References

1. RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2022. URL http://www.rstudio.com/.
2. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.
3. QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2022. URL http://qgis.osgeo.org.

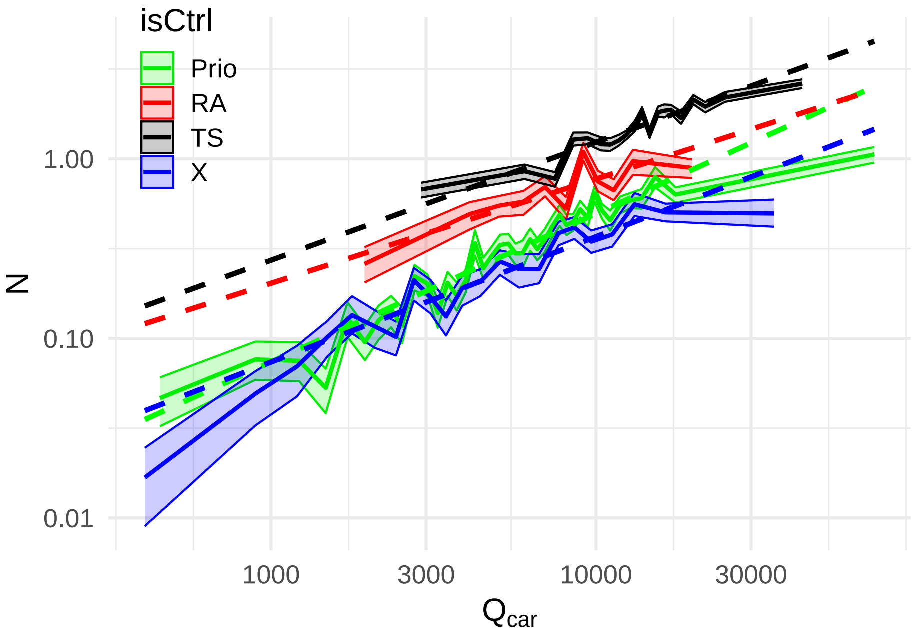

Figure 3. The result of the data-driven approach with boot-straped confidence intervals, compared against the glm-fits.

4. Daniels S, Nuyts E and Wets G. The effects of roundabouts on traffic safety for bicyclists: An observational study. *Accident Analysis & Prevention* 2008; 40(2): 518 – 526. DOI:10.1016/j.aap.2007.07.016. URL http://www.sciencedirect.com/science/article/pii/S0001457507001352.
5. Daniels S, Brijs T, Nuyts E et al. Explaining variation in safety performance of roundabouts. *Accident Analysis & Prevention* 2009; 42(2): 393–402. DOI:10.1016/j.aap.2009.08.019. URL https://www.sciencedirect.com/science/article/pii/S0001457509002280.
6. Jensen SU. Safe roundabouts for cyclists. *Accident Analysis & Prevention* 2017; 105: 30 – 37. DOI:10.1016/j.aap.2016.09.005. URL http://www.sciencedirect.com/science/article/pii/S0001457516303359. Improving cyclist safety through scientific research, ICSC2015.
7. Elvik R. Road safety effects of roundabouts: A meta-analysis. *Accident Analysis & Prevention* 2017; 99: 364–371. DOI:https://doi.org/10.1016/j.aap.2016.12.018. URL https://www.sciencedirect.com/science/article/pii/S0001457516304560.
8. DeStatis. German accident atlas, 2022. URL https://unfallatlas.statistikportal.de/.
9. NW. Strassen.nrw, 2022. URL https://www.strassen.nrw.de/en/startseite.html.

10. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2022.
11. Leich A, Fuchs J, Srinivas G et al. Traffic safety at german roundabouts—a replication study. *Safety* 2022; 8(3): 50. DOI:https://doi.org/10.3390/safety8030050.
12. Hauer E. Statistical Road Safety Modeling. *Transportation Research Record* 2004; 1897: 81–87. DOI:10.3141/1897-11.
13. Lord D and Mannering F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* 2010; 44: 291–305. DOI:10.1016/j.tra.2010.02.001.
14. Hughes B, Newstead S, Anund A et al. A review of models relevant to road safety. *Accident Analysis & Prevention* 2015; 74: 250 – 270. DOI:10.1016/j.aap.2014.06.003. URL http://www.sciencedirect.com/science/article/pii/S0001457514001766.
15. Ambros J, Jurewicz C, Turner S et al. An international review of challenges and opportunities in development and use of crash prediction models. *European Transport Research Review* 2018; 10(2): 35. DOI:10.1186/s12544-018-0307-7. URL https://doi.org/10.1186/s12544-018-0307-7.
16. Wagner P, Hoffmann R and Leich A. Observations on the relationship between crash frequency and traffic flow. *Safety* 2021; 7(1). DOI:10.3390/safety7010003. URL https://www.mdpi.com/2313-576X/7/1/3.