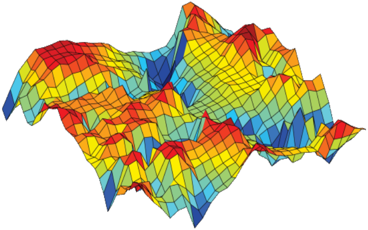
Predictive Analysis of Road Incidences

Exploratory Lines

BDigital

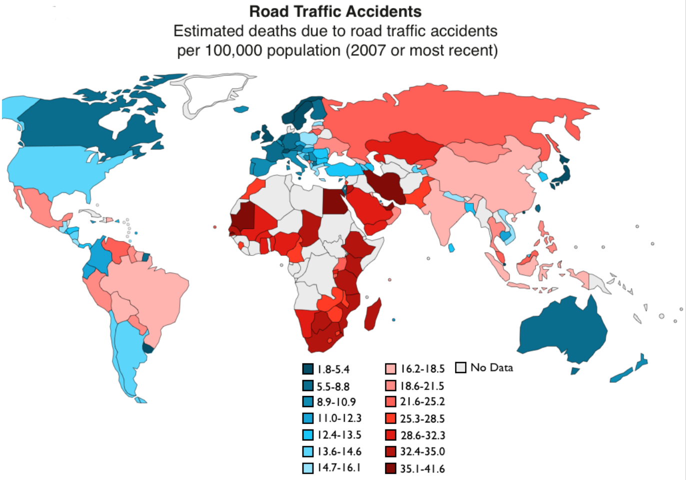
May 21, 2014



Introduction

- Despite significant improvements in vehicle technology and road engineering over the last 40 years, on a world-wide scale road accidents are still one of the main accidental causes of death and injury (WHO, 2004).
- In a study regarding the relation between extreme weather events (e.g.: flash floods) and road accidents it was found that motorists account for 84 % of the totality of the victims [cite].
- The assessment of the occurrence of road accidents has therefore attracted the attention of many researchers.

Introduction (cont.)



source: http://www.geocurrents.info

Research Question

What do we want to predict?

Ocurrence (0,1) and location (x,y) of incidents, as a function (pairwise or multivariant) of other variables;

- time of the day;
- weekday;
- traffic density;
- weather (rainfall, snowfall, visibility, winds, etc);
- traffic control devices (e.g.: traffic lights, stop signal, etc);
- roadway geometries (e.g.: steepness, intersection, degree of curvature, etc) ;
- roadway conditions (road surface, obstacles);
- vehicle type;
- human conditions (characteristics of the driver, fatigue, alcohol);
- involvement of pedestrians;

Some of these variables are continuous in space and most of them change over time.

Modelling Approaches: Classical Statistics

- The most common approach applied in early works is to model the interaction between each dependent variable and accident frequencies by means of conventional (multiple) linear regression models[cite]; these **do not take in account covariance between response variables**.

Modelling Approaches: Classical Statistics

- The most common approach applied in early works is to model the interaction between each dependent variable and accident frequencies by means of conventional (multiple) linear regression models[cite]; these **do not take in account covariance between response variables**.
- Most of these models use a Normal or Poisson distribution that is not well-suited to the scarce nature of accident data. Accident data is often characterized by small observed mean values and a large number of zero counts leading to the well discussed phenomenon of **over-dispersion**: the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given simple statistical model.

Modelling Approaches: Classical Statistics

- The most common approach applied in early works is to model the interaction between each dependent variable and accident frequencies by means of conventional (multiple) linear regression models[cite]; these **do not take in account covariance between response variables**.
- Most of these models use a Normal or Poisson distribution that is not well-suited to the scarce nature of accident data. Accident data is often characterized by small observed mean values and a large number of zero counts leading to the well discussed phenomenon of **over-dispersion**: the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given simple statistical model.
- None of these models addresses explicitly the question of **spatial autocorrelation**: the value of samples taken close to each other are more likely to have similar magnitude than by chance alone.

Modelling Approaches: Data Mining

Statistical models are particularly likely to be preferable when fairly simple models are adequate and the important variables can be identified before modeling. **This is not the case of the large and complex data set of road accidents**.

Data mining can be defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data* [cite].

- Scale differences: data sets can be much larger than in statistics.

Modelling Approaches: Data Mining

Statistical models are particularly likely to be preferable when fairly simple models are adequate and the important variables can be identified before modeling. **This is not the case of the large and complex data set of road accidents**.

Data mining can be defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data* [cite].

- Scale differences: data sets can be much larger than in statistics.
- Conceptual differences: compared with statistics, data mining pays less attention to the large-scale asymptotic properties of its inferences and more to the general philosophy of *learning*. For this fact, data mining has been criticized for being a "black box" [cite].

Modelling Approaches: Data Mining

Statistical models are particularly likely to be preferable when fairly simple models are adequate and the important variables can be identified before modeling. **This is not the case of the large and complex data set of road accidents**.

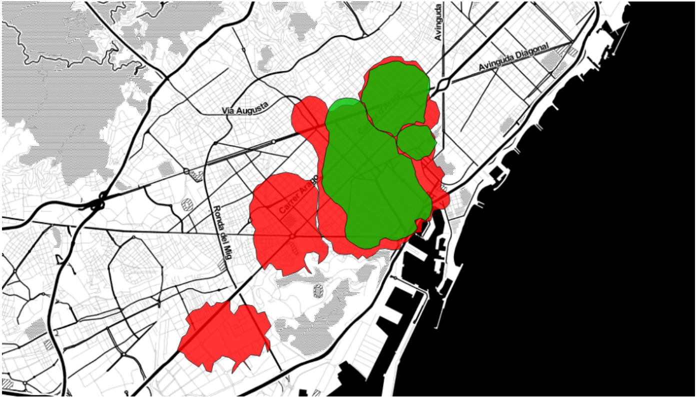
Data mining can be defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data* [cite].

- Scale differences: data sets can be much larger than in statistics.
- Conceptual differences: compared with statistics, data mining pays less attention to the large-scale asymptotic properties of its inferences and more to the general philosophy of *learning*. For this fact, data mining has been criticized for being a ”black box” [cite].
- Dataset collection differences: it is generally a form of secondary data analysis, as very likely the datasets have been collected for other purpose than the one of answering the research question.

Data Mining

There are two major groups of tasks in Data Mining [cite]:

- Description: affinity group and clustering;
- Prediction: classification and estimation;



DBSCAN clusters of tweets during MWC vs a *normal* day

Modelling Techniques

Often Data Mining Techniques are combined together, or with traditional statistical methods, to yield better results:

- **Multivariate regression analysis + Bayesian Probabilistic Networks (BPN)** [cite];
- Random Forest + Multivariate Adaptive Regression Splies (MARS) [cite];
- Classification and Regression Trees (CART) [cite];
- Genetic Mining Rule + Logit Model [cite];
- **Frequent Item Sets [cite]**;
- Support Vector Machines (SVM);
- **Self Organizing Maps (SOM)**;

Bayesian Approach

In this case study[cite] it is used a mixed-model to predict the ocurrence of road accidents:

The BPN is based on the regression model, but its parameters are updated by learning algorithms; the update *only replaces the response variables that have been observed in conjunction with the accident*. This approach allows to cope with different degrees of completness of the dataset.

The methodology is illustrated using data of the Austrian rural motorway network (+/- 80 000 vehicles/day) and the quality of the prediction was found to be 73%.

Bayesian Approach

In this case study[cite] it is used a mixed-model to predict the ocurrence of road accidents:

- Multivariate Poisson-lognormal regression analysis, which facilitates taking into account the covariance structure of the model response variables as well as over-dispersion effects.

The BPN is based on the regression model, but its parameters are updated by learning algorithms; the update *only replaces the response variables that have been observed in conjunction with the accident*. This approach allows to cope with different degrees of completness of the dataset.
The methodology is illustrated using data of the Austrian rural motorway network (+/- 80 000 vehicles/day) and the quality of the prediction was found to be 73%.

Bayesian Approach

In this case study[cite] it is used a mixed-model to predict the ocurrence of road accidents:

- Multivariate Poisson-lognormal regression analysis, which facilitates taking into account the covariance structure of the model response variables as well as over-dispersion effects.
- Bayesian Probabilistic Networks (BPN) that take into account aleatory and epistemic uncertainties as well as possibly non-linear dependencies between the risk indicating variables and the response variables.

The BPN is based on the regression model, but its parameters are updated by learning algorithms; the update *only replaces the response variables that have been observed in conjunction with the accident*. This approach allows to cope with different degrees of completness of the dataset.

The methodology is illustrated using data of the Austrian rural motorway network (+/- 80 000 vehicles/day) and the quality of the prediction was found to be 73%.

Association Algorithm

In this case study[cite] an association algorithm is used to obtain a descriptive analysis of accident high risk areas. This algorithm **identifies the accident circumstances that frequently occur together**. We are able to target the highest frequencies through a *minimum support value*.

This algorithm was applied to road accident data in Belgian (1997-1999). Some factors that were frequently linked to accidents:

- "Left" turns;
- Uneven views when approaching an intersection;
- Rainy conditions;
- Involvement of young pedestrians;

Some of these risk factors could be reduced with signing or roundabouts in these zones;

What About Space?

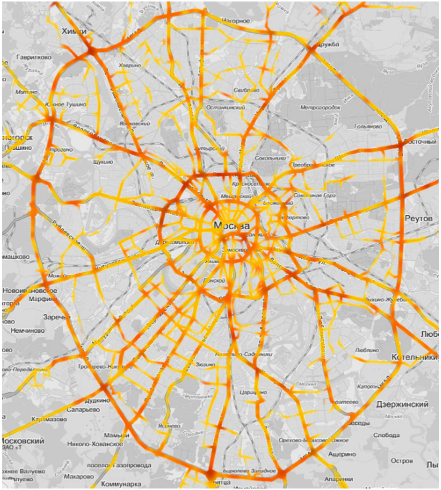


Figure: Heat map of road accidents in Moscow; source:
http://mapsforhumans.com/2011/07/road-accidents-heat-map-of-moscow/

What About Space? (cont.)

Road accidents happen in time and **space**. The contact with the spatial dimension of the phenomena starts with the input dataset. The most frequent approach to represent the road structure is to use networks. Networks represent lanes as *links* and intersections as *nodes*.

Figure: Road network in Hillsborough [cite]

For longer roads, it may be actually useful to split them into **homogeneous sections**.

What About Space? (cont.)

But the space dimension does not end up in the representation of the input variables (implicit). Space **relationships should also be explicitly stated in the model**.

- Spatial Auto correlation (as opposed to spatial independence) is an arrangement of the accidents where the locations are related to each other.
- Numerous studies have shown that accounting for spatial correlation among the analyzed observations is a big step toward a better safety assessment [cite].
- This is actually a *weakness* of many models; on the Baysean Network model [cite], although acknowledged, spatial correlations were not considered in the development of the risk model; on conditionally autoregressive models (CAR) spatial effects are introduced as random.

Geographical Information Systems (GIS) can be a valuable tool for representing/integrating the different variables on a geographical space.

Considering Space: A Simple Example

Study of how circumstances affect people in episodes of bad weather [cite].Data from newspappers was collected for a period of 10 yrs, in the region of Calabria.

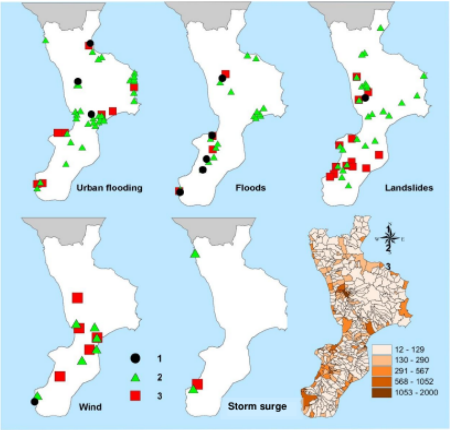


Figure: Localisation of damage to people sorted according to the type of damaging agent. (1) Victims, (2) injured, and (3) involved people [cite]

Considering Space: A Simple Example (cont)

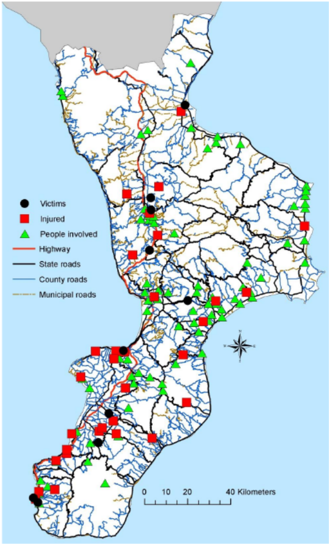


Figure: Localisation of damage to people along the Calabria road network [cite]

- Urban flooding and flash floods were found to be the most damaging events;
- Drivers accounted for 84% of the victims during these episodes.
- The mapping of damaging effects pointed out the regional sectors for which the high frequency of damaging events suggests planning further in-depth examinations;
- The identification of these crictical points can support local regulator interventions, that might change damage incidences in the future.

Considering Space: A more complex example

Self Organizing Maps

Targeting More Specific Aspects of Incidences:

- Identify and describe secondary crashes;
- Identify hots pots or ”black zones”, and model their behaviour;
- Model crashes at specific type of road entity (e.g.: unsignalised intersections);
- Classify incidences (e.g.: victims, injured, involved people);
- Predict the type of accidents (casualties or damages);
- ...

Other Relevant Studies:

- Predict traffic levels in unmeasured locations (Kriging)[cite];
- Estimation of road traffic congestion (MCE, ANN)[cite];
- Traffic simulation model (CA)[cite];
- Traffic Forecast System[cite];
- Predict travel time [cite];

Visualization

3D viz: give some ”eye candy” examples

References