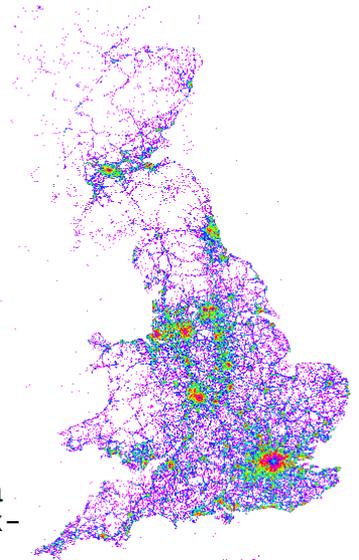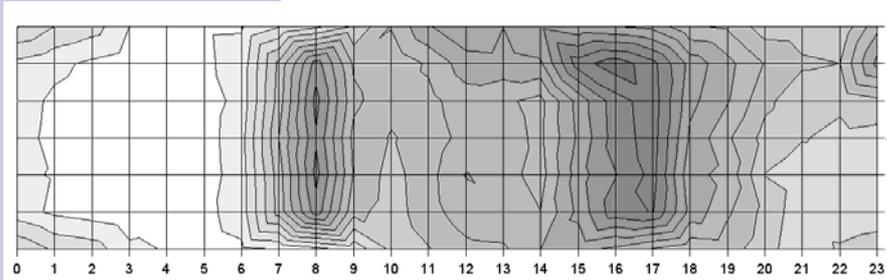**The Art of Doing by Knowing**

- **the challenge:** a 1.5GB database containing 21 years' worth of UK traffic accident reports
- **the team:** a large international consortium of 7 data mining groups applied their extensive data mining expertise
- **the results:** a range of models that helped identify common types of accidents and particularly dangerous circumstances

Improving road safety with data mining

Sol-EU-Net: Data Mining and Decision Support

Sol-EU-Net: Data Mining and Decision Support

Dr Peter Flach
Department of Computer Science
University of Bristol
The Merchant Venturers Building
Woodland Road
Bristol BS8 1UB, United Kingdom

Phone: +44-117-954-5162
Fax: +44-117-954-5208
Email: Peter.Flach@bristol.ac.uk

The SolEuNet consortium worked with Hampshire County Council (UK) in order to obtain a better insight into how the characteristics of accidents may have changed over the past 20 years as a result of improvements in highway and vehicle design. A range of data mining techniques has been applied including innovative visualization techniques, text mining, subgroup discovery, and association rules. These techniques helped pointing out data quality issues, identified common types of accidents, and particularly dangerous roads and circumstances.
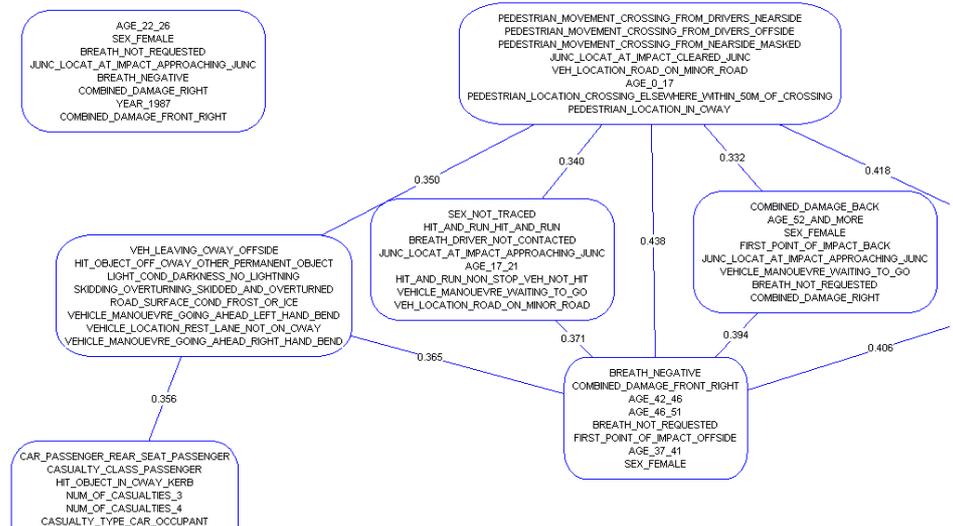
Background

Hampshire County Council, a local authority in the United Kingdom, wants to obtain a better insight into how the characteristics of traffic accidents may have changed over the past 20 years as a result of improvements in highway design and in vehicle design. Data was available in the form of a 1.5GB database containing police traffic accident reports for all UK accidents that happened in the period 1979-1999. The database records details about the accident, together with the vehicles and casualties involved, which in principle can be easily discerned by the police either at the scene of the accident, or when they are reported to the police at a later date after the accident.

Objectives

The primary objective of this project was to improve the understanding of road safety in order to reduce the occurrences and severity of accidents. The following broad areas for study were identified by the end-user: influence of road surface condition; influence of skidding; influence of location (for example: junction approach); and influence of street lighting. Each of these areas were interesting for trend analysis: long-term overall trends, regional trends, urban trends, and rural trends. Also, the comparison of different kinds of locations is interesting: for example, rural versus metropolitan versus suburban. Additional data mining targets include: finding particular types of accident that become more prevalent; trend analysis on types of vehicle damage; and correlation analysis between accident characteristics and age of drivers or speed of cars.

Results

Inevitably in a database of this magnitude, the data is of variable quality. The consortium identified a number of data quality issues that need to be taken into account before models can be obtained from the data. These issues include: particularly noisy location data over the years, splitting and merging of local authorities, evolution of the traffic accident report form, and a relatively large proportion of missing or unknown data. Identification of these data quality issues is a valuable result in itself as it will assist future data analysis methods.

In close collaboration with the end-user, a range of data mining techniques were applied to this dataset, including text mining, clustering of time series, subgroup discovery, multi-relational data mining, and association rule learning. These techniques were particularly suited for this domain as the data mining task was exploratory and descriptive rather than predictive. Using these techniques, the consortium identified clusters of common accidents; clusters of local authorities displaying similar trends; subgroups of similar severe or fatal accidents; and conditions under which serious accidents were more likely to happen.