Ciudad 2020
Exploratory lines and further work

Joana Simoes, BDigital
27/05/2014

bdigital BARCELONA DIGITAL TECHNOLOGY CENTRE

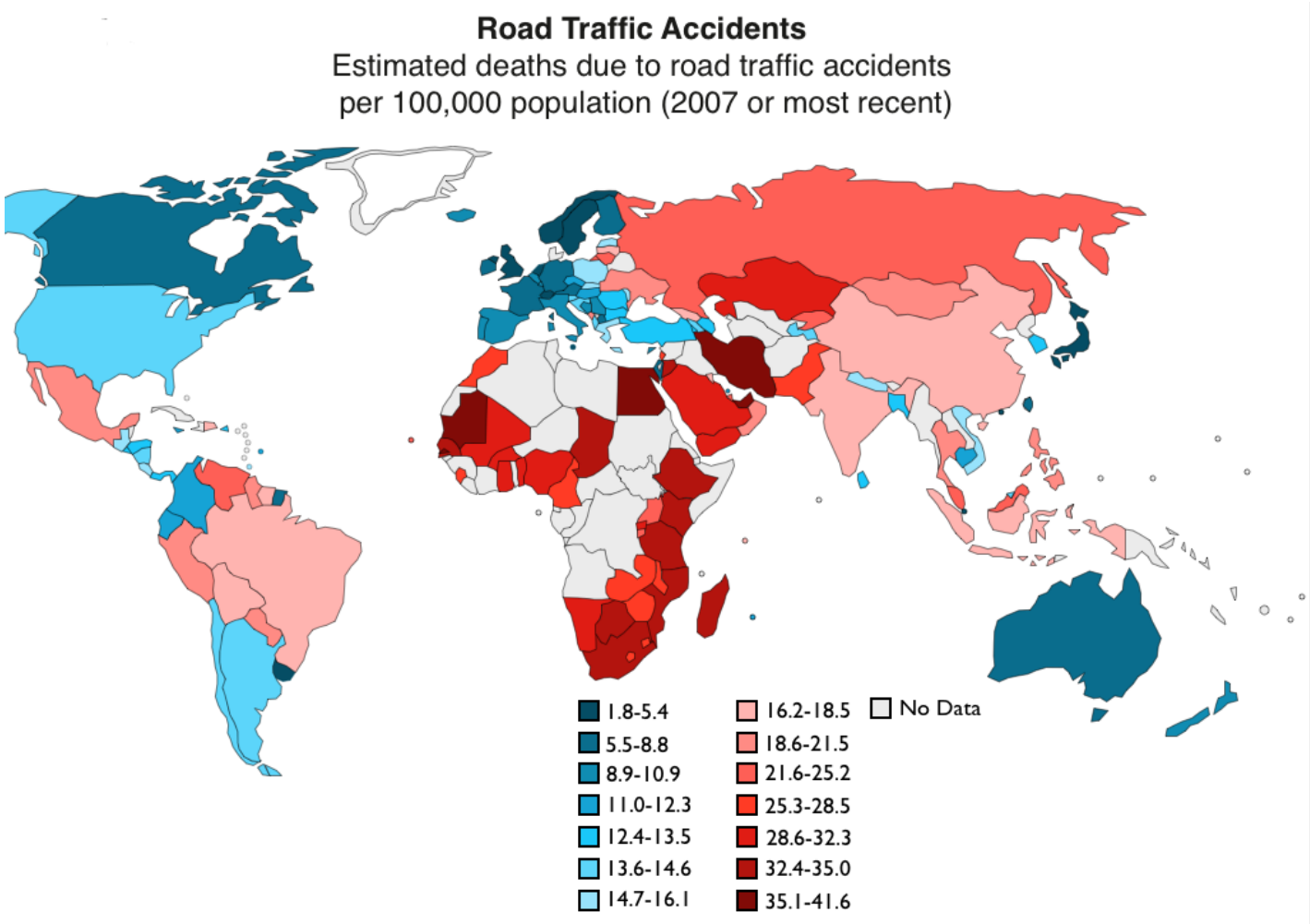
TECNIO
Be tech. Be competitive

Introduction

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 [1].
- The assessment of the occurrence of road incidences in general, and accidents in particular, has 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;

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

General Characteristics of Incidences Datasets:

- small observed mean values and a large number of zero counts (leads to **over-dispersion**);
- **nonlinear** relationships between explanatory and response variables;
- **covariance** between explanatory variables;
- **spatial autocorrelation**: the value of samples taken close to each other are more likely to have similar magnitude than by chance alone;

Figure: data for a freeway in Taiwan, with 373 Km [2]

Table 1
Sample summary of statistics of characteristics of road sections

	Minimum	Maximum	Mean	Standard deviation
Accident frequency (per year)	0	6	0.72	0.95
Degree of horizontal curve (angle, in degree, subtended by a 100 m arc, equal to 18,000/(π × Radius))	0	14.3	1.66	1.99
Vertical grade (percent)	−5.3	5.3	0	1.44
ADT (in 1000's of vehicles)	29.31	116.56	52.52	18.98
Truck ADT (in 1000's of vehicles)	1.00	13.56	5.99	2.07
Tractor-trailer ADT (in 1000's of vehicles)	1.36	8.29	4.96	1.39
Bus ADT (in 1000's of vehicles)	0.37	11.76	1.84	1.16
Peak hour factor	0.77	0.97	0.91	0.04
Number of days with precipitation	49	174	93.5	26.6
Annual precipitation (millimeters)	1195	3749	1669.1	692.2

Context Overview

Modelling Approaches: Classical Statistics

- The most common approach applied in early works is to model the interaction between each dependent variable and incidence frequencies by means of conventional (multiple) linear regression models [3]; these **do not take in account covariance between explanatory variables**.

Modelling Approaches: Classical Statistics

- The most common approach applied in early works is to model the interaction between each dependent variable and incidence frequencies by means of conventional (multiple) linear regression models [3]; these **do not take in account covariance between explanatory variables**.
- Most of these models use a Normal or Poisson distribution that is not well-suited to the scarce nature of incidence data. Often there is a greater variability (statistical dispersion) in a data set than would be expected based on a given simple statistical model (**over-dispersion**).

Modelling Approaches: Classical Statistics

- The most common approach applied in early works is to model the interaction between each dependent variable and incidence frequencies by means of conventional (multiple) linear regression models [3]; these **do not take in account covariance between explanatory variables**.
- Most of these models use a Normal or Poisson distribution that is not well-suited to the scarce nature of incidence data. Often there is a greater variability (statistical dispersion) in a data set than would be expected based on a given simple statistical model (**over-dispersion**).
- None of these models addresses explicitly the question of **spatial autocorrelation**;

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 incidences**.

Data mining can be defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data* [4].

- 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 incidences**.

Data mining can be defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data* [4].

- 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" [5].

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 incidences**.

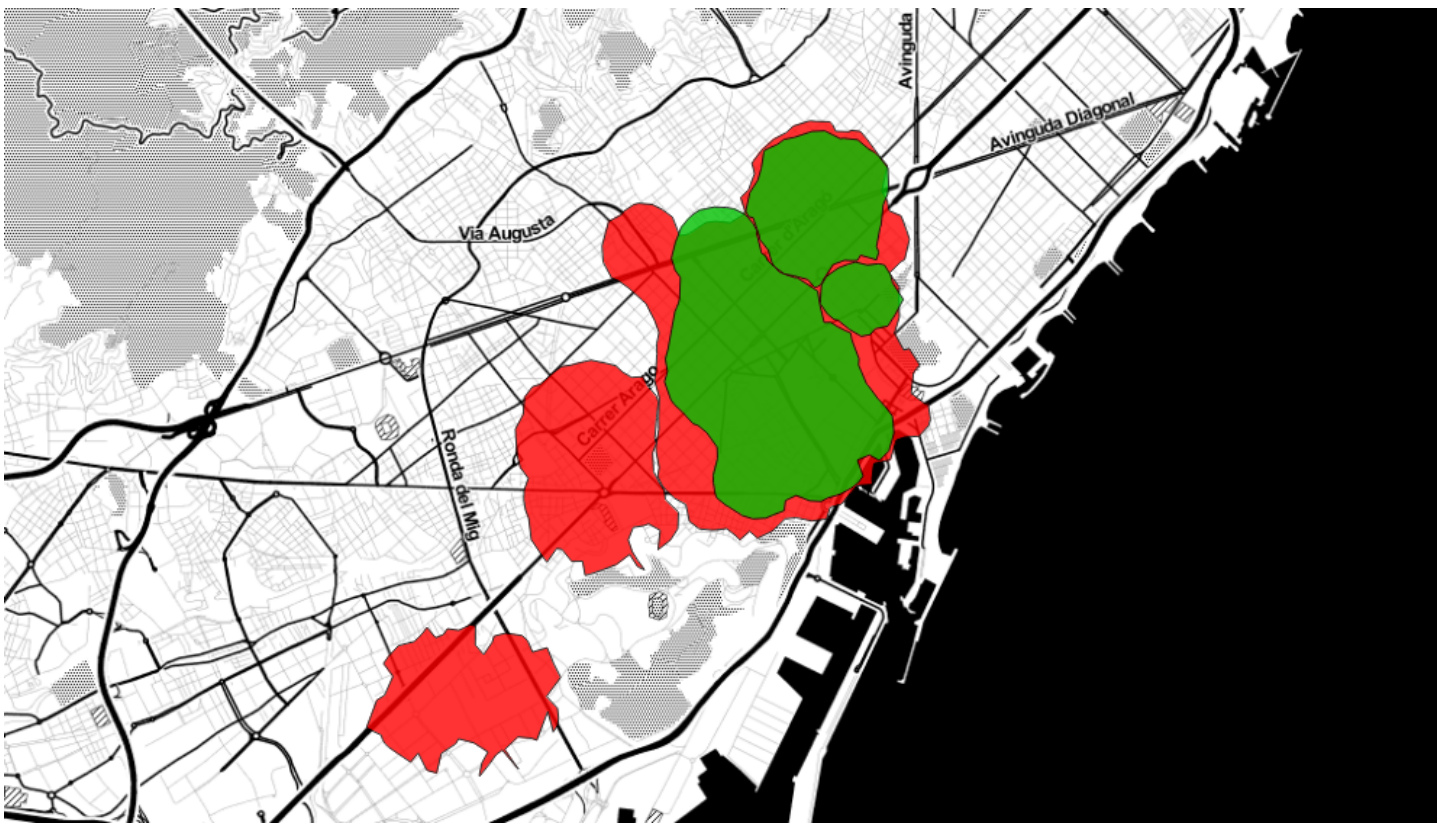
Data mining can be defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large amounts of data* [4].

- 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" [5].
- 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 [4]:

- 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)** [3];
- Random Forest + Multivariate Adaptive Regression Splies (MARS) [7];
- Classification and Regression Trees (CART) [2];
- Genetic Mining Rule + Logit Model [6];
- **Frequent Item Sets** [4];
- Support Vector Machines (SVM);
- **Self Organizing Maps (SOM)**;

Case Study: Bayesian Approach

In this case study [3] 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%.

Case Study: Association Algorithm

In this case study [4] 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 by adding simple signs, or roundabouts in these zones;

What About Space?

Everything is related to everything else, but closer things are more closely related.

Tobler, 1970

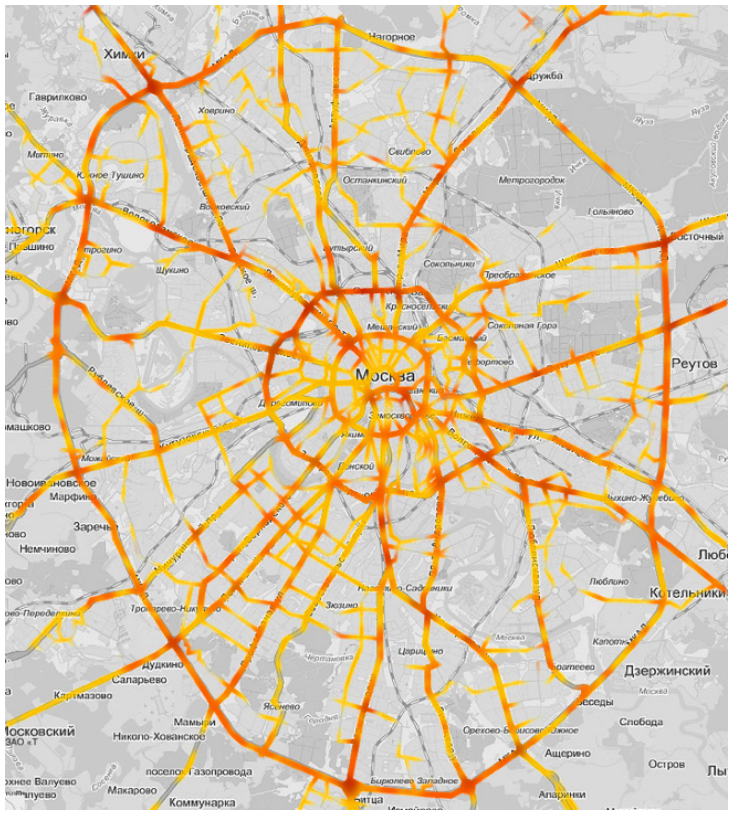


Figure: Heat map of road accidents in Moscow; source: http://mapsforhumans.com/2011/07/road-accidents-heat-map-of-moscow/

What About Space? (cont.)

Road incidences 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 [5]

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

What About Space? (cont.)

But the space dimension does not cease 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 [5].
- This is actually a *weakness* of many models; on the Baysean Network model [3], 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 [1]. Data from newspappers was collected for a period of 10 yrs, in the region of Calabria.

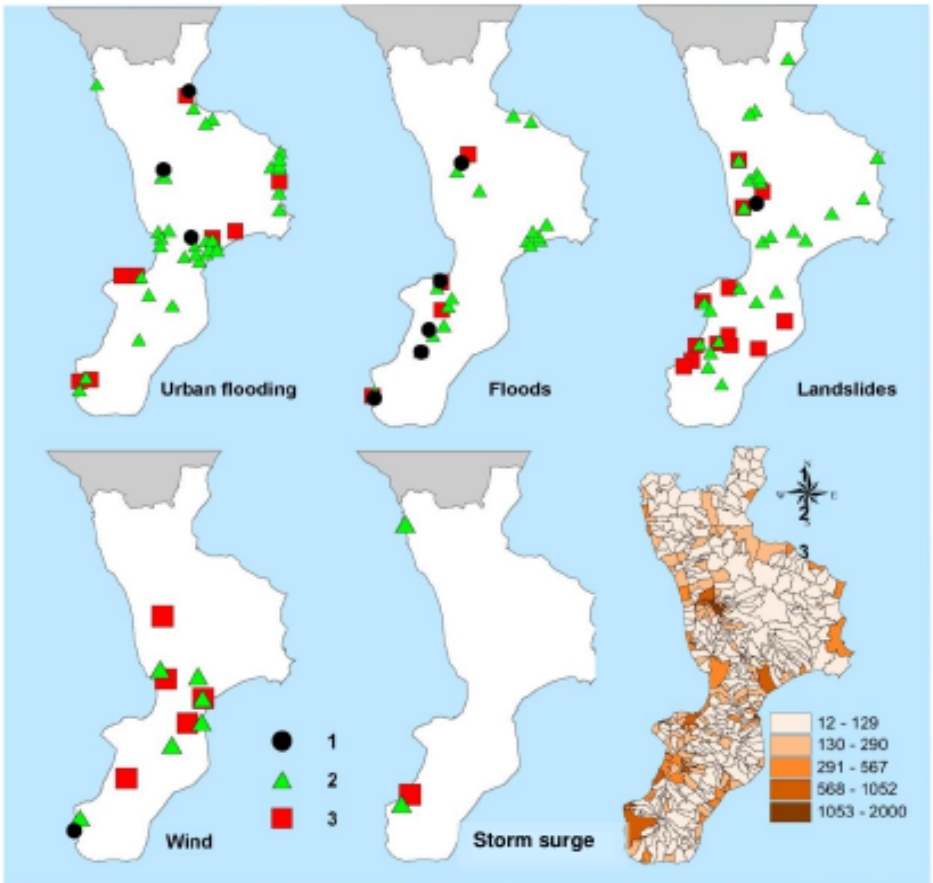


Figure: Localisation of damage to people sorted according to the type of damaging agent. (1) Victims, (2) injured, and (3) involved people [1]

Considering Space: A Simple Example (cont)

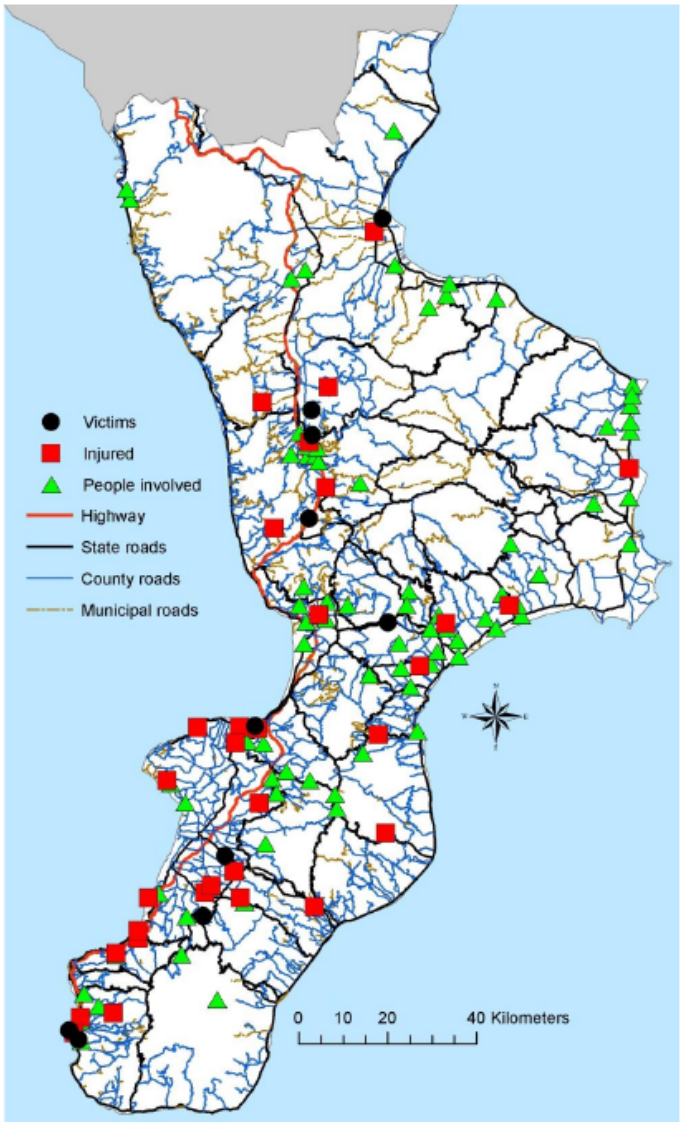


Figure: Localisation of damage to people along the Calabria road network [1]

- 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

The Self Organizing Map (SOM) is an **artificial neural network** based on an **unsupervised learning process** that performs a gradual and **nonlinear** mapping of high dimensional input data onto an ordered and structured array of nodes, generally of lower dimension [8].

- The SOM compresses information and reduces dimensionality;

Considering Space: A more complex example

The Self Organizing Map (SOM) is an **artificial neural network** based on an **unsupervised learning process** that performs a gradual and **nonlinear** mapping of high dimensional input data onto an ordered and structured array of nodes, generally of lower dimension [8].

- The SOM compresses information and reduces dimensionality;
- It transforms nonlinear statistical relationships in geometric relationships;

Considering Space: A more complex example

The Self Organizing Map (SOM) is an **artificial neural network** based on an **unsupervised learning process** that performs a gradual and **nonlinear** mapping of high dimensional input data onto an ordered and structured array of nodes, generally of lower dimension [8].

- The SOM compresses information and reduces dimensionality;
- It transforms nonlinear statistical relationships in geometric relationships;
- It is a visualization method for multidimensional data;

Considering Space: A more complex example

The Self Organizing Map (SOM) is an **artificial neural network** based on an **unsupervised learning process** that performs a gradual and **nonlinear** mapping of high dimensional input data onto an ordered and structured array of nodes, generally of lower dimension [8].

- The SOM compresses information and reduces dimensionality;
- It transforms nonlinear statistical relationships in geometric relationships;
- It is a visualization method for multidimensional data;
- Unlike other NN models, it **preserves topological relationships** units that are close in the output space, are also next to each other in the data space.

Considering Space: A more complex example

The Self Organizing Map (SOM) is an **artificial neural network** based on an **unsupervised learning process** that performs a gradual and **nonlinear** mapping of high dimensional input data onto an ordered and structured array of nodes, generally of lower dimension [8].

- The SOM compresses information and reduces dimensionality;
- It transforms nonlinear statistical relationships in geometric relationships;
- It is a visualization method for multidimensional data;
- Unlike other NN models, it **preserves topological relationships** units that are close in the output space, are also next to each other in the data space.
- Unlike other models (e.g.: Support Vector Machines), as the training is unsupervised, it can prevent the analyst's arbitrary perspective.

Self Organizing Maps

The process of creating SOMs undergoes two main stages [9]:

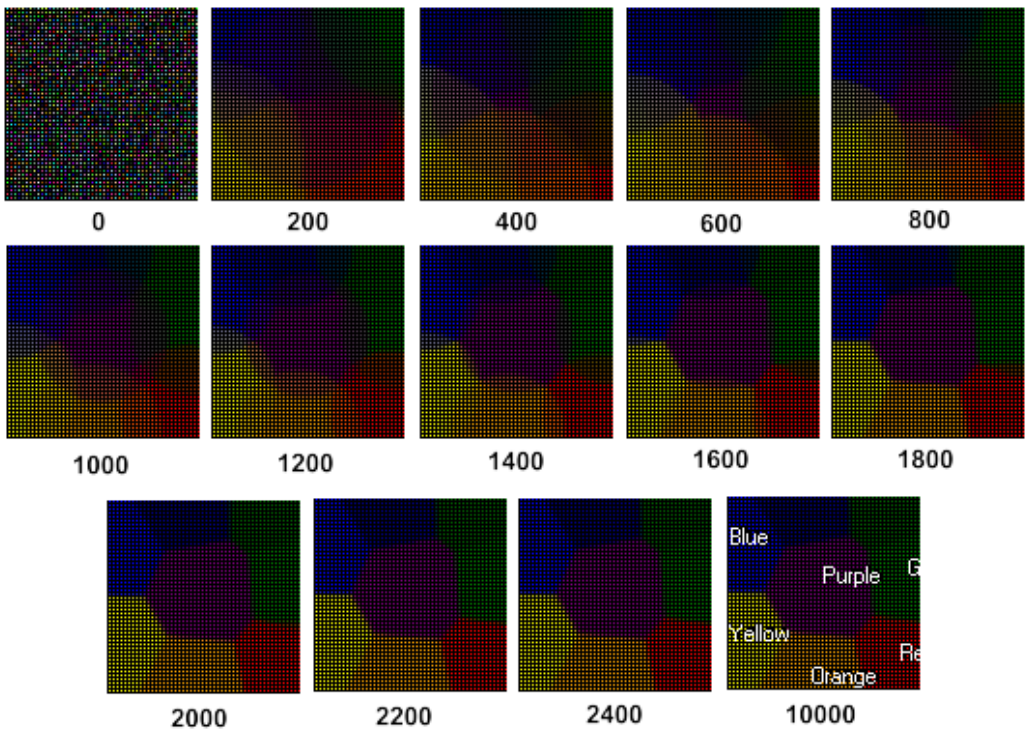


Figure: source: `http://www.mql5.com/en/articles/283`

Self Organizing Maps

The process of creating SOMs undergoes two main stages [9]:

- training: vector quantitization; the neurons *adjust* their weights to the input vector;

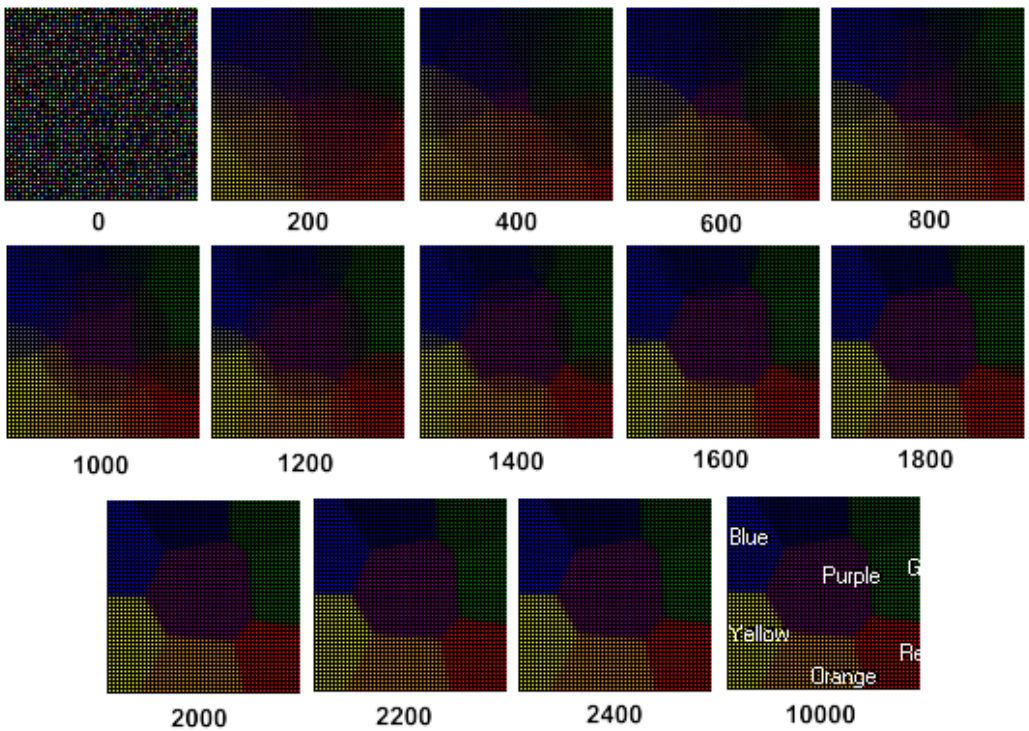


Figure: source: `http://www.mql5.com/en/articles/283`

Self Organizing Maps

The process of creating SOMs undergoes two main stages [9]:

- training: vector quantitization; the neurons *adjust* their weights to the input vector;
- mapping: vector projection; the network classifies a new input dataset;

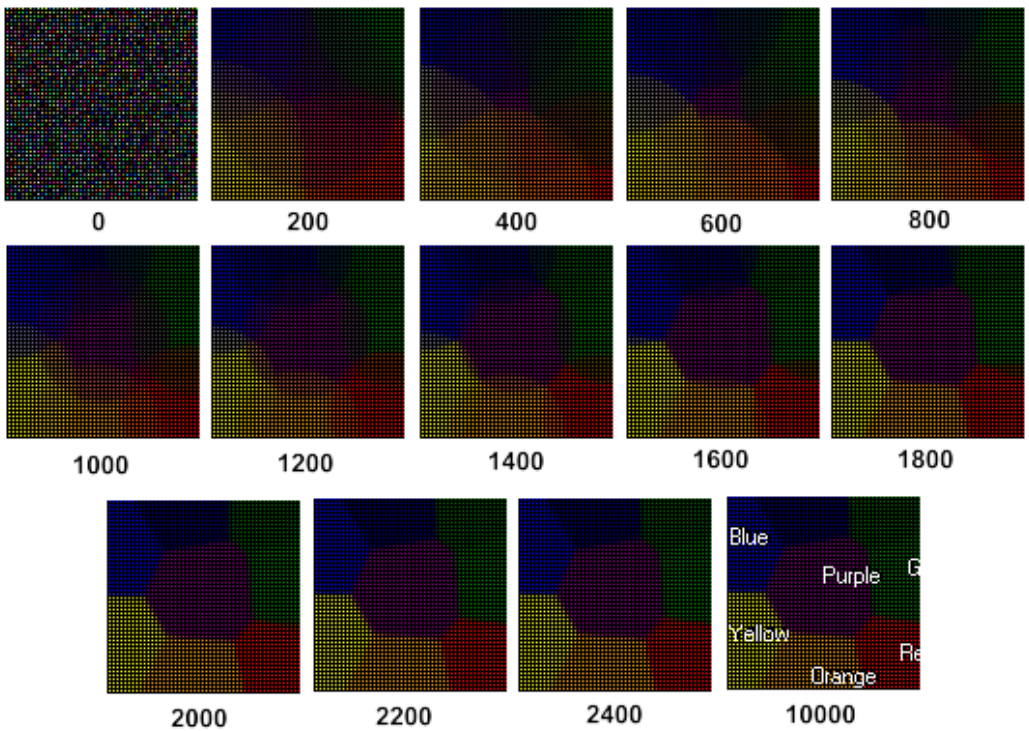


Figure: source: `http://www.mql5.com/en/articles/283`

Self Organizing Maps (cont.)

Apart from the predictive capacity, the algorithm also works as a multidimensional clustering algorithm and a valuable visualization tool (U-Matrix and component planes).

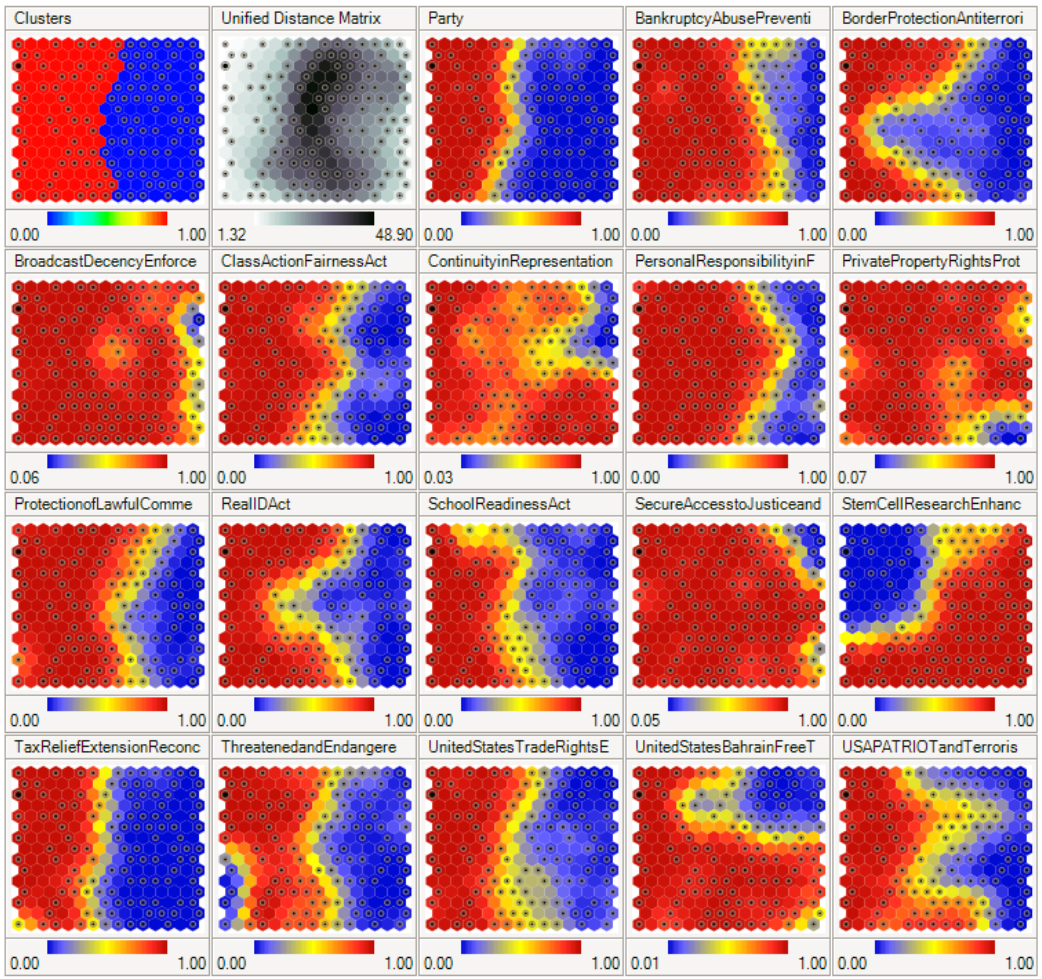
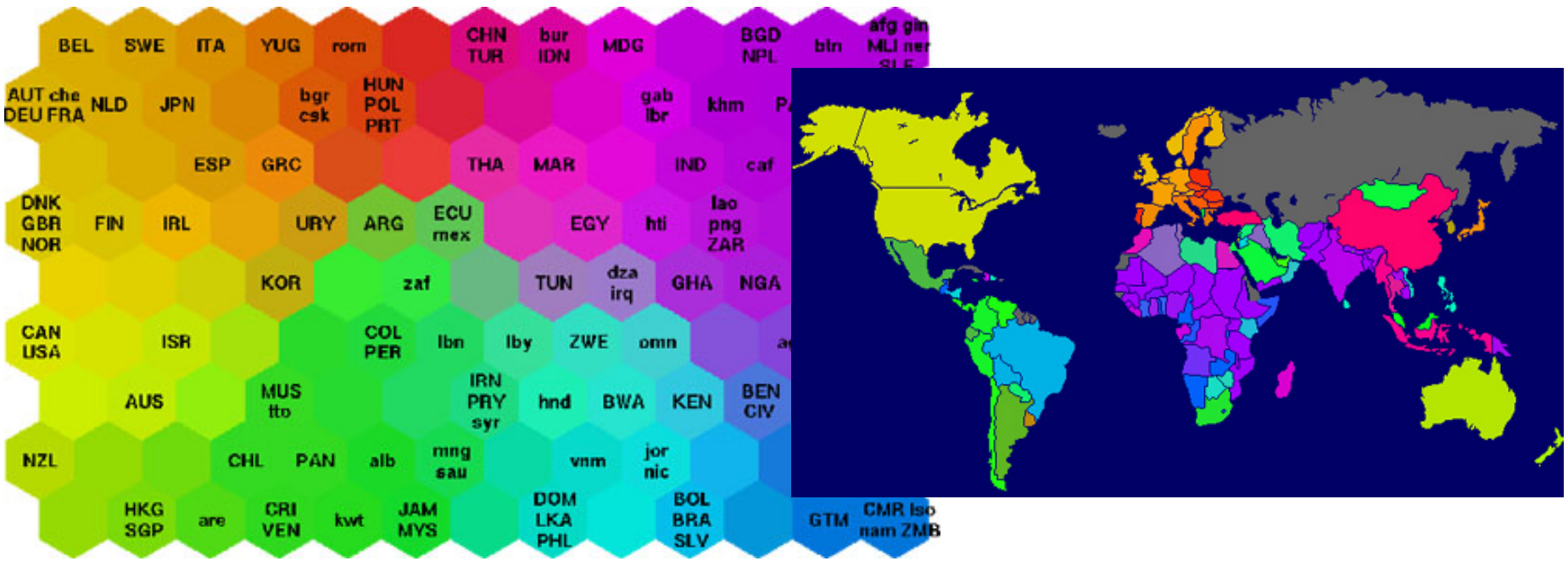


Figure: source: `https://upload.wikimedia.org/wikipedia/commons/7/70/Synapse_Self-Organizing_Map.png`

Self Organizing Maps (cont.)

When applied to geo-referenced data, this technique may allow the explanation of complex structures and phenomena, in a spatial perspective [8].

Going back to geographical space:



source: `http://www.ai-junkie.com/ann/som/som5.html`

Self Organizing Maps (cont.)

SOMs have been used for estimating:

- the risk for cancer patients (*demonstrated accuracy in classification*) [10];
- underlying physical parameters from near infrared planetary spectra (83%-100% accuracy) [12];
- the distance of an astronomic object based on its color (*when compared to other techniques offer a competitive choice in terms of low-RMSE, and percentage of outliers*) [11];
- productivity of goat and sheep farms, based on herd management practices (*the results are coherent with the animal science criteria that are common for this problem*) [9];

The approach has been showed to be an **evidence-based predictive tool with high-knowledge-generation capabilities**, very valuable to **support decision-making** [10].

Model Proposal

SOMs and Road Incidences

Real-world road incidences are characterized by complicated multi-dimensionality mutual relationships.

- We want to propose a SOM to capture and extract the essence of these relationships by **visualizing the results of incidence clustering**.
- Moreover, we want to be able to **predict these clusters of incidences**, for a new set of input variables.

SOMs have been used in a study to estimate the effectiveness of Crash Avoidance technologies [13]. They identified and characterized clusters of accidents, based on the relationships between (normalized) variables.

SOMs and Road Incidences

Real-world road incidences are characterized by complicated multi-dimensionality mutual relationships.

- We want to propose a SOM to capture and extract the essence of these relationships by **visualizing the results of incidence clustering**.
- Moreover, we want to be able to **predict these clusters of incidences**, for a new set of input variables.

SOMs have been used in a study to estimate the effectiveness of Crash Avoidance technologies [13]. They identified and characterized clusters of accidents, based on the relationships between (normalized) variables.

- 1 year of data (2010);

SOMs and Road Incidences

Real-world road incidences are characterized by complicated multi-dimensionality mutual relationships.

- We want to propose a SOM to capture and extract the essence of these relationships by **visualizing the results of incidence clustering**.
- Moreover, we want to be able to **predict these clusters of incidences**, for a new set of input variables.

SOMs have been used in a study to estimate the effectiveness of Crash Avoidance technologies [13]. They identified and characterized clusters of accidents, based on the relationships between (normalized) variables.

- 1 year of data (2010);
- 16,180 fatal passenger car drivers;

SOMs and Road Incidences

Real-world road incidences are characterized by complicated multi-dimensionality mutual relationships.

- We want to propose a SOM to capture and extract the essence of these relationships by **visualizing the results of incidence clustering**.
- Moreover, we want to be able to **predict these clusters of incidences**, for a new set of input variables.

SOMs have been used in a study to estimate the effectiveness of Crash Avoidance technologies [13]. They identified and characterized clusters of accidents, based on the relationships between (normalized) variables.

- 1 year of data (2010);
- 16,180 fatal passenger car drivers;
- 48 variables;

Discussion

- What incidence dataset could we use? (e.g.: time and space dimensions, magnitude of incidence, type of incidence);
- Incidences vs Accidents;
- Would it be possible to have two datasets: one for training and one for validation?
- What type of predictive dataset do we have (e.g.: road geometry, road characteristics, traffic density) ?
- How accurate are these datasets (measuring errors, etc)?

Recommended Dataset Characteristics (aprox.):

- specific type of incident (e.g.: road accidents);
- One year of data;
- +/- geo-located 500 incidences per day;
- Traffic density dataset for the same time period/geographic area;
- Finest possible geographical granularity;
- Similar dataset for calibration;

References I


O. Petrucci and A. A. Pasqua *Damaging events along roads during bad weather periods: a case study in Calabria (Italy)*. Nat. Hazards Earth Syst. Sci., 12, 365-378, 2012.


Li-Yen Chang and Wen-Chieh Chen *Data mining of tree-based models to analyze freeway accident frequency*. Journal of Safety Research 36, 365-375, 2005.


Markus Deublein and Matthias Schubert and Bryan T. Adey and Jochen Kohler and Michael H. Faber *Prediction of road accidents: A Bayesian hierarchical approach*. Accident Analysis & Prevention 51, 274-291, 2013.


Karolien Geurts and Isabelle Thomas and Geert Wets *Understanding spatial concentrations of road accidents using frequent item sets*. Accident Analysis and Prevention 37, 787-799, 2005.


Qiang Zeng and Helai Huang *Bayesian spatial joint modeling of traffic crashes on an urban road network*. Accident Analysis & Prevention 67, 105-112, 2014.


Yu-Chiun Chiou and Lawrence W. Lan and Wen-Pin Chen *A two-stage mining framework to explore key risk conditions on one-vehicle crash severity*. Accident Analysis & Prevention 50, 405-415, 2013.


Mohamed Abdel-Aty and Kirolos Haleem *Analyzing angle crashes at unsignalized intersections using machine learning techniques*. Accident Analysis & Prevention 43, 461-470, 2011.


Jorge Manuel Lourenco Gorricha *Visualization of Clusters in Geo-referenced Data Using Three-dimensional Self-Organizing Maps*. Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Statistics and Information Management, Instituto Superior de Estatistica e Gestao de Informacao da Universidade Nova de Lisboa, 2009.


R. Magdalena and C. Fernandez and J. D. Martin and E. Soria and M.Martinez and M. J. Navarro and C. Mata *Qualitative analysis of goat and sheep production data using self-organizing maps*. Expert Systems, The Journal of Knowledge and Engineering, vol. 26, n 2, 2009.

References II

Leonid Churilov and Adyl Bagirov and Daniel Schwartz and Kate Smith and Michael Dally *Data Mining with Combined Use of Optimization Techniques and Self-Organizing Maps for Improving Risk Grouping Rules: Application to Prostate Cancer Patients*. Journal of Management Information Sy.vton.v/Spring 2005, Vol. 21, No. 4. pp. 85-100, 2005.

M. J. Way and C. D. Klose *Can Self-Organizing Maps Accurately Predict Photometric Redshifts?*. (s): M. J. Way and C. D. Klose Publications of the Astronomical Society of the Pacific, Vol. 124, No. 913 pp. 274-279, 2012.

Lili Zhang and Erzzebet Merenyi and William M. Grundy and Eliot F. Young *An SOM-Hybrid Supervised Model for the Prediction of Underlying Physical Parameters from Near-Infrared Planetary Spectra*. Advances in Self-Organizing Maps, 7th International Workshop, WSOM 2009, USA, pp. 362-372, 2012.

Hitoshi Uno and Yusuke Kageyama and Akira Yamaguchi and Tomosaburo Okabe *Effectiveness study of Crash Avoidance technologies by using Clusteringand Self Organizing Map*. Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV), Seoul, South Korea, 2013.