Module INM433 – Visual Analytics

Lecture 06

Further abilities and topics of visual analytics

given by

prof. Gennady Andrienko and

prof. Natalia Andrienko

Content and objectives

- We talk about predictive visual analytics, i.e., the use of interactive visual interfaces in building and application of predictive models. You will see an example of using visual analytics techniques for predictive traffic modelling and simulation.
- We give a brief overview of the existing visual analytics approaches to analysing further types of data that could not be considered in detail: relationships (networks), images, videos, and texts.
- We present a list of existing visual analytics software systems you can use in practical analysis.
- We wrap up by summarising the main principles of visual analytics.

Predictive visual analytics (by example of traffic)

From analysis of available data to predictive modelling and simulation

Predictive analytics

General notes

- Two main purposes of data analysis:
  - to understand the piece of reality represented (partly!) in the data;
  - to forecast the properties and/or behaviour of this piece of reality beyond the part represented in the data
    - e.g., for other time moments or periods; for other locations; for other objects.
- Statistics and machine learning develop methods for building predictive models from data:
  - Formulas, rules, decision trees, or other formal or digital constructs, which have input and output variables.
  - When some values are assigned to the input variables, the model gives (predicts) the corresponding values of the output variable(s).
- Simulation models developed in various domains aim at forecasting behaviours of objects and phenomena under various conditions.
  - Often based not on data analysis but on theories and/or analogies.

Predictive analytics and visualisation

- Many software packages provide tools for building predictive models
  - R, MatLab, SAS, Weka, JMP, …
- These packages also include visualisation tools
  - ✓ Show data or final results of the modelling.
  - – Do not provide interactive techniques for active involvement of human analysts in the model building process.
    - ☹ The process of model building is a “black box” to human analysts.

Predictive visual analytics

- Predictive visual analytics = building of predictive models with the use of visual analytics approaches.
- Principles:
  - conscious preparation of data (cleaning, transforming, partitioning, …)
  - conscious decomposition of the modelling task
    - a combination of several partial models may be better than a single global model
  - conscious selection of variables, modelling methods, and parameters; creation and comparison of model variants
  - conscious evaluation of model quality
    - Instead of relying on a single numeric measure, study the distribution of the model error over the set of inputs and identify where the model performs poorly.
  - conscious refinement of models
    - targeted improvement in the parts where the performance is poor, e.g., through (further) decomposition

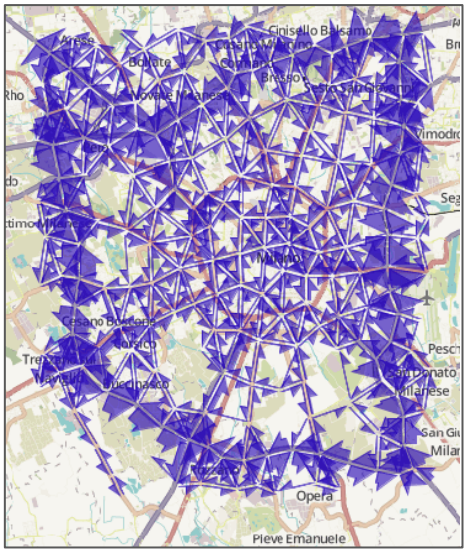
Predictive visual analytics by example

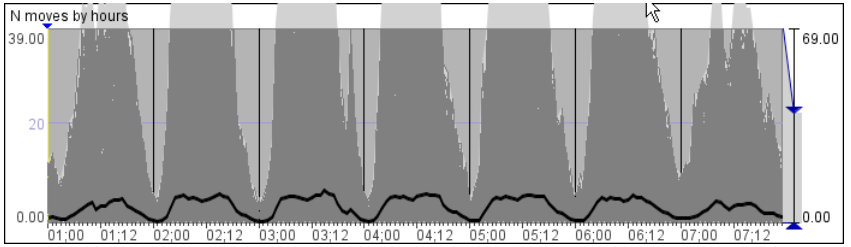
- Given: historical traffic data (vehicle trajectories) supposedly representing movements under usual conditions.
- Question 1: How to utilize these data for predicting regular traffic flows?
- Question 2: How to utilize these data for predicting extraordinary mass movements in special cases?

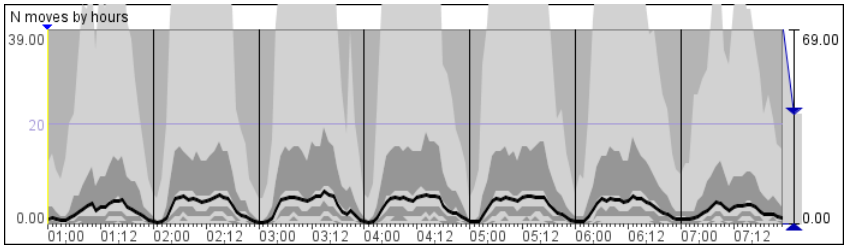
- Example dataset: GPS tracks of cars in Milan

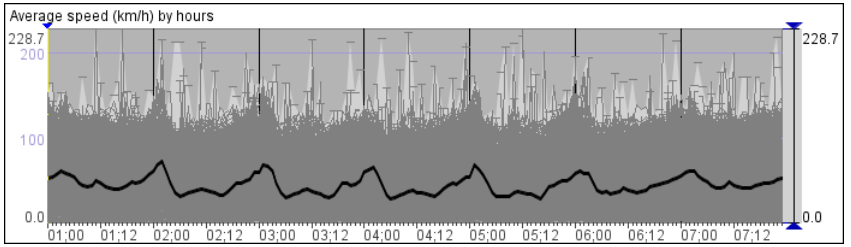
Data transformation: ST aggregation

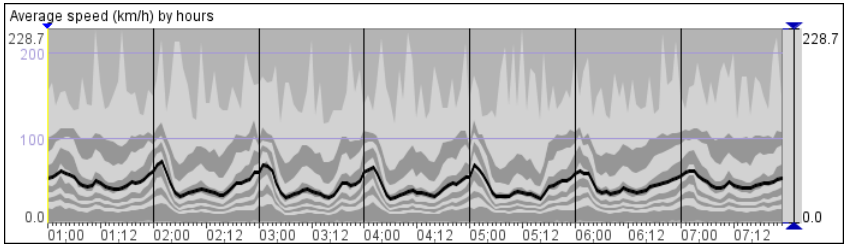
- Divide the territory into cells.
- Divide the time into hourly intervals.
- For each time interval and each ordered pair of neighbouring cells P → Q count the vehicles that moved from P to Q and compute their mean speed.









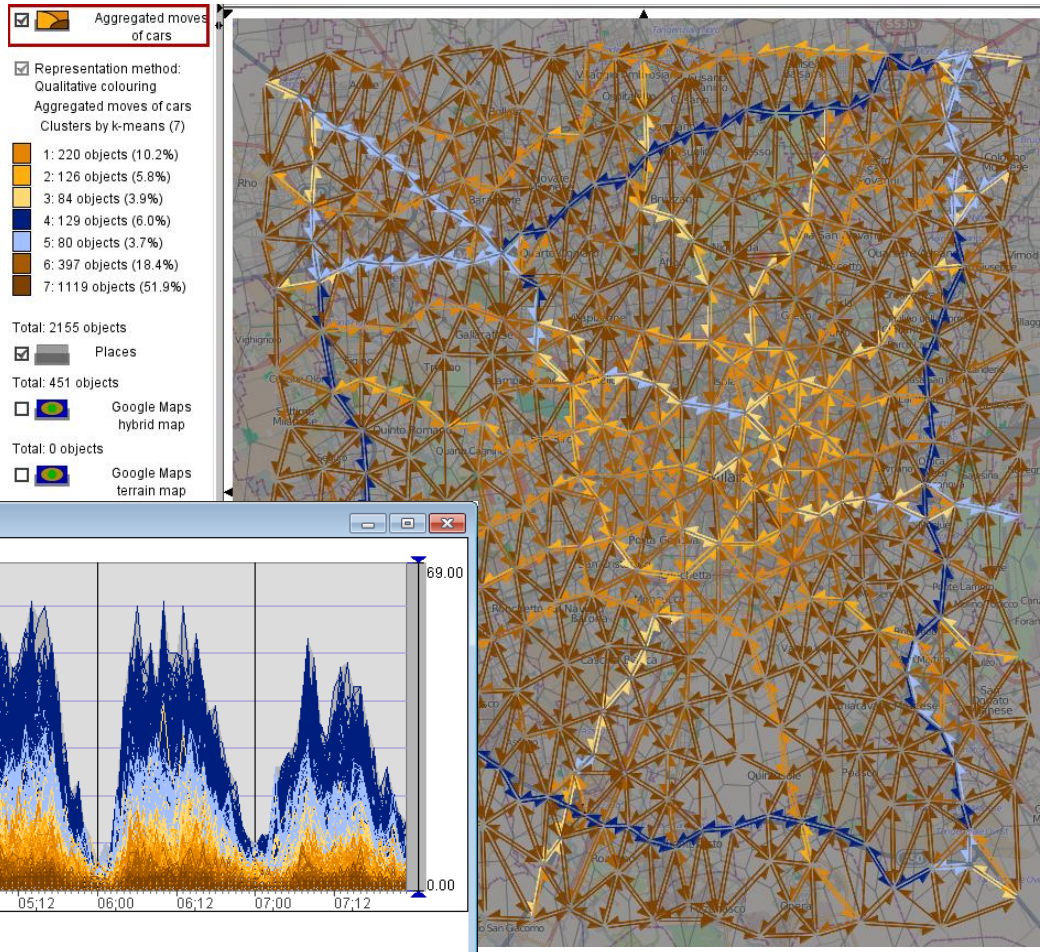


Part 1. Prediction of regular traffic flows

1) Partition-based clustering of the links

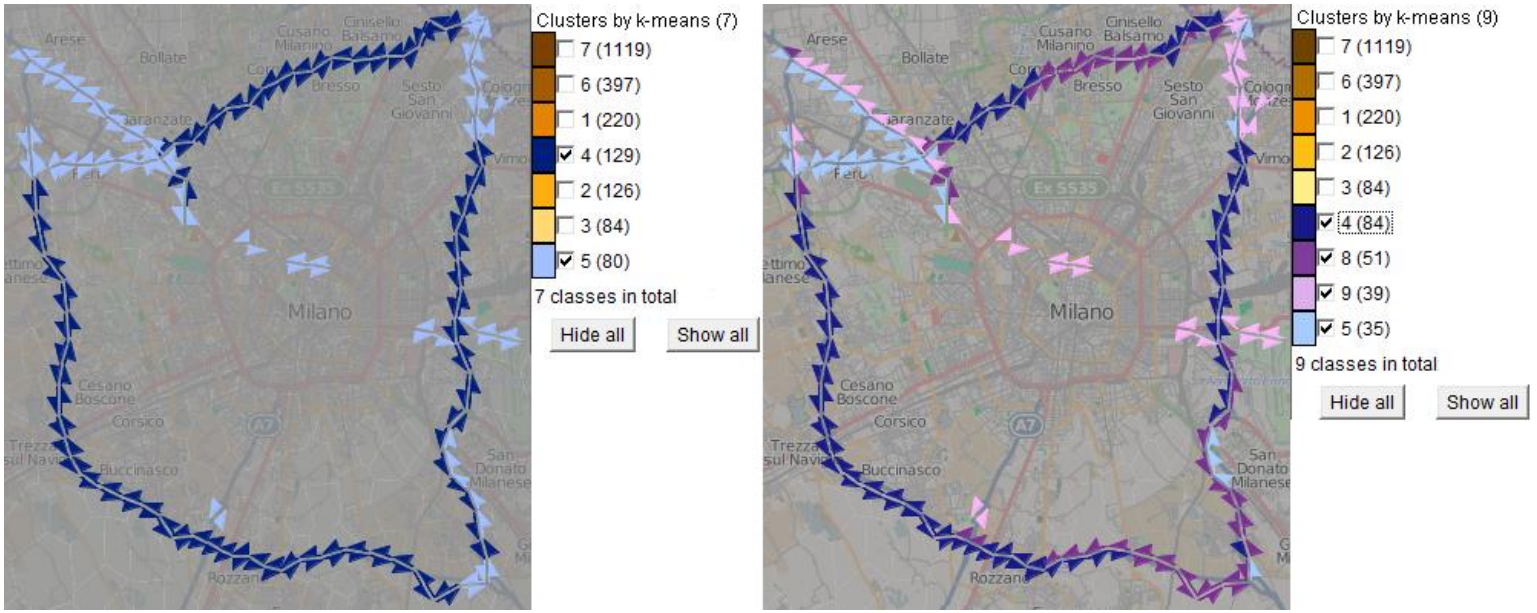
by similarity of the TS of the hourly move counts

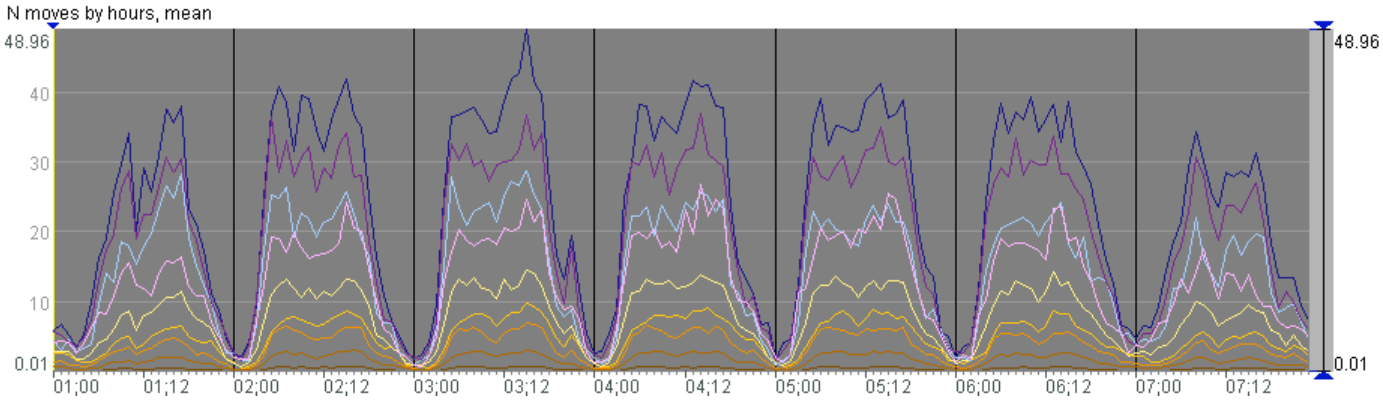
- Clustering method: k-means
- Tried different k from 5 to 15
- Immediate visual response facilitates choosing the most suitable k (i.e., giving interpretable and clear results).



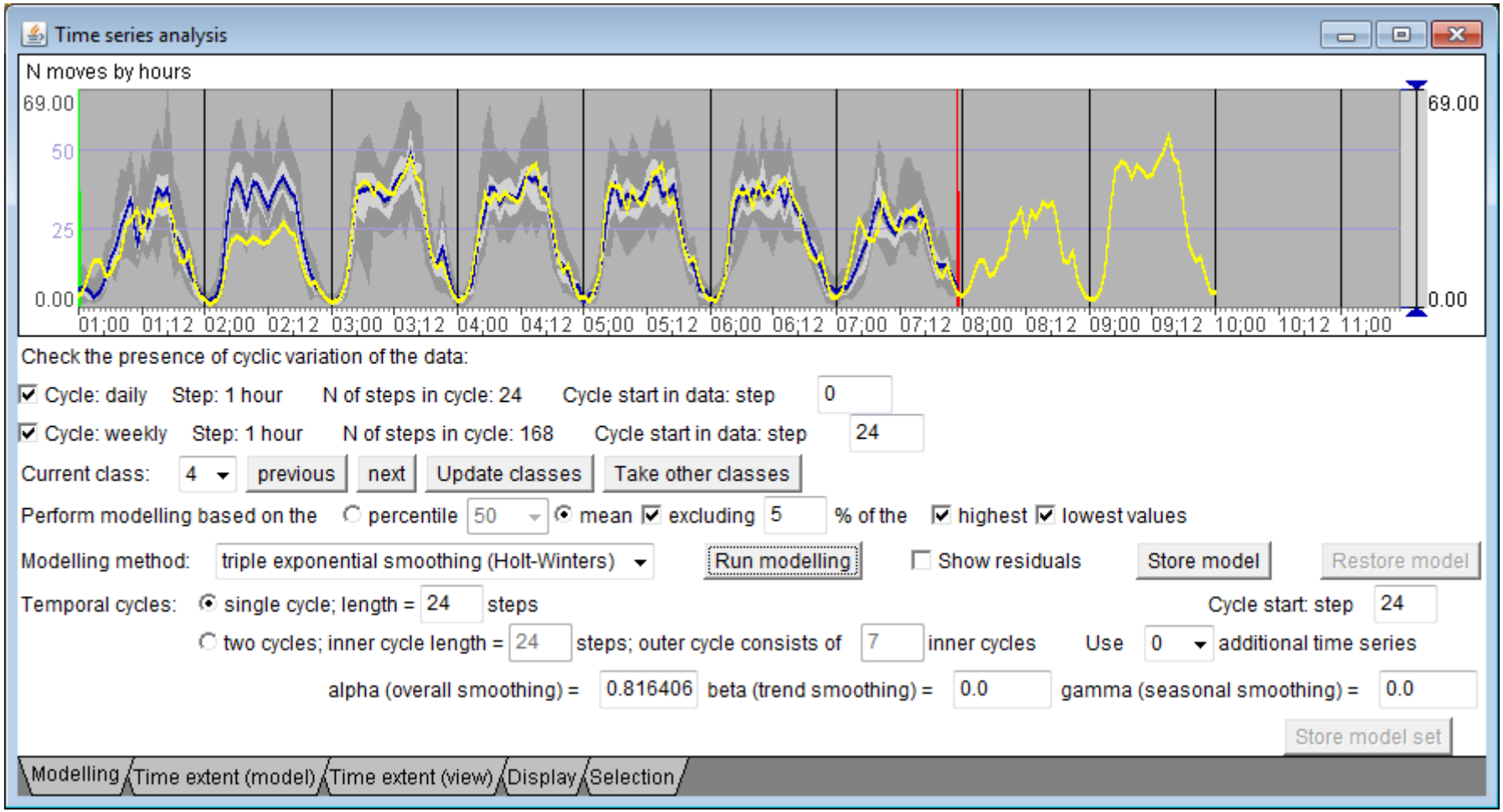
1.a) Re-grouping by progressive clustering

for reducing internal variation in clusters

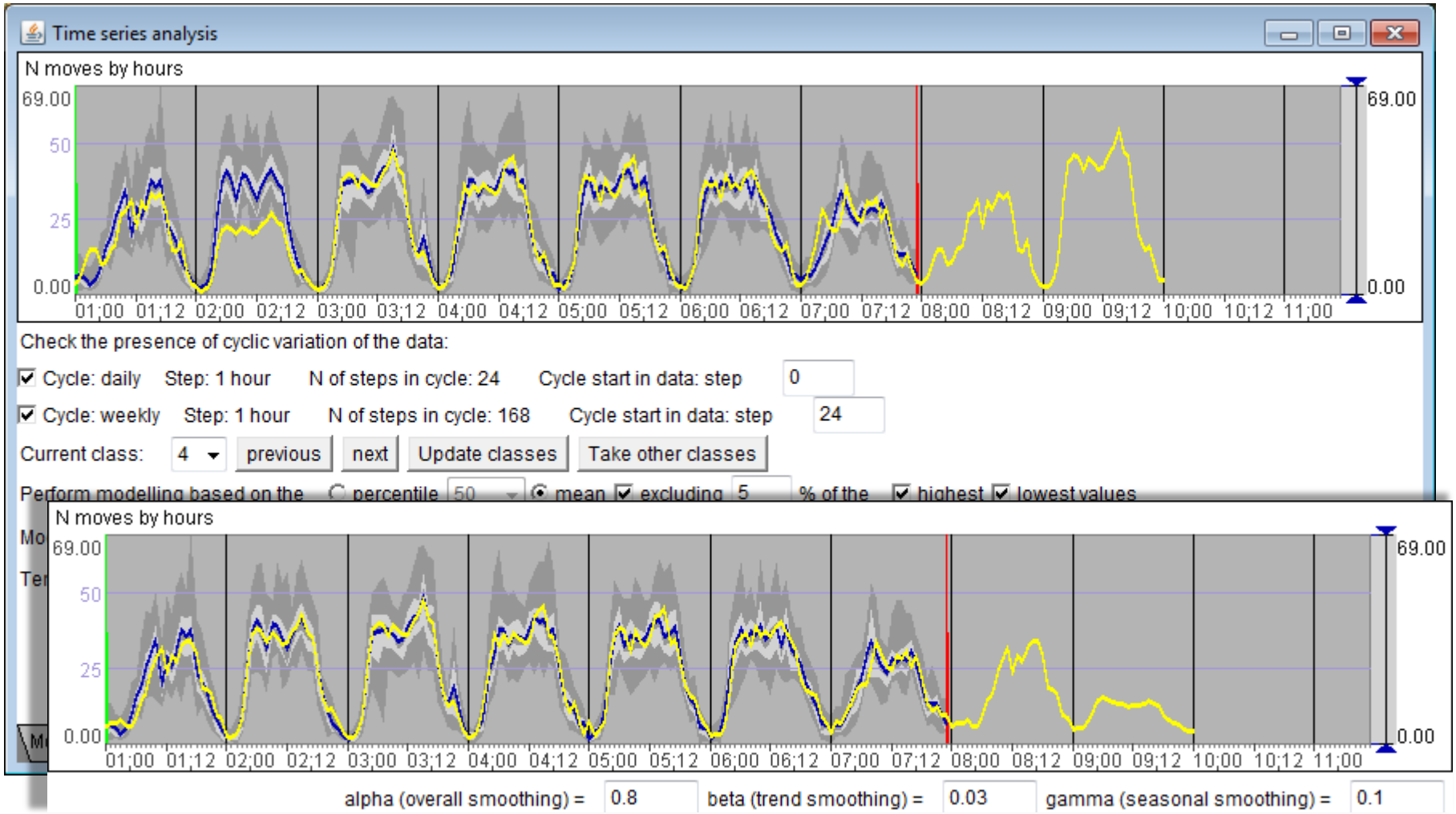




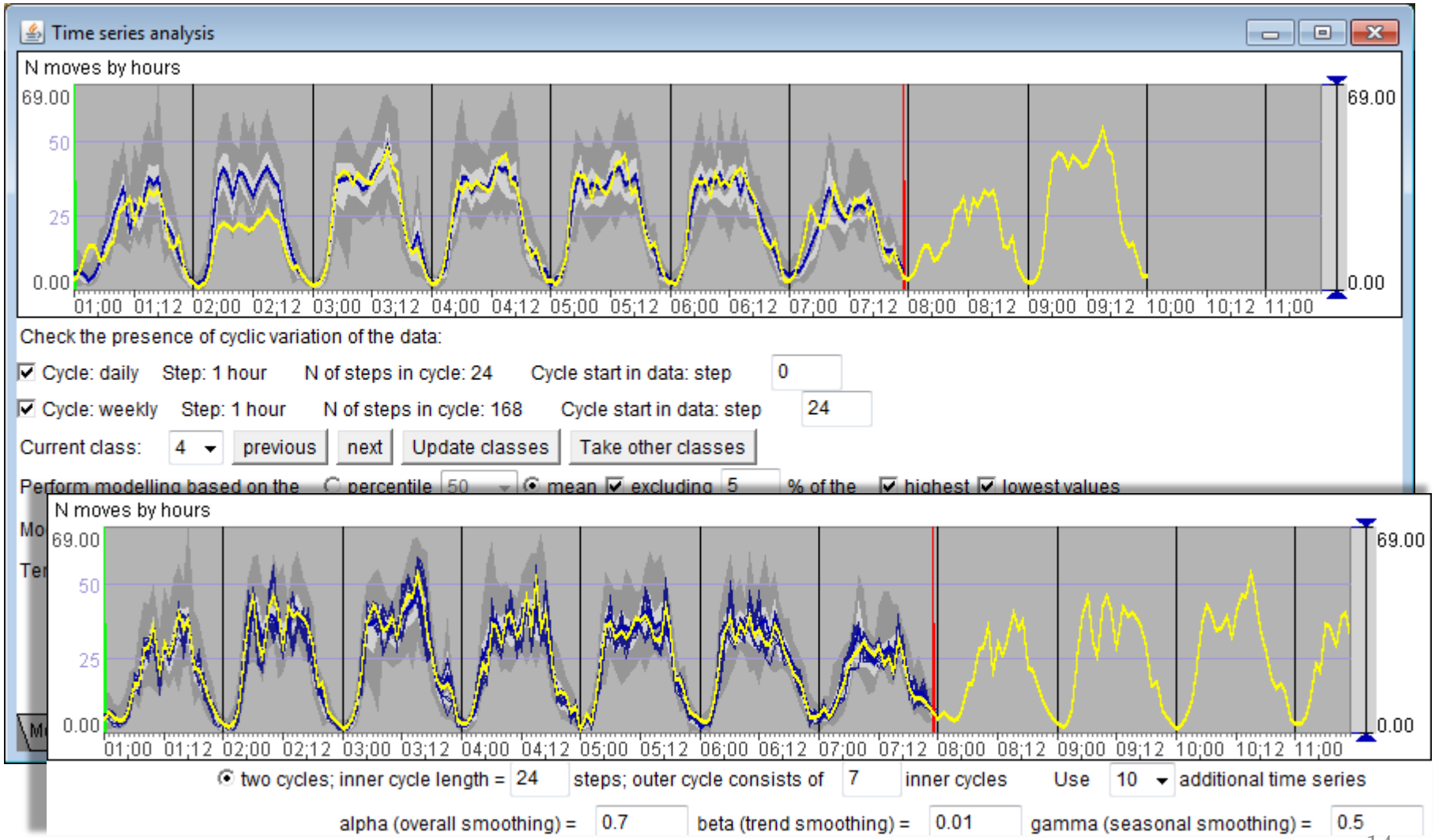
2) Cluster-wise time series modelling



2) Cluster-wise time series modelling



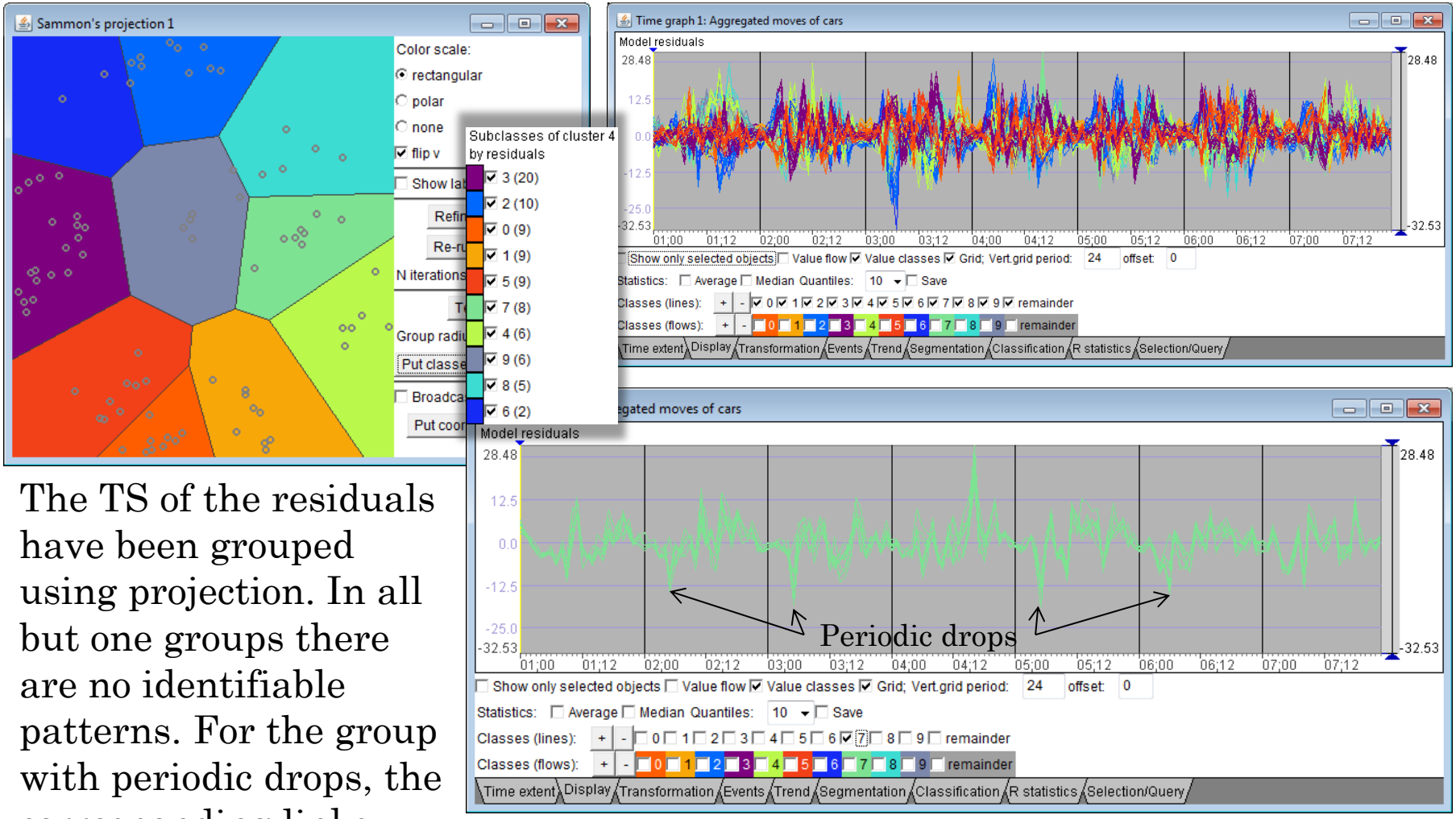
2) Cluster-wise time series modelling



3) Model evaluation (analysis of residuals)

- The goal is not to minimise the residuals
  - The model should not reproduce all fluctuations and outliers present in the data
  - This should be an abstraction capturing the characteristic features of the temporal variation
  - High values of the residuals do not mean low model quality
- The goal is to have the residuals randomly distributed in space and time (no detectable patterns)
  - This means that the model correctly captures the characteristic, non-random features of the temporal variation

Visual analysis of residuals



The TS of the residuals have been grouped using projection. In all but one groups there are no identifiable patterns. For the group with periodic drops, the corresponding links need to be considered separately ⇒ return back to the link re-grouping stage.

4) Use of the TS models for prediction of regular traffic

- After obtaining good models for all link clusters (possibly, after subdividing some of them based on the residual analysis), the models can be used for predicting the expected car flows in different times throughout the week.
  - The model capture the periodic (daily and weekly) variation of the traffic properties.
  - The variation pattern is expected to regularly repeat each week.
- However, each model as such gives the same prediction for all cluster members.
  - Although it would be technically possible to build an individual model for each link, such a model would be over-fitted (i.e., representing in detail fluctuations rather than capturing the general pattern). The cluster-wise modelling provides appropriate abstraction and generalisation.

$\Rightarrow$ The prediction needs to be individually adjusted for the members.

Adjustment of model predictions

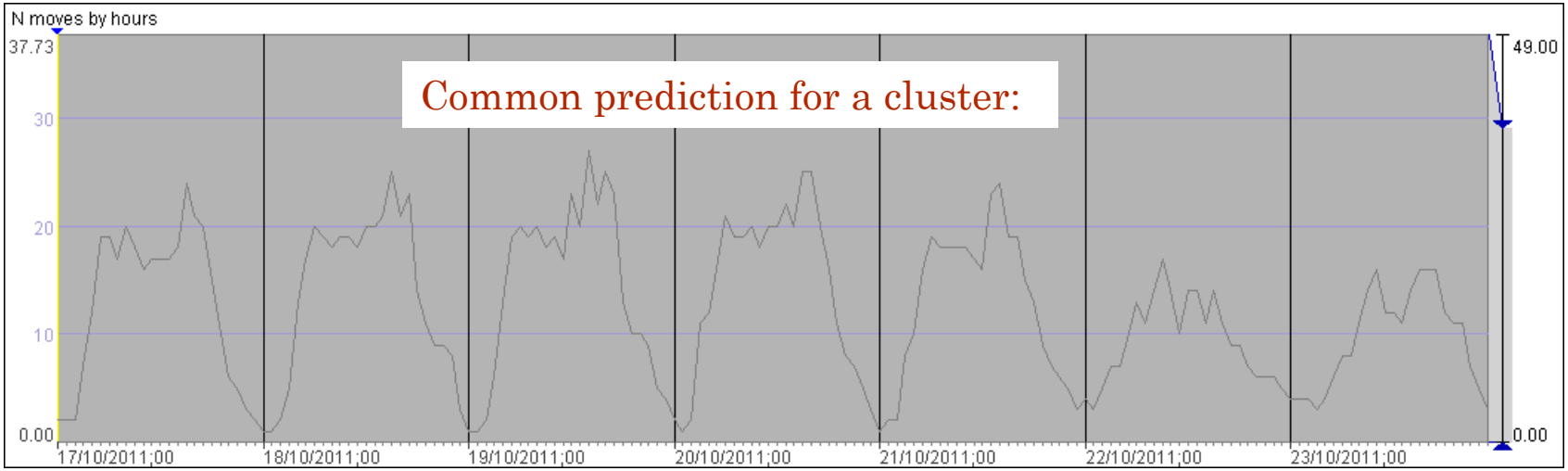
- For each link i, compute and store the basic statistics (quartiles) of the original values: $Q1_i$, M_i, $Q3_i$ (1st quartile, median, 3rd quartile)
- Compute the basic statistics of the model predictions for the whole cluster: Q1, M, Q3 (common for all cluster members)
- Shift *(level adjustment)*: $S_i = M_i - M$
- Scale factors *(amplitude adjustment)*:

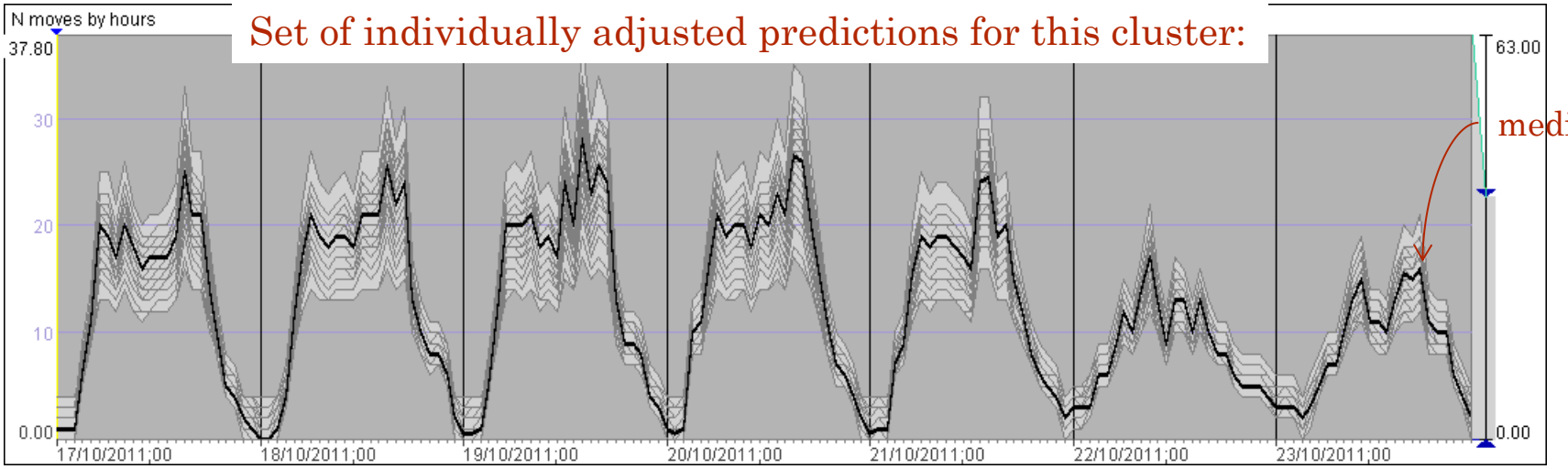
$$F_i^{low} = \frac{M_i - Q1_i}{M - Q1} \qquad F_i^{high} = \frac{Q3_i - M_i}{Q3 - M}$$

- For time step t, given a predicted value v^t (common for the cluster), the individually adjusted value for link i is

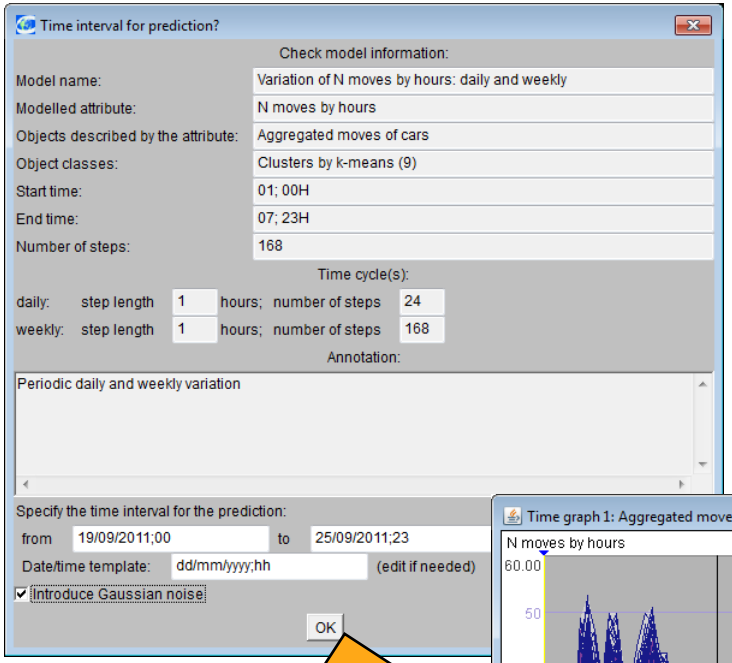
$$v^t_i = \begin{cases} M + F_i^{low} \cdot (v^t - M) + S_i, \text{ if } v^t < M \\ M + F_i^{high} \cdot (v^t - M) + S_i, \text{ otherwise} \end{cases}$$

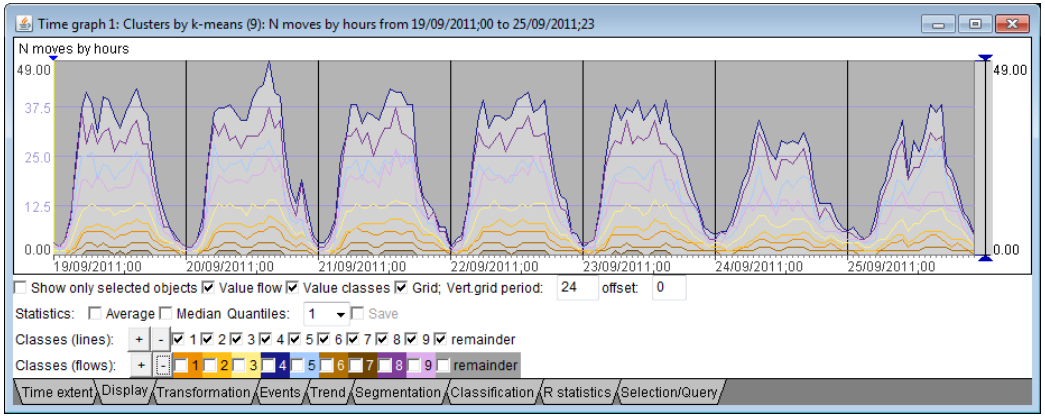
Example of individual adjustment

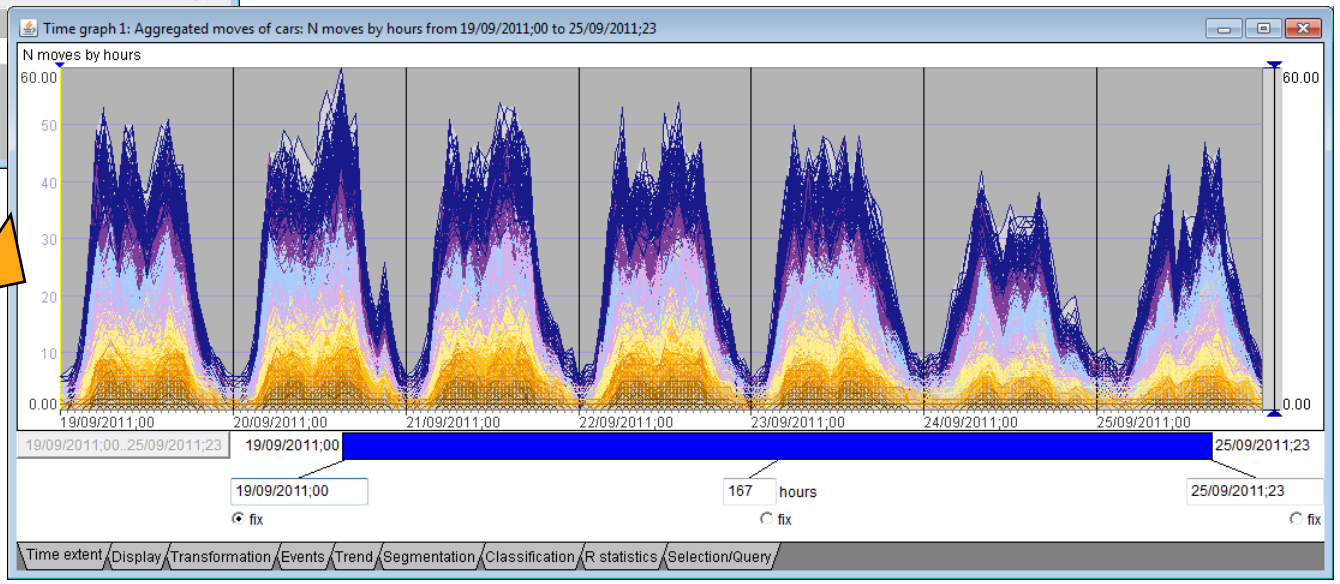




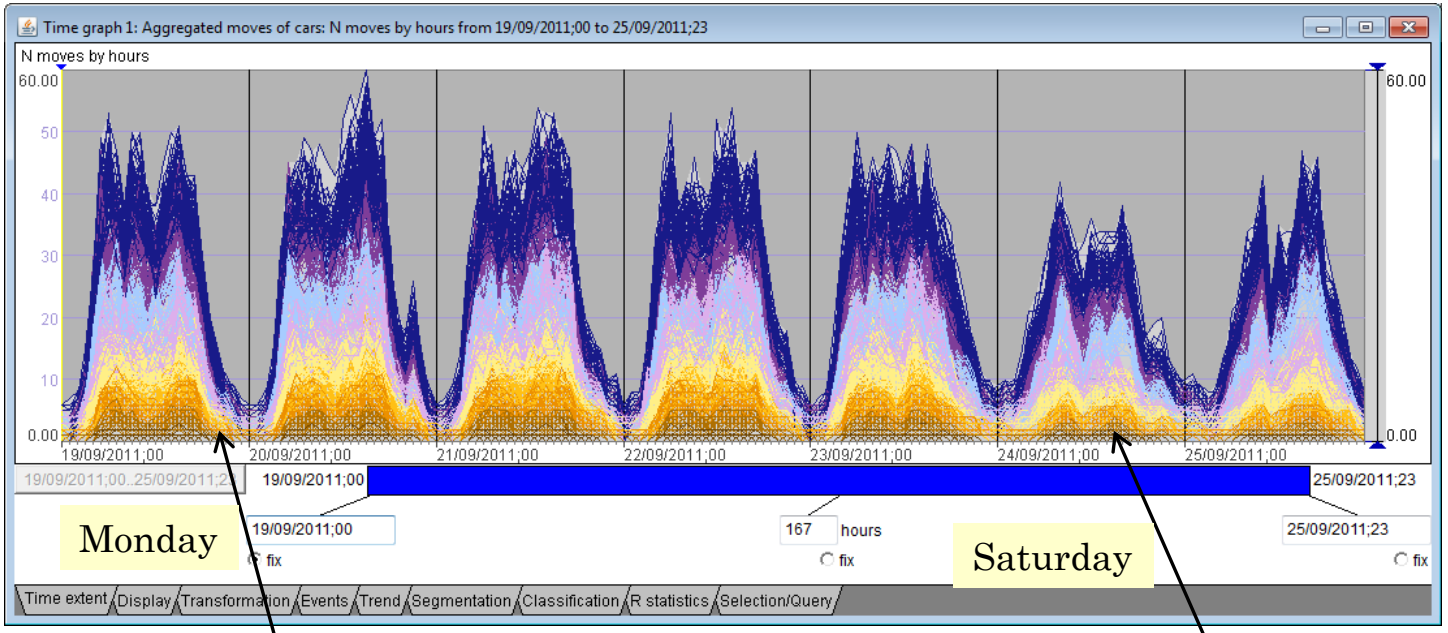
Example of prediction



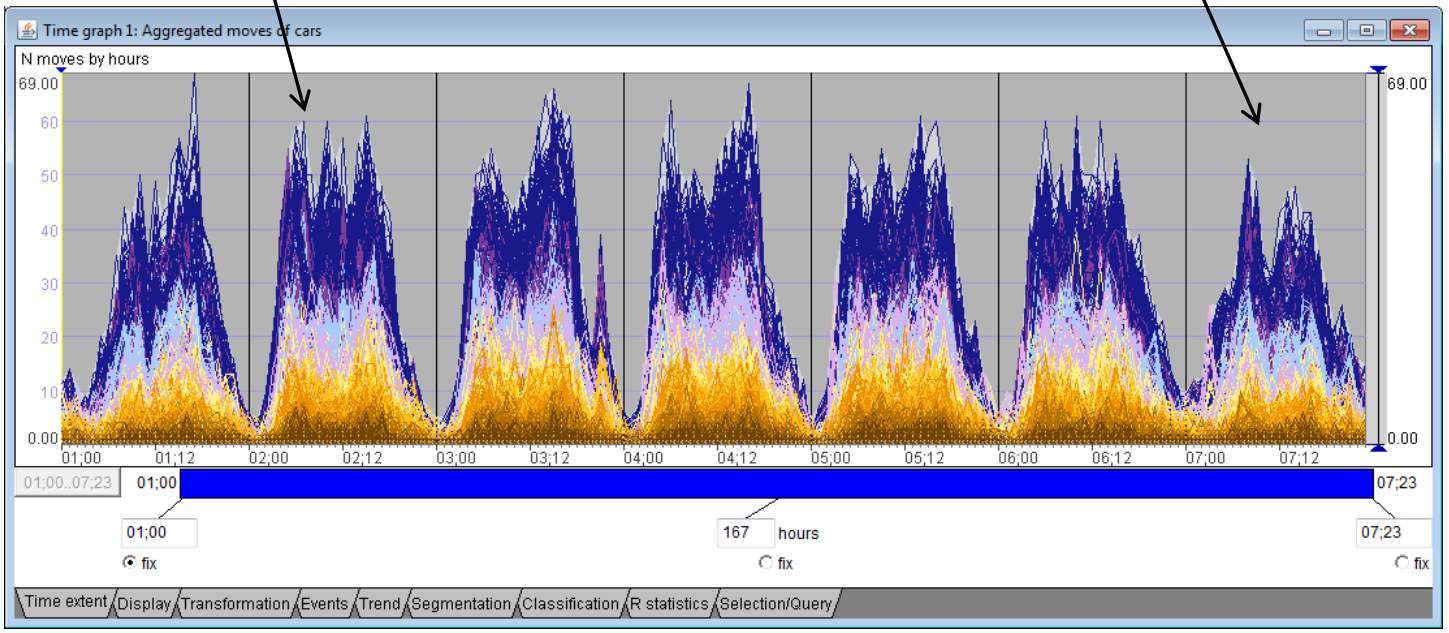




Predicted:



Original:



Where to read more

Natalia Andrienko, Gennady Andrienko

A Visual Analytics Framework for Spatio-temporal Analysis and Modelling

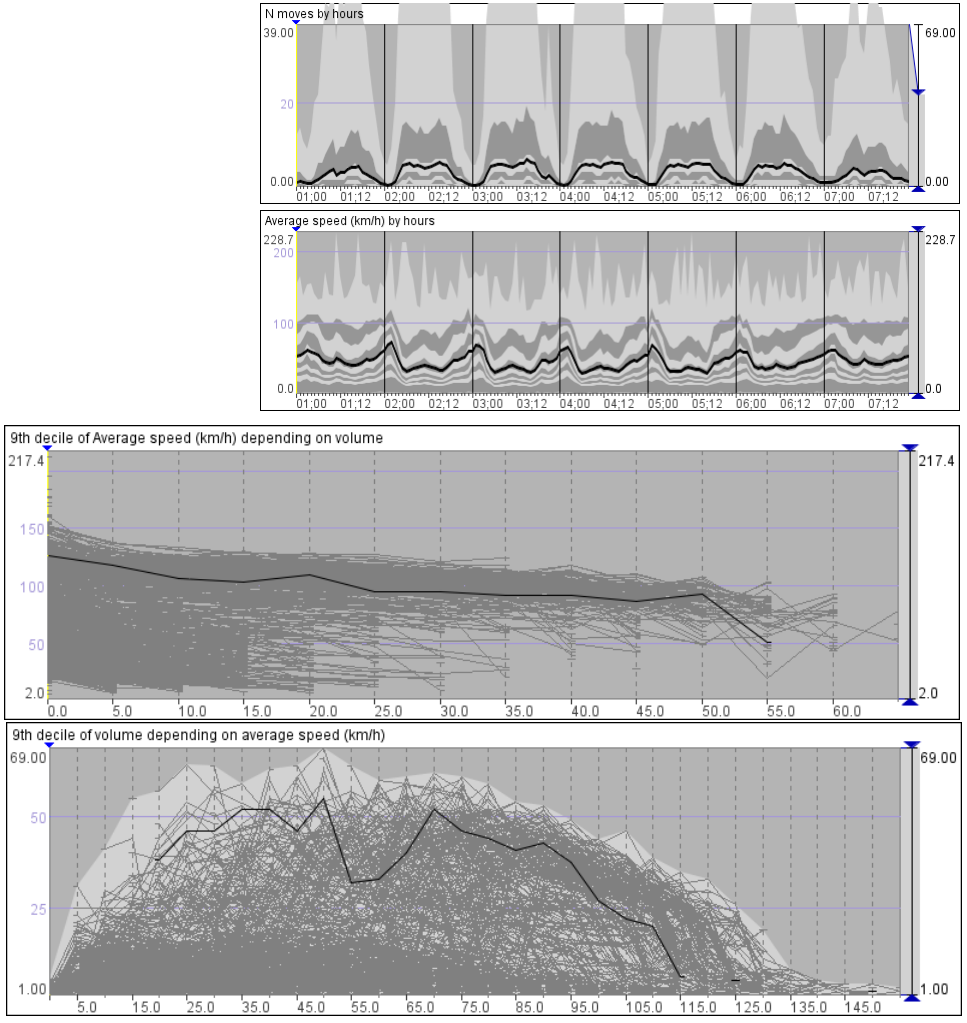
Data Mining and Knowledge Discovery, vol. 27(1), pp.55-83, 2013

http://dx.doi.org/10.1007/s10618-012-0285-7

Prediction of extraordinary traffic flows

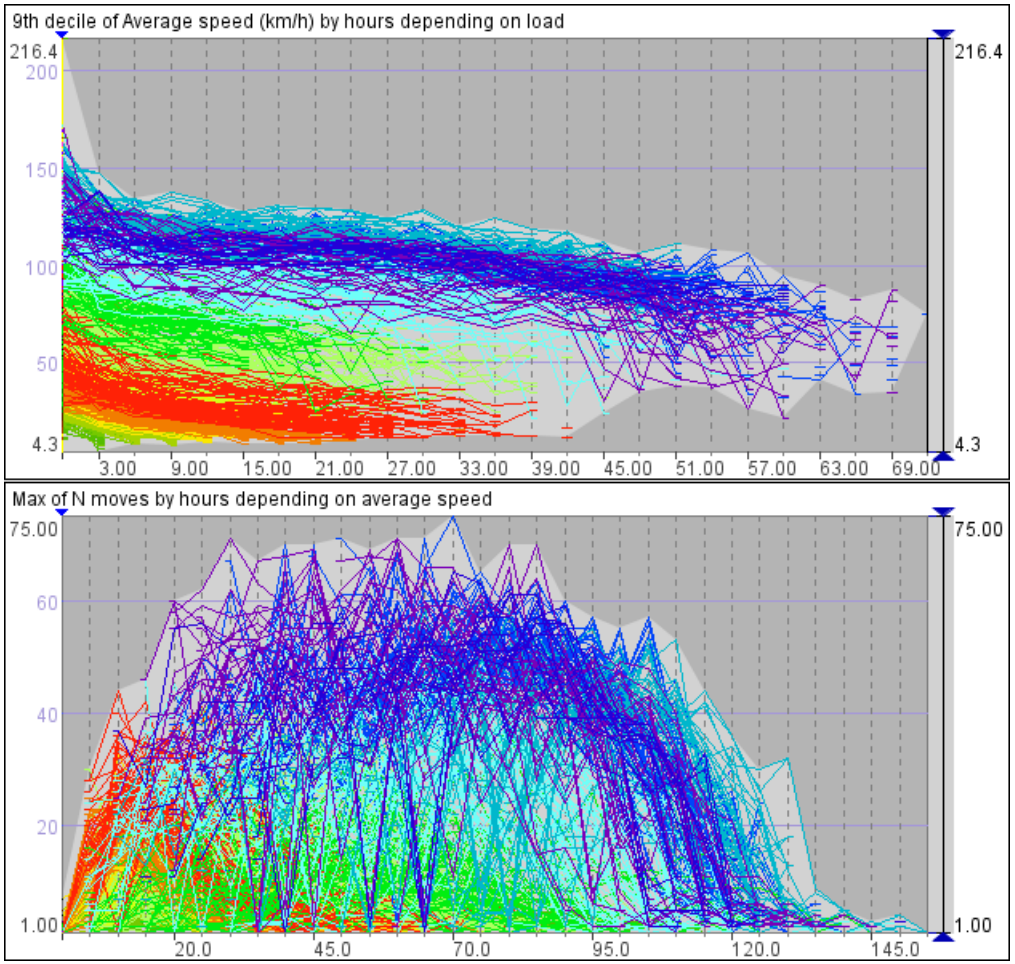
Volume-speed interdependencies

- The general interdependencies between the traffic volume and mean speed can be observed from the displays of the time series.
- If we explicitly capture the interdependencies and represent them by models, we will be able to predict the traffic dynamics under usual and unusual conditions.

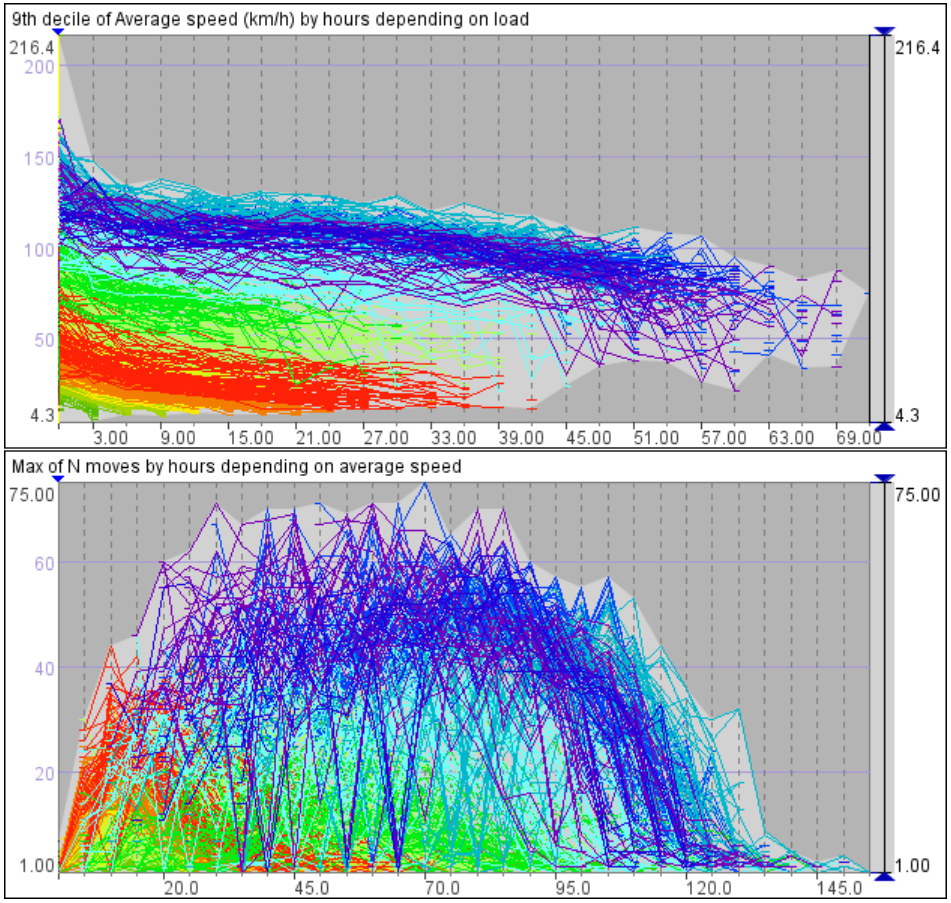


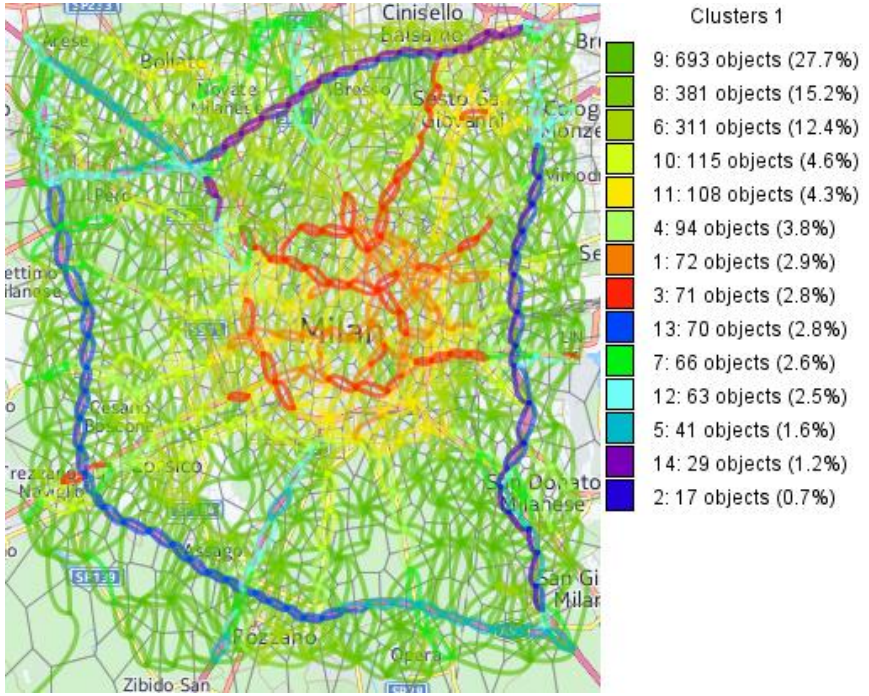
1) Data transformation

- Dependency of attribute A(t) on attribute B(t):
  - Divide the value range of B into intervals
  - For each interval, collect all values of A that co-occur with the values of B from this interval
  - Compute statistics of the values of A: minimum, maximum, median, mean, percentiles …
  - For each of these, there is a series B → A, or A(B)

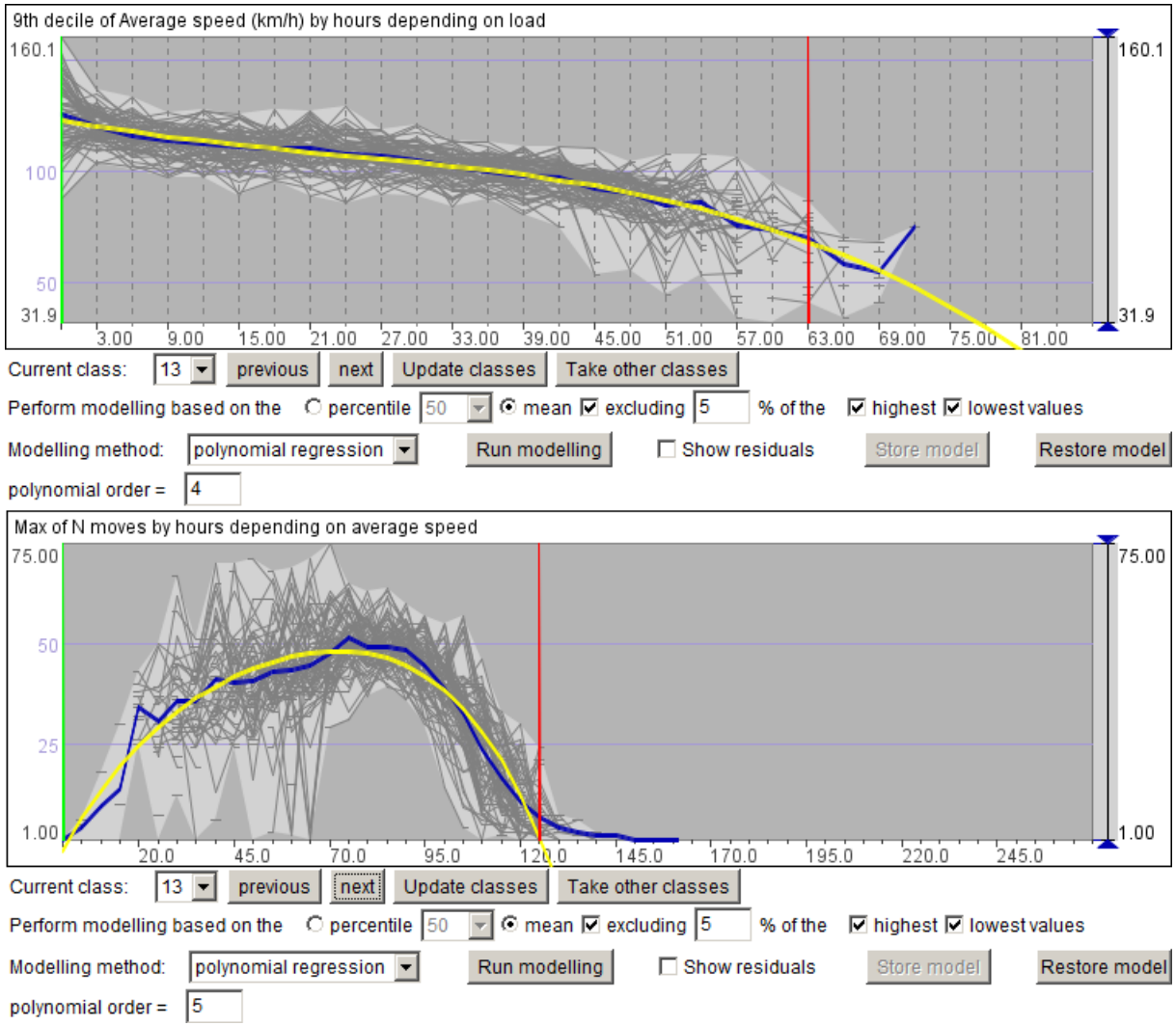


2) Partition-based clustering of the links by the similarity of the speed-volume dependencies



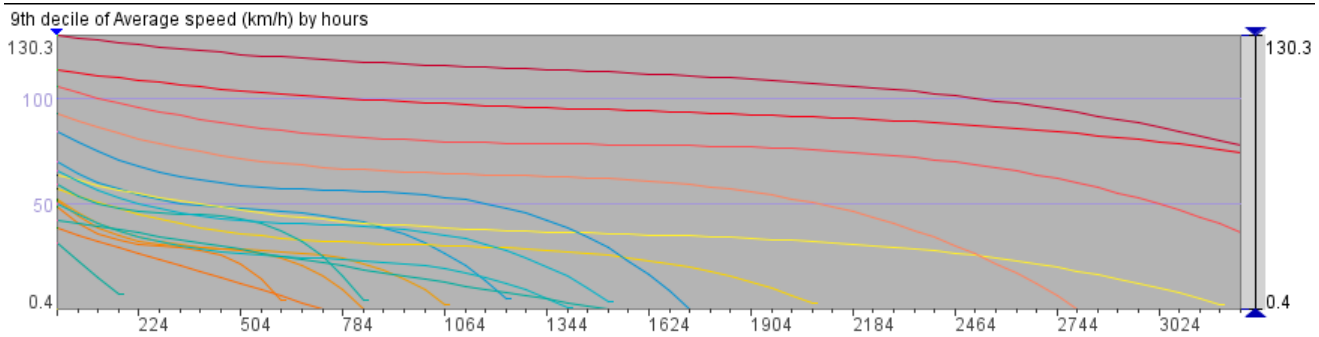


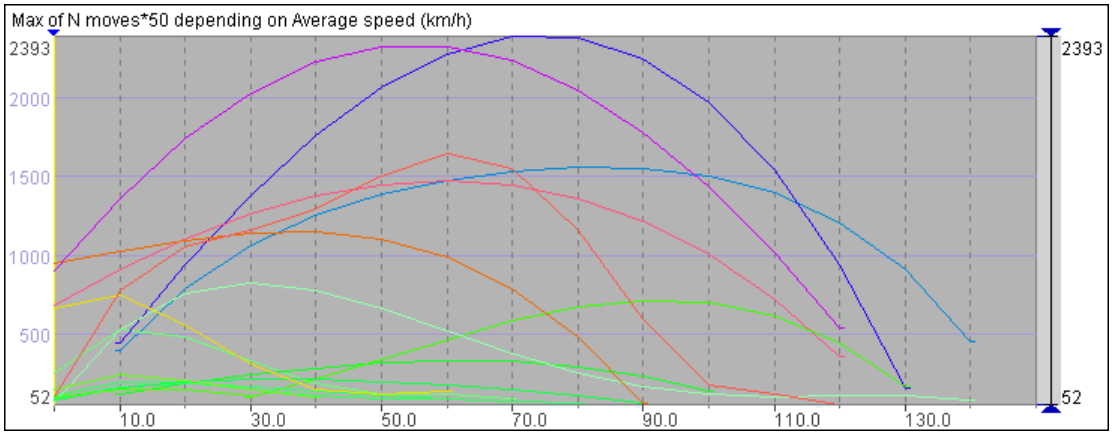
3) Representing the interdependencies by formal models



Models of the dependencies are built similarly to the time series modelling, but another modelling method is chosen: polynomial regression instead of double exponential smoothing. As previously, models are built for link clusters rather than individual links, to reduce the workload, minimise the impact of outliers, and avoid over-fitting.

Models built for scaled* data





* The original dataset does not contain the trajectories of all cars that moved over Milan but contains only trajectories of a sample of the cars. The sample size is estimated to be about 2% of the total number of cars.

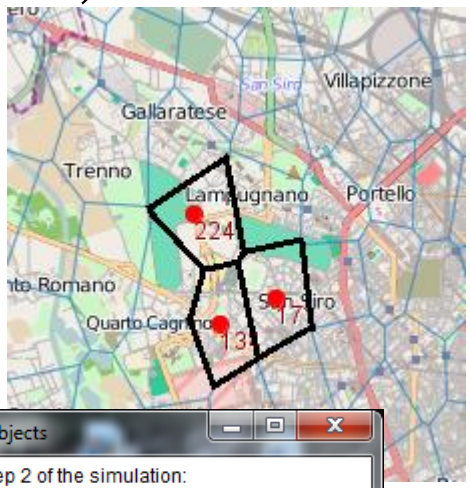
⇒ The aggregation of the original dataset does not give the true flow volumes for the links but about 2% of the true volumes. To obtain more realistic flow volumes, the computed volumes need to be multiplied by 50.**

** When additional data are available, such as traffic volumes measured by traffic counters in different places, scaling may be done in a more sophisticated and more accurate way.

4) Forecasting unusual traffic (traffic simulation)

- General idea of the simulation method:
  - For each link P → Q, determine the number of vehicles that wish to move from P to Q in the current minute.
  - Determine the possible speed of these vehicles (model volume → speed).
  - Determine the number of vehicles that will be actually able to move from P to Q with this speed (model speed → volume).
  - Promote this number of vehicles from P to Q; suspend the remaining vehicles in P.
- An interactive visual interface supports defining simulation scenarios, "what if" analysis, and comparison of results of different simulations.

Example: simulation of movement of 10,000 cars from around San Siro stadium



Set prediction models

The simulation requires the following prediction models:

1. (Place_1, Place_2, Time) -> N of cars

A set of time series models predicting the regular number of moves (flow) from one place to another by time intervals.

Variation of N moves by hours *50: daily and weekly

Select from available models

2. (Place_1, Place_2, N of cars) -> Possible speed

2) A set of dependency models predicting the maximal average speed of moving from one place to another depending on the place link load, i.e., number of cars that try to move.

Variation of Max of Average speed (km/h) depending on N moves

Select from available models

3. (Place_1, Place_2, Possible speed) -> N of cars

A set of dependency models predicting the maximal number of cars (flow) that will be able to move from one place to another within a given time interval depending on the maximal average speed with which the cars can move.

Variation of Max of N moves*50 depending on Average speed (km/h)

Select from available models

Scale factor for the model-predicted values: 1.0

Done Cancel

Transition times?

Select the attribute defining the transition times.

Start ID
End ID
N of moves
Length
Average move duration (minutes); total
Average speed (km/h); total
Average path length; km
Average path length ratio to link length
N trajectories; total *50
N moves; total *50

Use the weights of the links defined by the attribu

Length
Average move duration (minutes); total
Average speed (km/h); total
Average path length; km
Average path length ratio to link length
N trajectories; total *50
N moves; total *50
Average N moves by hours *50
Median of N moves by hours *50
Max N moves by hours *50

OK Car

Distribute moving objects

Step 2 of the simulation:

Distribute moving objects among the destinations and routes

A given number of moving objects will be distributed among the possible destinations, i.e., places from the layer Places.

The places need to have weights defined by some numeric attribute.

Select the attribute defining the weights:

N visits
N starts
N ends
N visitors total
N visits total
N ends after 18:00

The number of moving objects in the selected place(s) of origin:

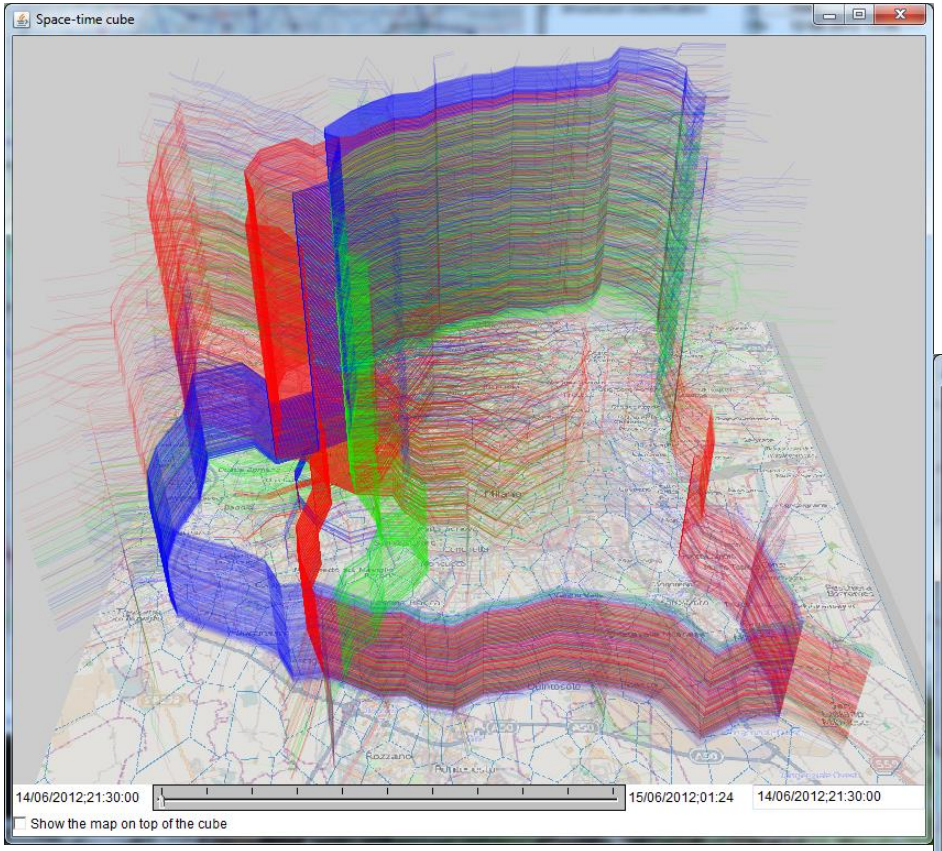
In place 171: 3000
In place 134: 4000
In place 224: 3000

Localize the places on map

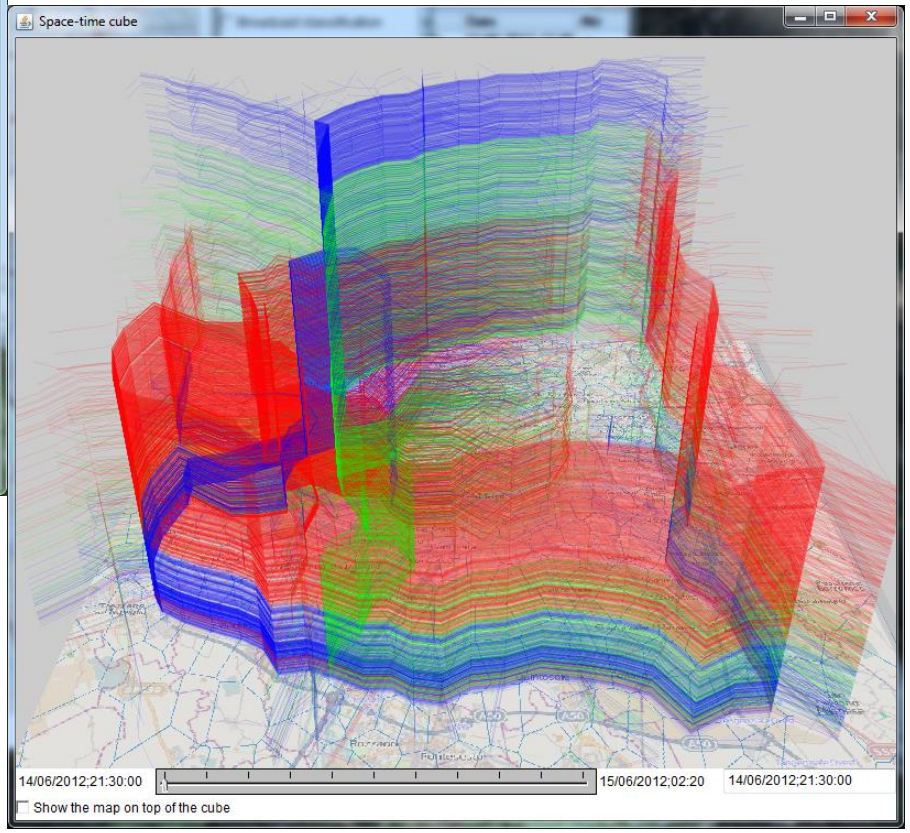
The given number of objects will be distributed among the 3 selected places of origin.

Continue Stop the process

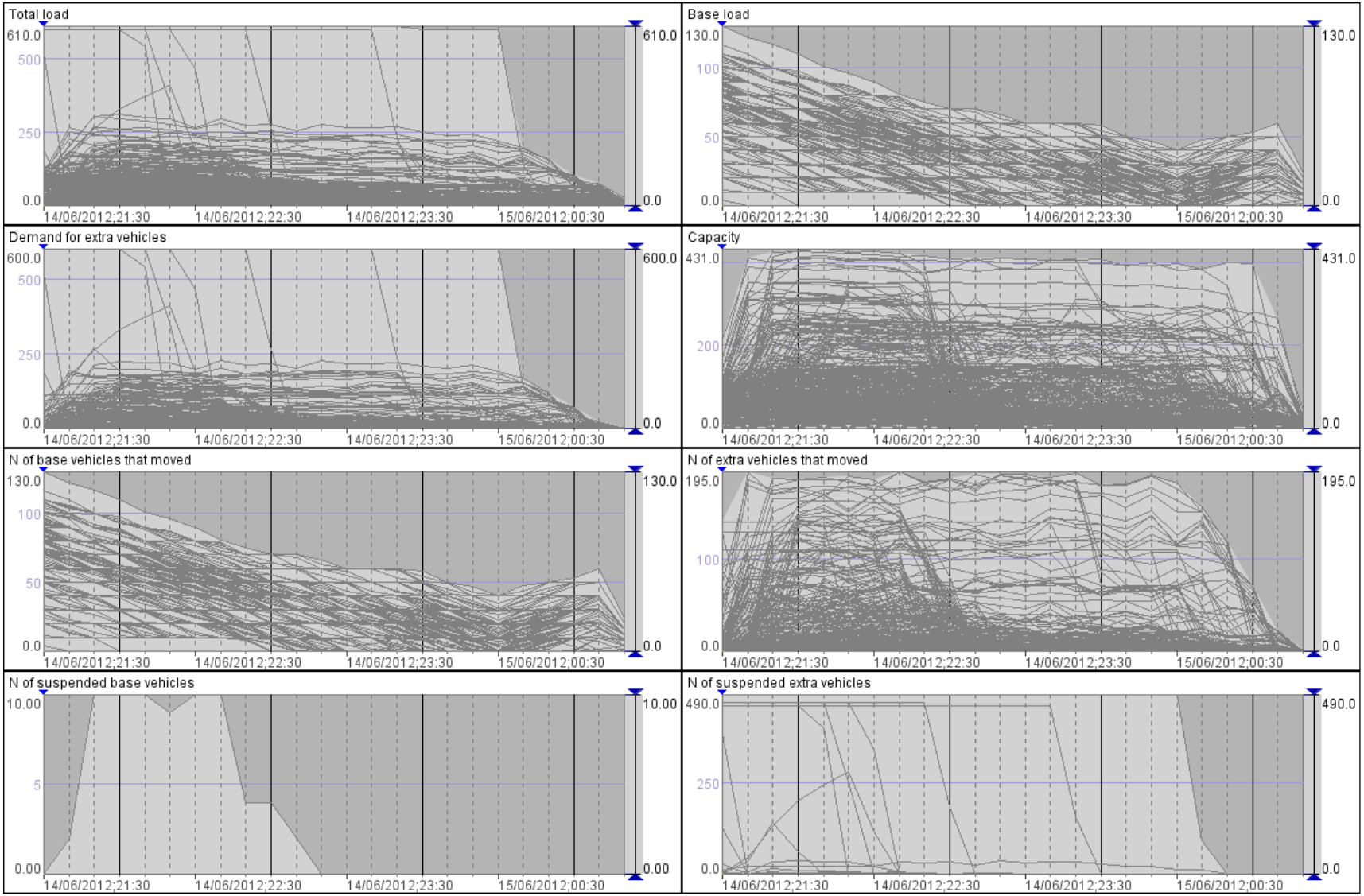
Simulated trajectories



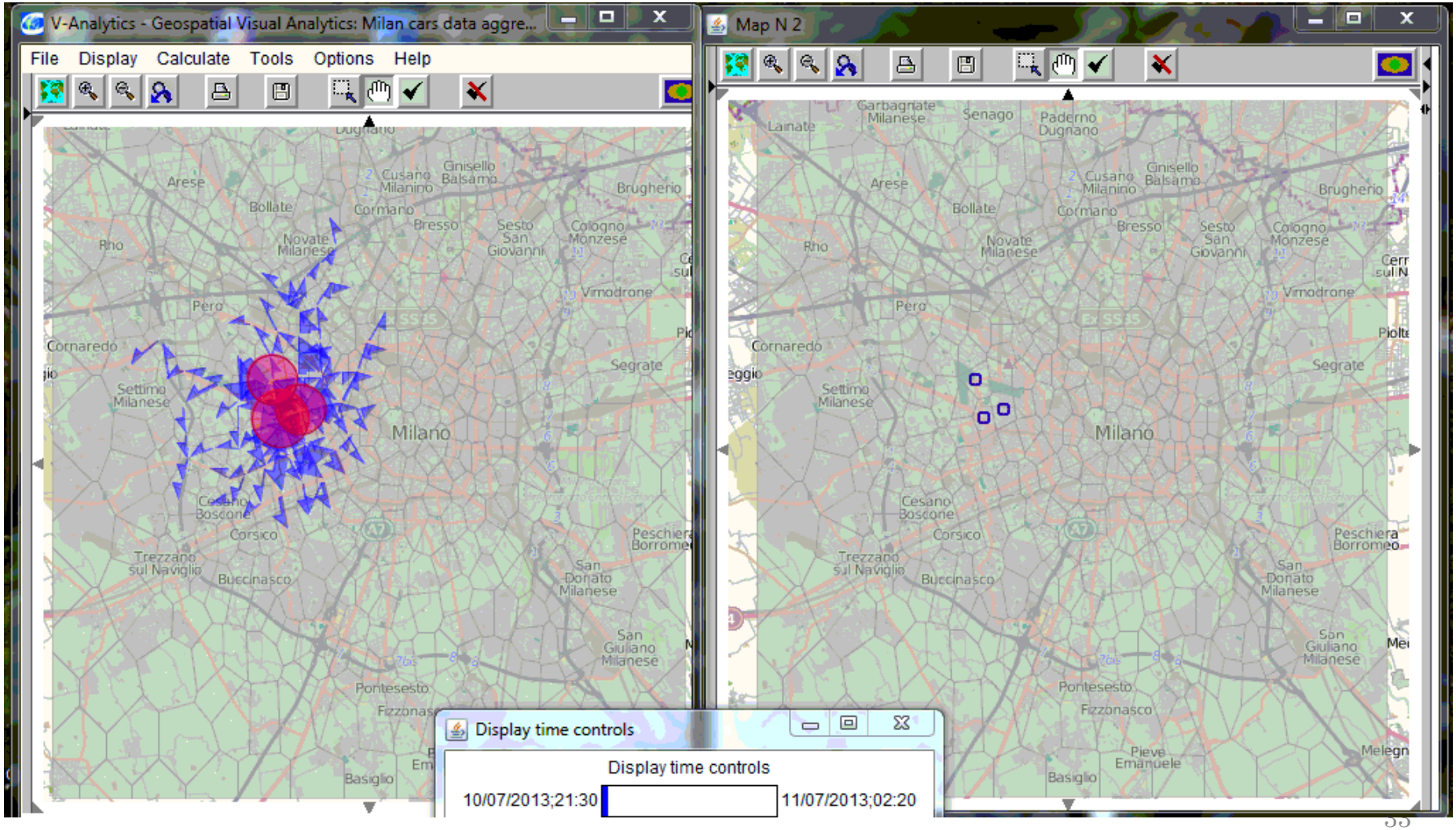
What will be the effect of re-routing a part of the traffic to the south?



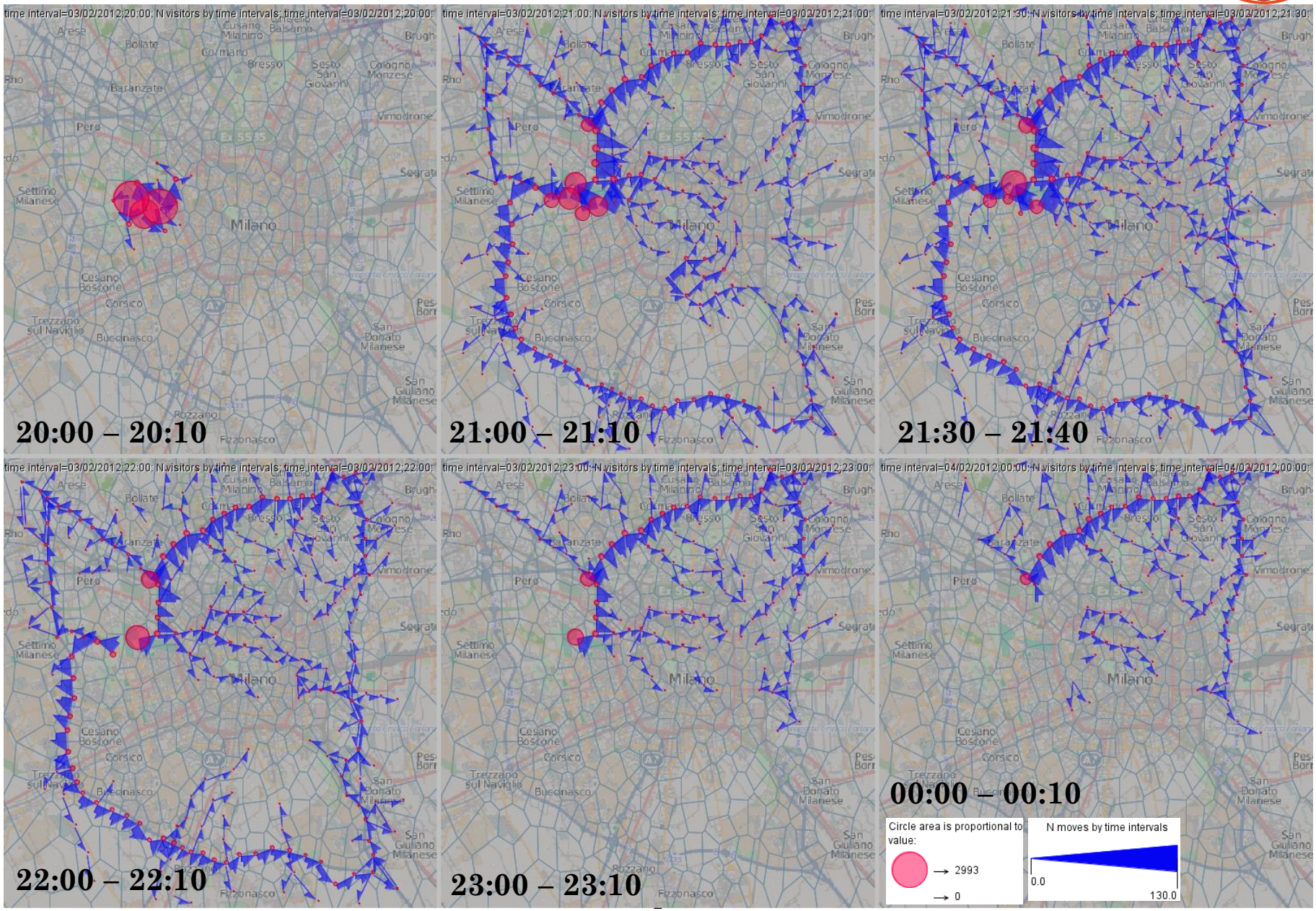
Aggregated representation of simulation results: time graphs



Aggregated representation of simulation results: map animation



Presence and flows for selected time intervals



Implementation of the principles of predictive visual analytics

- conscious preparation of data (cleaning, transforming, partitioning, …)
  - ST aggregation, expression of interdependencies, clustering
- conscious decomposition of the modelling task
  - cluster-wise model building
- conscious selection of variables, modelling methods, and parameters; creation and comparison of model variants
  - interactive visual interface for trying different methods and parameter settings
- conscious evaluation of model quality
  - interactive visual exploration of model residuals
- conscious refinement of models
  - further decomposition through progressive clustering or interactive division

Where to read more

Natalia Andrienko, Gennady Andrienko, and Salvatore Rinzivillo

Leveraging Spatial Abstraction in Traffic Analysis and Forecasting with Visual Analytics

Information Visualization, vol. 15(2), pp.117-153, 2016

http://dx.doi.org/10.1016/j.is.2015.08.007

Visual analytics support to predictive modelling

- Various VA research prototypes support the process of building predictive models in a way adhering to the main principles.
  - Each system is oriented to a particular type of data and particular modelling method or class of methods.

☹When it comes to model building in practice, it may be hard to find a ready-to-use system providing suitable visual analytics support.

⇒Analysts should try to implement the main principles by themselves

- Use interactive visualisations to explore available data and decompose the modelling through data partitioning.
- Try different modelling tools, methods, and parameter settings.
- Use interactive visualisations to explore model predictions and errors and to find possibilities for model refinement (e.g., by further data cleaning or partitioning, choosing another method, modifying parameter settings, …).
- The process is iterative rather than sequential.

Selected papers on predictive VA

- Focus: classification models
  - M. Gleicher. "Explainers: Expert Explorations with Crafted Projections", *IEEE Trans. Visualization and Computer Graphics*, 19(12): 2042-2051, 2013
  - S. van den Elzen and J.J. van Wijk, "BaobabView: Interactive construction and analysis of decision trees", In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST'11)*, pp. 151-160, 2011.
- Focus: regression models
  - T. Mühlbacher and H. Piringer. "A Partition-Based Framework for Building and Validating Regression Models", *IEEE Trans. Visualization and Computer Graphics*, 19(12): 1962-1971, 2013.
  - Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski. "Integrating Predictive Analytics and Social Media". In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST'14)*, 2014.

Selected papers on predictive VA (continued)

- Focus: time series models
  - M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind, "Visual Analytics for Model Selection in Time Series Analysis", *IEEE Trans. Visualization and Computer Graphics*, 19(12): 2237-2246, 2013
  - M.C. Hao, H. Janetzko, S. Mittelstädt, W. Hill, U. Dayal, D.A. Keim, M. Marwah, and R.K. Sharma, "A Visual Analytics Approach for Peak-Preserving Prediction of Large Seasonal Time Series", *Computer Graphics Forum*, 30(3): 691-700, 2011
- Focus: support of forecasting by means of simulation models
  - S. Afzal, R. Maciejewski, and D.S. Ebert. "Visual analytics decision support environment for epidemic modeling and response evaluation". In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST'2011)*, pp. 191–200, 2011
  - H. Ribicic, J. Waser, R. Fuchs, G. Blöschl, E. Gröller, "Visual Analysis and Steering of Flooding Simulations", *IEEE Trans. Visualization and Computer Graphics*, 19(6): 1062-1075, 2013

Questions?

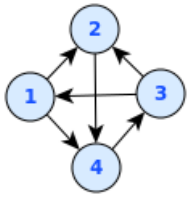
Predictive visual analytics

Visual analytics of other data types

Graphs (networks), texts, images, video

Data type: Graphs (networks)

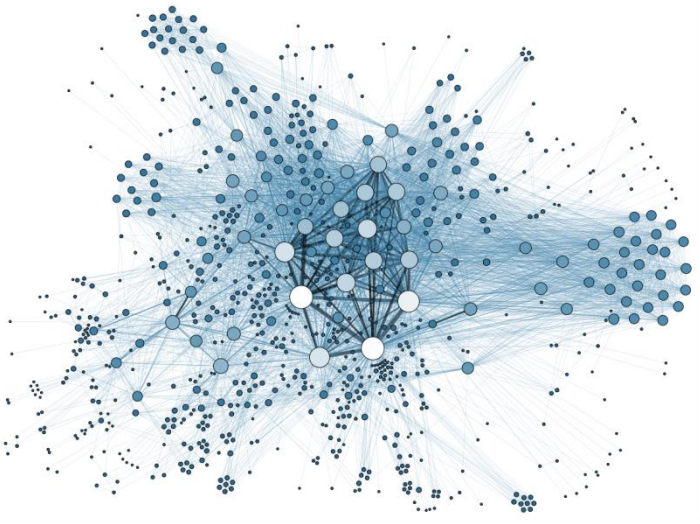
- Represent binary relationships between entities
  - Bonds between atoms in a molecule, friendship between people, …
- Consist of nodes (a.k.a. vertices) representing entities and links (a.k.a. edges) representing relationships.
  - Each link connects two nodes.
  - Links may be directed or undirected.
  - Links may have weights representing the strengths of the relationships.
  - Nodes and links may have various attributes.



- Graphs may evolve over time:
  - Nodes and/or links may appear and disappear; weights or other attributes may change
    - E.g., imagine the evolution of friendship relationships between people

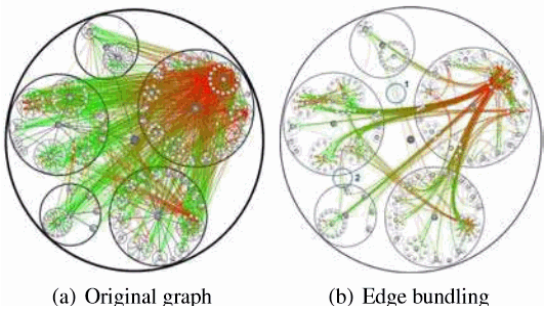
Graph-specific analysis tasks and problems

- Tasks:
  - Analyse the structure of a graph
  - Find particular structural patterns in a graph
  - Compare the structures of multiple graphs
  - Analyse the temporal evolution of a graph
- Problems:
  - Visualisation of graphs
    - In particular, intersecting links in node-link diagrams
  - Large graphs
  - Numerous graphs
  - Evolving graphs



Some approaches to supporting graph analysis

- Visual simplification:
  - Graph layouts that minimise crossings; edge bundling; matrix representation; combined node-link and matrix representation; …
  - Interactive filtering
  - Aggregation of nodes and/or links (interactive or automated)



(a) Original graph (b) Edge bundling

- Computational techniques:
  - Identify important nodes (e.g., hubs)
  - Find the shortest path between two nodes
  - Find specific substructures, e.g., highly connected communities
  - Group (cluster) the nodes according to the links (partition-based, density-based); aggregate the graph based on the grouping
  - Group (cluster) multiple graphs by similarity

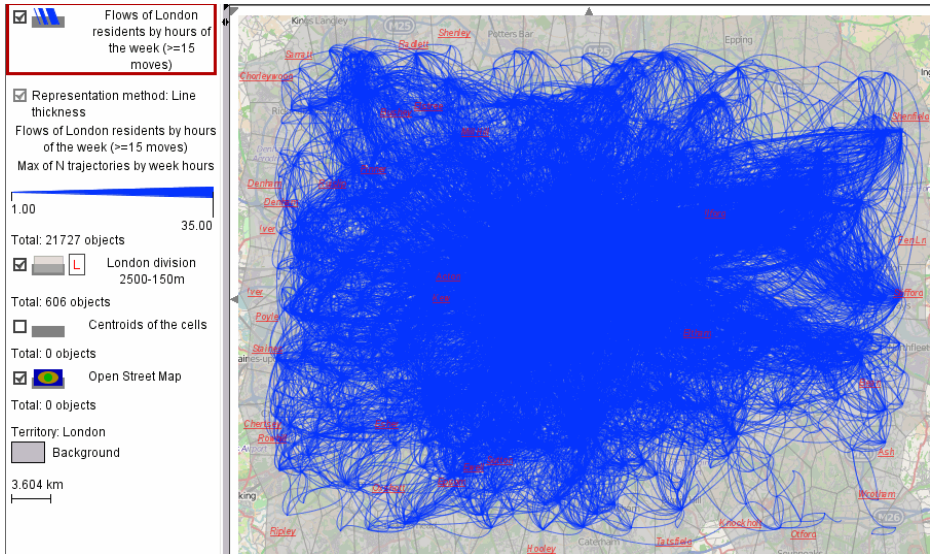
Recommended reading

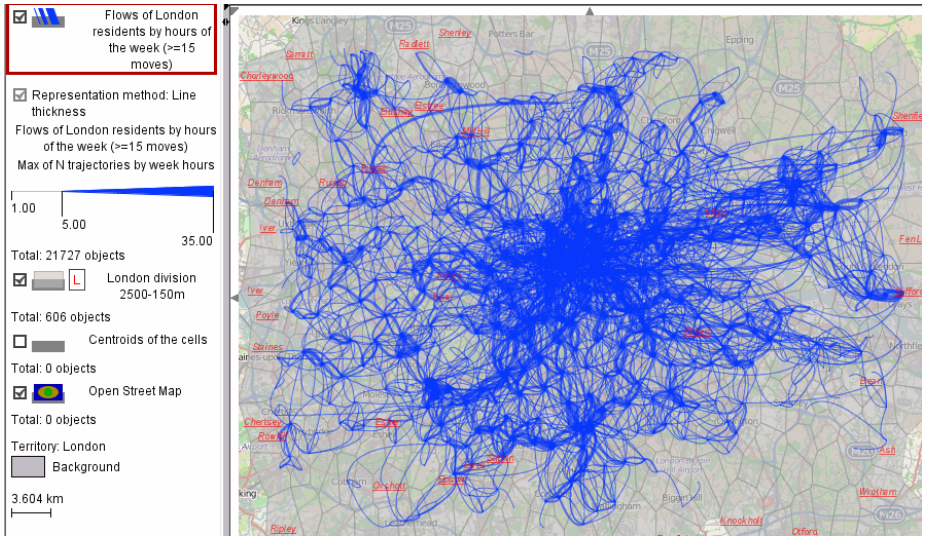
T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J.J. van Wijk, J.-D. Fekete, and D.W. Fellner

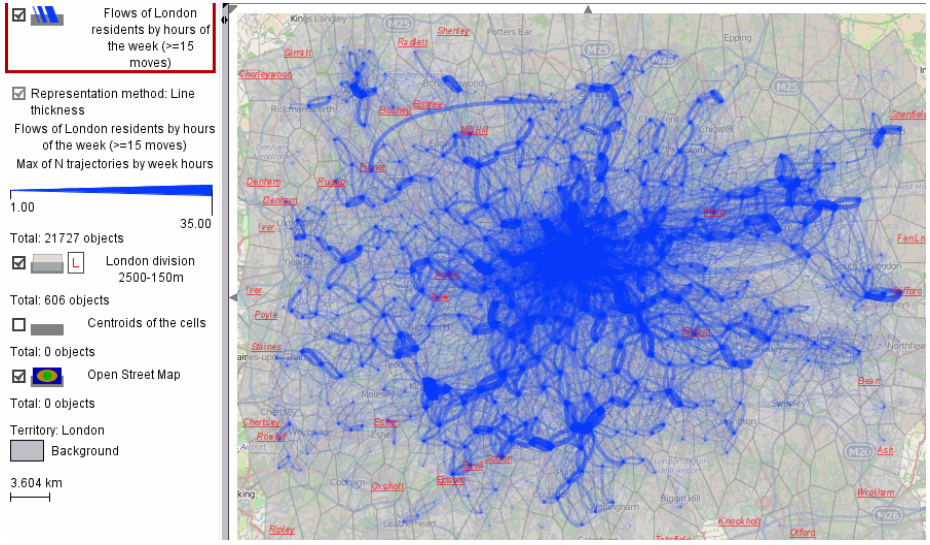
Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges

Computer Graphics Forum, Vol. 30 (6), pp. 1719–1749, September 2011

Application of graph visualisation and analysis techniques to spatial links (flows)



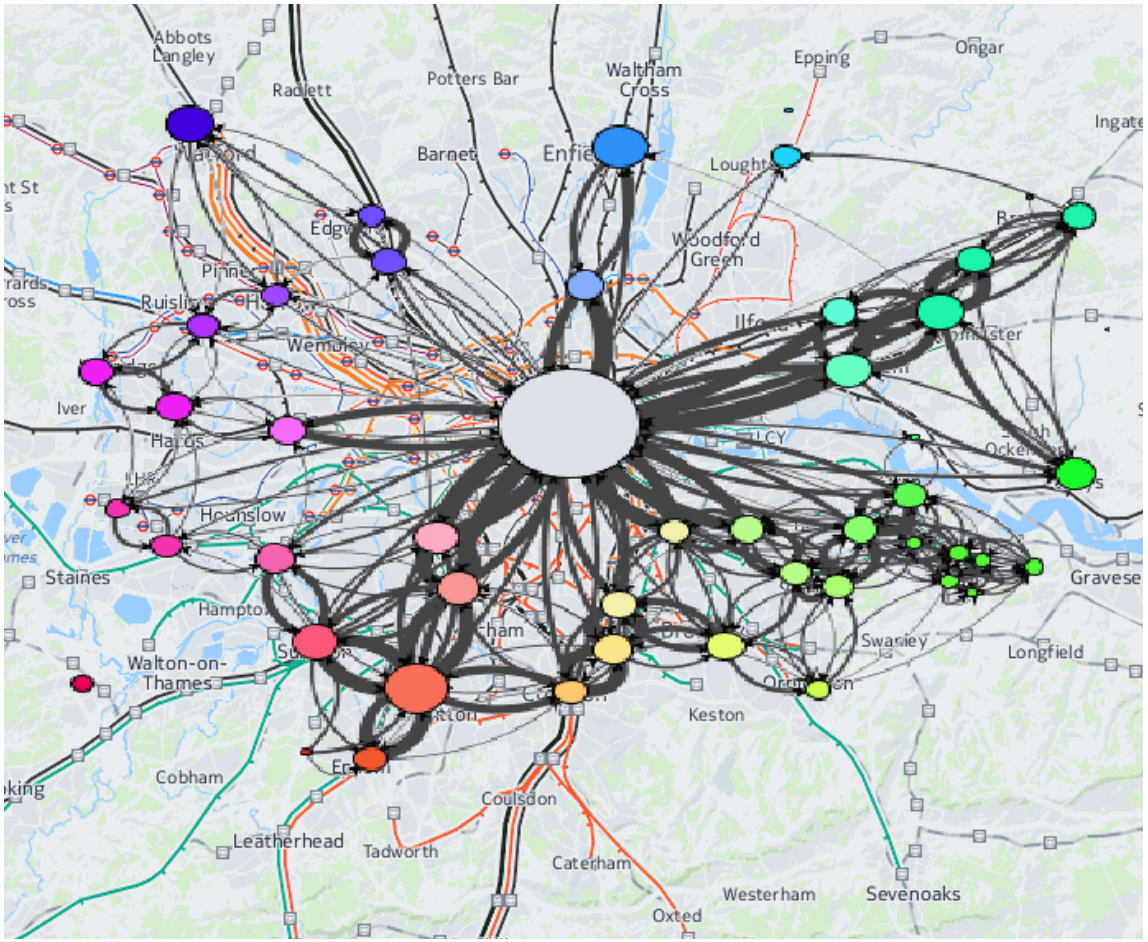




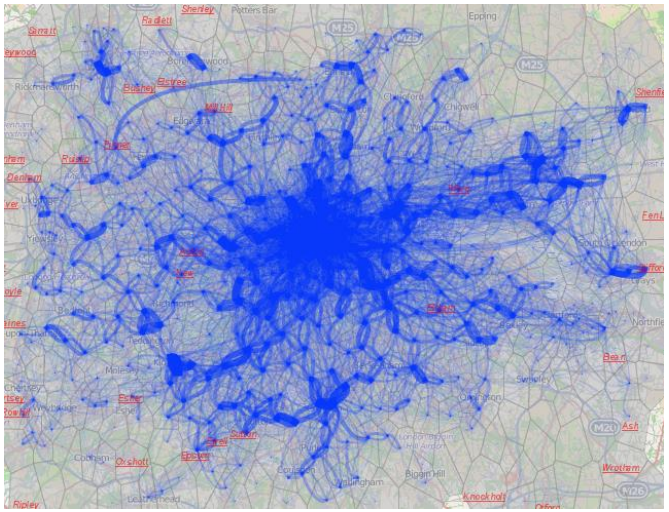
Flow maps obtained from episodic trajectories or OD moves are often unreadable due to visual clutter.

A flow map may be treated as a graph where places are nodes and spatial links are edges. Techniques for node clustering and aggregation may be helpful, but it is important to account for the spatial neighbourhood!

Application of graph visualisation and analysis techniques to spatial links (flows)



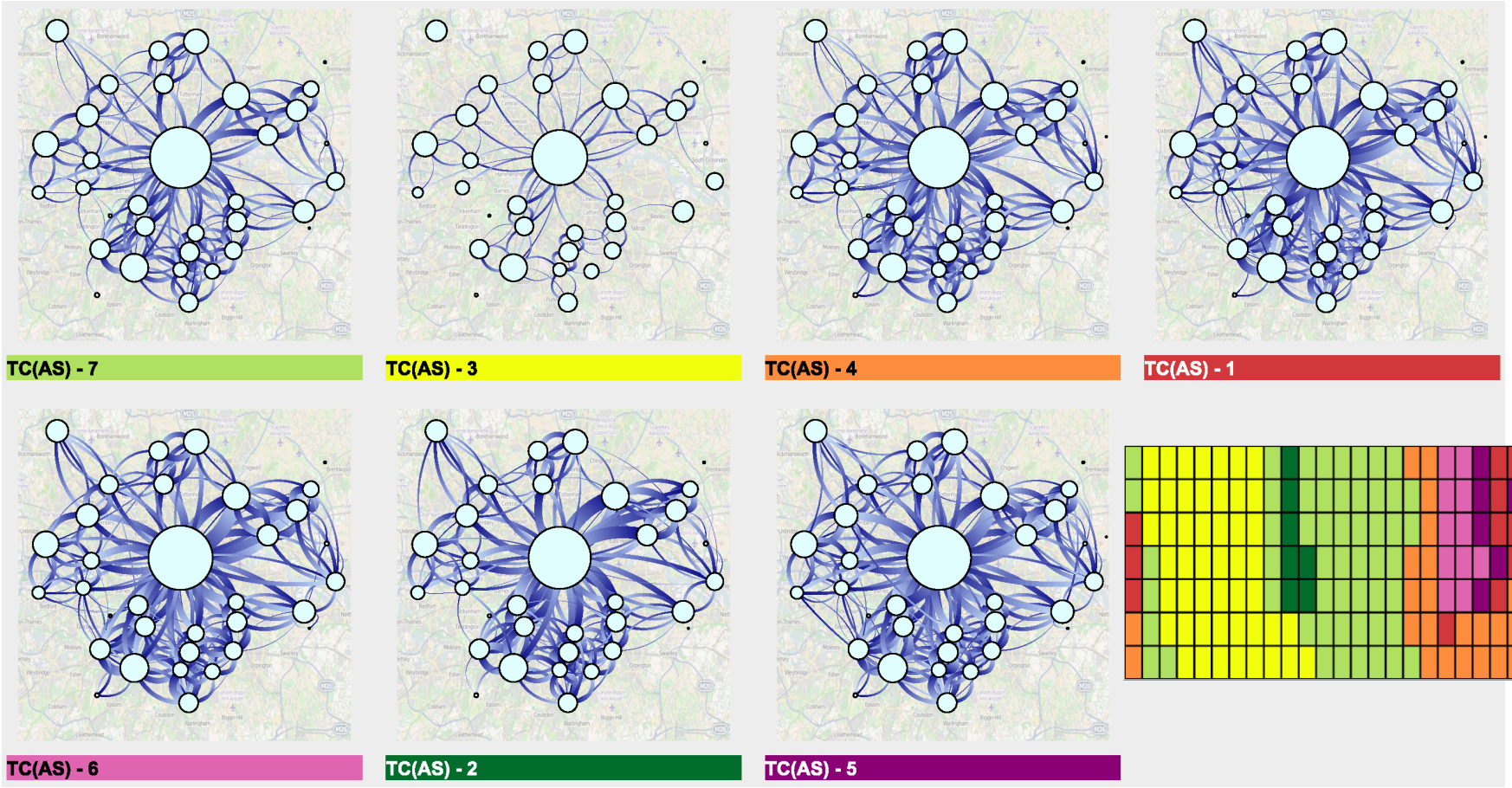
This graph makes the overall pattern of mobility much clearer.



Each graph node represents a group (cluster) of strongly connected neighbouring places. The edges show the flows between the clusters.

The central cluster unites many places in the centre which are highly interconnected. Peripheral regions are strongly connected to the centre.

Application of graph visualisation and analysis techniques to spatial links (flows)



Analysis of the graph evolution over time (weekly time cycle). Time intervals (week hours) have been clustered by similarity of the spatial situations in terms of the people presence in the regions and flows between the regions.

Application of graph visualisation and analysis techniques to spatial links (flows)

Comparison of the spatial situations corresponding to different time clusters

T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, A. Kerren
MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering
IEEE Transactions on Visualization and Computer Graphics, 22(1):11-20, 2016
http://dx.doi.org/10.1109/TVCG.2015.2468111, video: https://vimeo.com/136303590

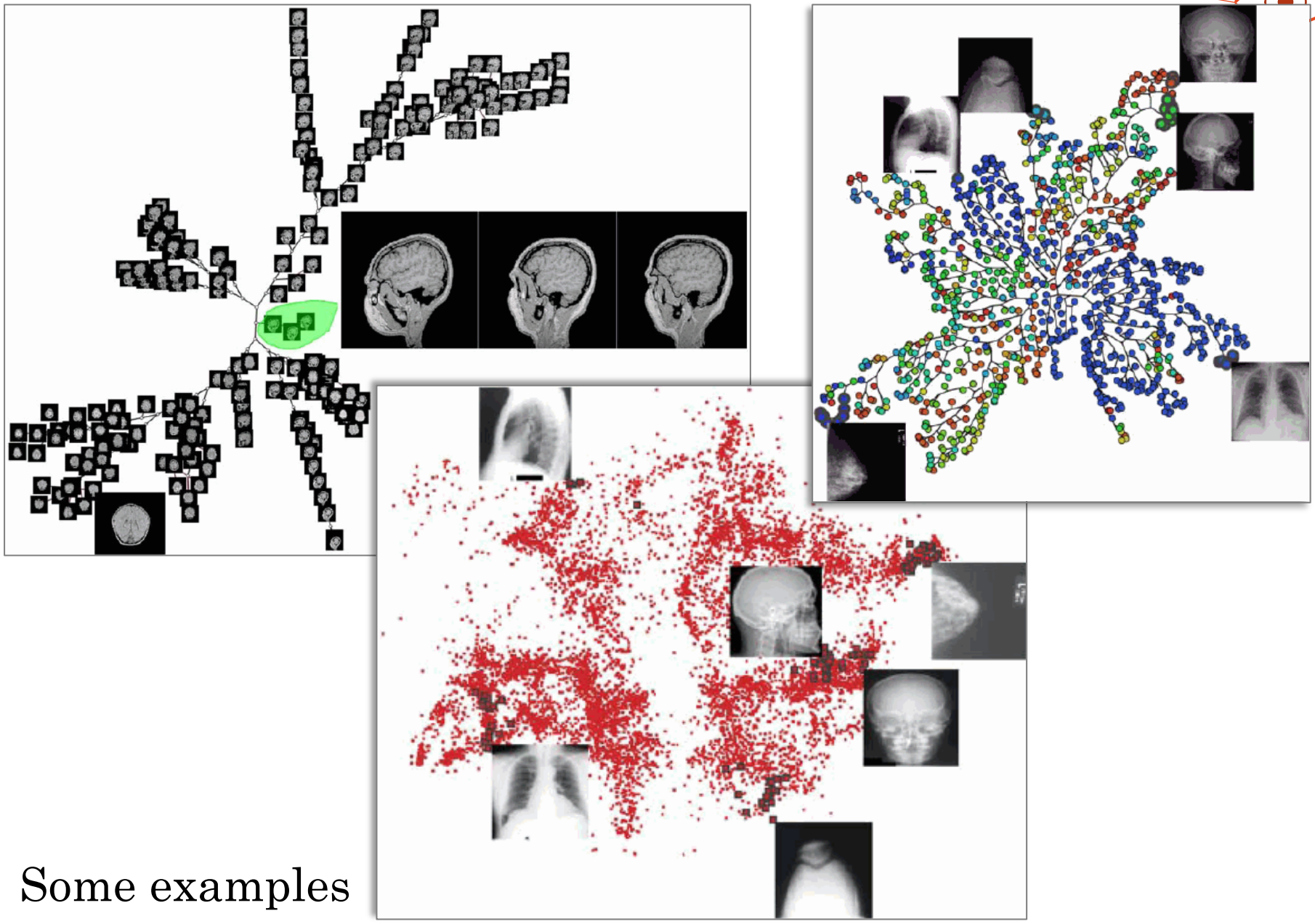
Data type: Images

- E.g., photographs or medical images
- Images are intended for human perception and analysis – humans can usually do this well.
  - Interpretation and analysis of some images may require professional training (medical images).
- Tasks requiring special support:
  - Exploration of or search in large collections of images.
  - Comparison of images for detection of fine differences.

Support to dealing with image collections

General approach

- Compute various aggregate attributes (*features*) from the pixels of each image
  - E.g., average lightness, hue, saturation. Possible refinements:
    - Divide each image into parts by a grid and compute the averages for the parts.
    - Derive more summary statistics, e.g., quartiles, percentiles.
    - Divide the value ranges into bins and derive frequency histograms.
- Two possible approaches to visualisation and interactive analysis:
  - Visualisation and interactive analysis of the derived features as usual multidimensional numeric data
    - Users may select images for viewing based on their features.
  - Images are grouped and organised (in a hierarchy or on a plane) based on the features. Special visualisations represent the organised images.
    - Images are typically represented by icons (thumbnails); the features are not explicitly shown. Users may "expand" the icons for detailed viewing.



Some examples

Representative papers

- Kresimir Matkovic, Denis Gracanin, Wolfgang Freiler, Jana Banova, and Helwig Hauser

  Large Image Collections - Comprehension and Familiarization by Interactive Visual Analysis

  Proceedings of SmartGraphics 2009/ Springer Lecture Notes in Computer Science, 5531, pp. 15-26.

- D.M. Eler, M.Y. Nakazaki, F.V. Paulovich, D.P. Santos, G.F. Andery, M.C.F. Oliveira, J.B. Neto, and R. Minghim

  Visual analysis of image collections

  The Visual Computer, 25 (10), 2009, pp. 923-937

- Y. Gu, C. Wang, J. Ma, R. Nemiroff, and D. Kao

  iGraph: a graph-based technique for visual analytics of image and text collection

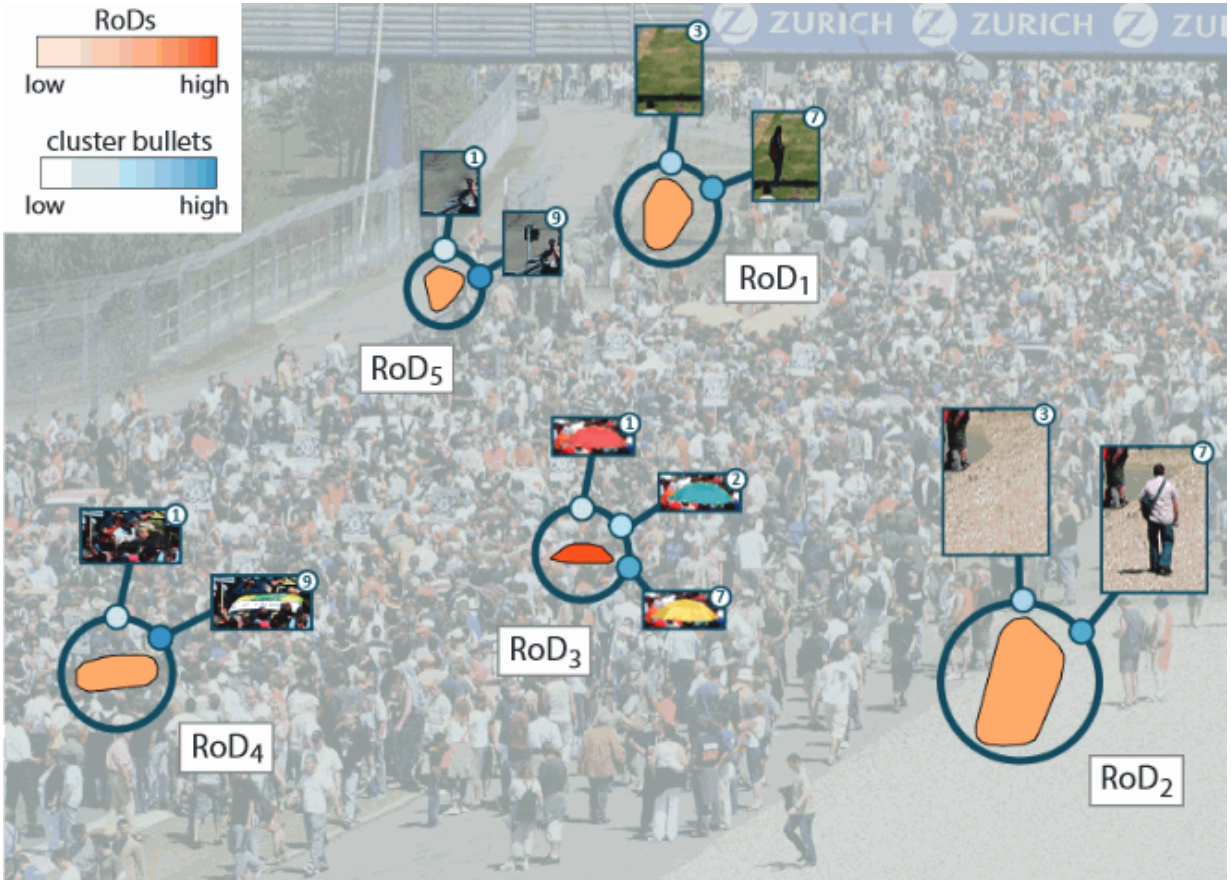
  IS&T/SPIE Visualization and Data Analysis 2015 (Feburary 2015)

Support to image comparison

- Problem setting:
  - There is a (large) collection of images that have much in common but may differ in some parts.
  - Tasks: explore the variations across the set of images; find regions of high variation; find outliers (images that differ much from the rest).
- Approach:
  - For each pair of images, compute colour differences between corresponding pixels; select the pixels for which the difference exceeds a threshold.
  - Take the selected pixels from all comparisons; make regions by grouping neighbouring pixels (RoD – regions of difference).
  - For each RoD, cluster the set of images by similarity in this RoD. The RoD may differ in the number of image clusters obtained, depending on the degree of variation in each region.
  - Visualise the RoD and enable interactive operations (an example follows).

Original images:

Differences:



Recommended reading

J. Schmidt, M.E. Gröller, and S. Bruckner

VAICo: visual analysis for image comparison

IEEE Transactions on Visualization and Computer Graphics, Vol. 19 (12), pp. 2090-2099, December 2013

Data type: Video

- Problems in visual analysis of video by a human:
  - A video has to be viewed sequentially, which may take much time.
  - Change blindness: a human often misses important changes when the eyes are focused on a different area.
- Two areas of research in visualisation and visual analytics on supporting viewing and analysing video data:
  - Summarisation and compact representation of the video content.
  - Extraction and representation of important changes, in particular, object movements.
- In both areas, lots of computational techniques for image processing are utilised.
  - In particular, for detecting moving objects and separating them from the background.

Video summarisation

Representative papers

- Connelly Barnes, Dan B Goldman, Eli Shechtman, and Adam Finkelstein
  Video Tapestries with Continuous Temporal Zoom
  ACM Transactions on Graphics (Proc. SIGGRAPH), August 2010.
  https://vimeo.com/13403055
  https://www.youtube.com/watch?v=uadKlZnaBmE
- Carlos D. Correa and Kwan-Liu Ma
  Dynamic Video Narratives
  ACM Transactions on Graphics, vol. 29, no. 4, 2010
  https://www.youtube.com/watch?v=0zm3lauoOfo

Connelly Barnes, Dan B Goldman, Eli Shechtman, and Adam Finkelstein

Video Tapestries with Continuous Temporal Zoom

A multiscale tapestry represents an input video as a seamless and zoomable summary image, which can be used to navigate through the video. This visualisation eliminates hard borders between frames, providing spatial continuity and also continuous zoom to finer temporal resolutions. The figure illustrates three zoom levels.

https://vimeo.com/13403055; https://www.youtube.com/watch?v=uadKlZnaBmE

Carlos D. Correa and Kwan-Liu Ma

Dynamic Video Narratives

The user can click on objects in a summarised image of a video to see their motion. The user can also interactively edit the summary.

https://www.youtube.com/watch?v=0zm3lauoOfo

Extraction and visualisation of object movements from video

Representative papers

- Amir H. Meghdadi and Pourang Irani

  Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization

  IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, December 2013.

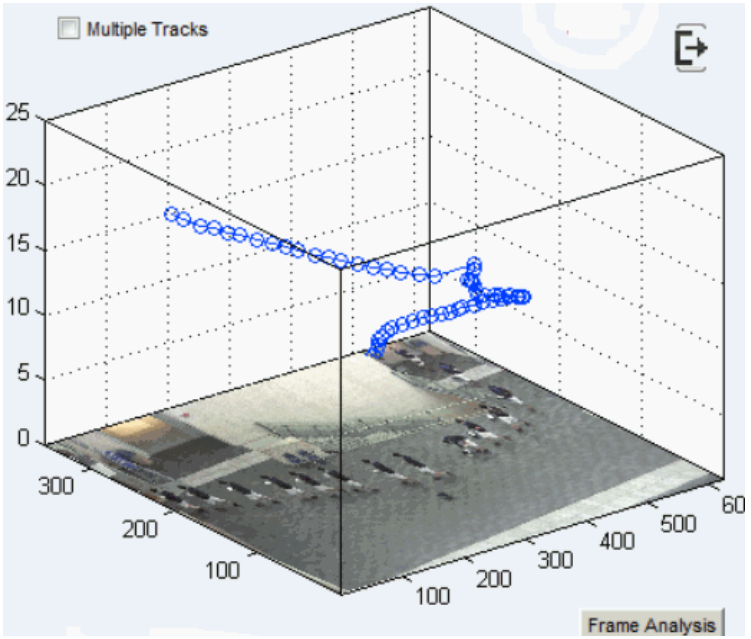
- Markus Höferlin, Benjamin Höferlin, Gunther Heidemann, and Daniel Weiskopf

  Interactive Schematic Summaries for Faceted Exploration of Surveillance Video

  IEEE Transactions on Multimedia, Vol. 15, No. 4 (2013)

Amir H. Meghdadi and Pourang Irani

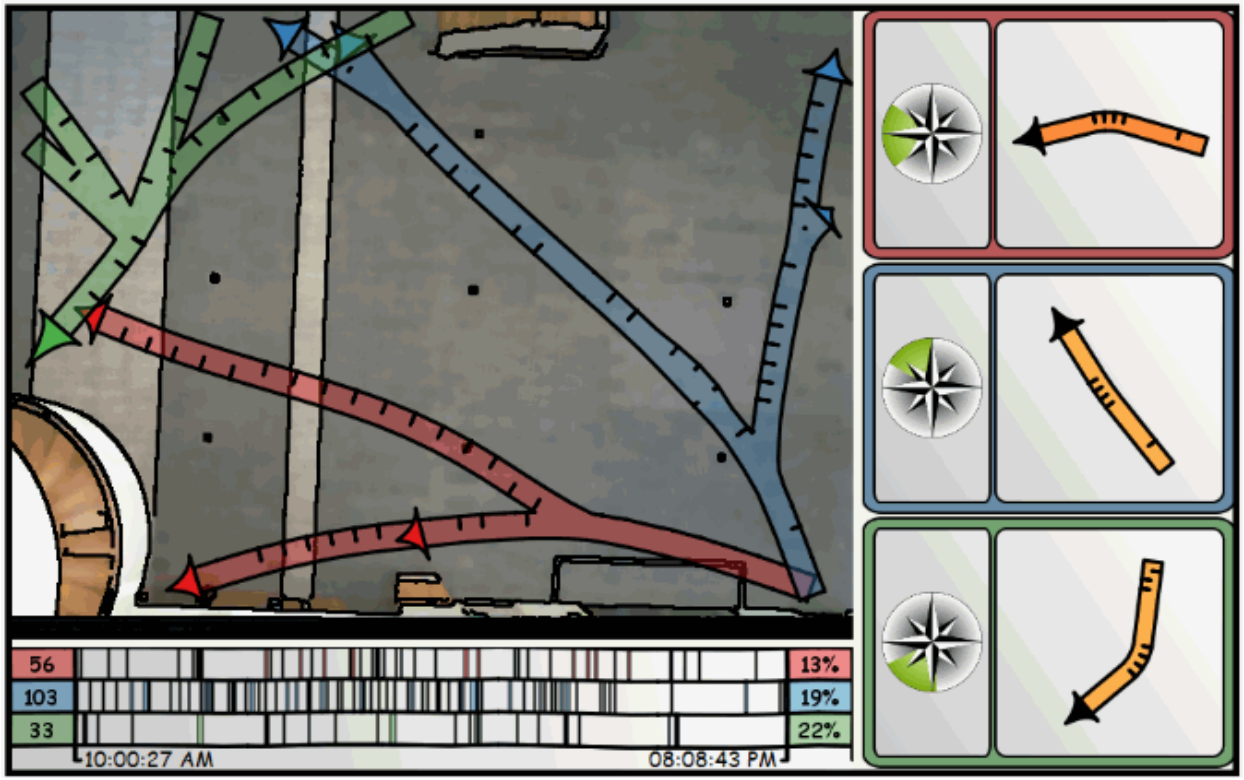
Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization



The movements of each object are summarised in one image containing multiple states of the object and also represented as a trajectory line in a space-time cube. The system provides tools for spatial and temporal filtering based on interactively defined regions of interest.

Markus Höferlin, Benjamin Höferlin, Gunther Heidemann, and Daniel Weiskopf

Interactive Schematic Summaries for Faceted Exploration of Surveillance Video



An interactive schematic summary of trajectories of moving objects extracted from a video. The trajectories have been clustered by similarity; the clusters are represented in a summarised form as flows.

Data type: Texts

- Text analysis problems addressed in visual analytics:
  - Efficient analysis of one large text document.
  - Analysis of and finding relevant information in a large collection of texts.
  - Comparison of text documents and document collections.
  - Analysis of text streams (texts appearing over time).
  - Analysis of texts in time and space.
  - Investigative analysis of text collections.

Data pre-processing

- All VA approaches involve pre-processing of texts using computational techniques.
  - Purpose: transform the unstructured data into (more) structured.
- Classes of pre-processing operations (in the order of increasing sophistication):
  - Calculation of numeric measures: word lengths, sentence lengths, …
  - Extraction of significant keywords
  - Probabilistic topic modelling
  - Application of NLP (Natural Language Processing) techniques for
    - Identification of named entities (people, places, organisations, etc.)
    - Sentiment analysis (identification of emotions and attitudes)

Analysis based on numeric measures

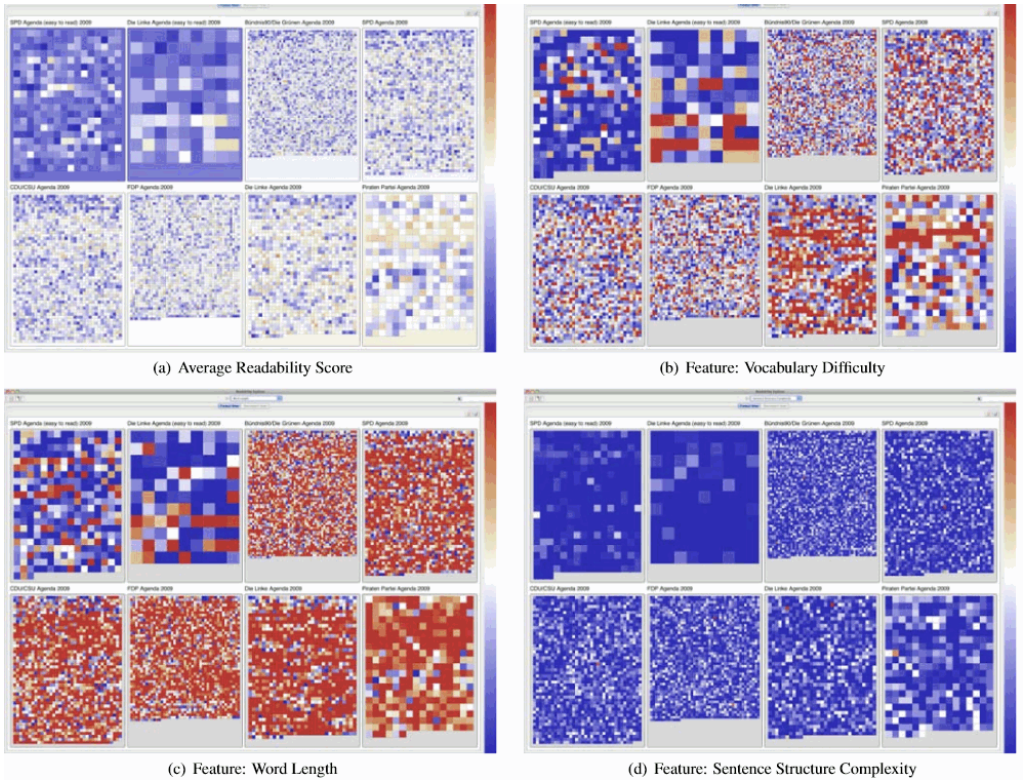
An example

- Daniela Oelke, David Spretke, Andreas Stoffel, and Daniel A. Keim

  Visual Readability Analysis: How to Make Your Writings Easier to Read

  IEEE Transactions on Visualization and Computer Graphics, Vol. 18, No. 5, pp. 662-674, 2012

Each document is represented by a sequence of pixels (arranged in rows), each pixel representing a word, sentence, or paragraph. Numeric values of a selected measure (feature) are encoded in pixel colours using a diverging colour scale. The visualisation allows comparison of multiple documents and identification of "difficult" parts of a text.



(a) Average Readability Score

(b) Feature: Vocabulary Difficulty

(c) Feature: Word Length

(d) Feature: Sentence Structure Complexity

Extraction of significant keywords

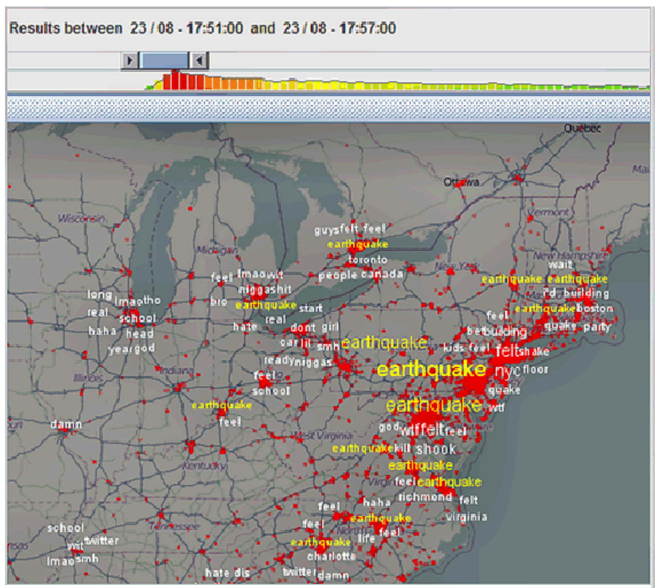
- Words that occur frequently are not informative.
- From words that usually occur rarely but are more frequent in a given text, we can guess what this text is about.
- There are text processing tools that remove frequent words (a.k.a. “stop words”) from texts and keep more informative words.
- Another approach is search for predefined terms of interest.

Analysis based on keyword extraction

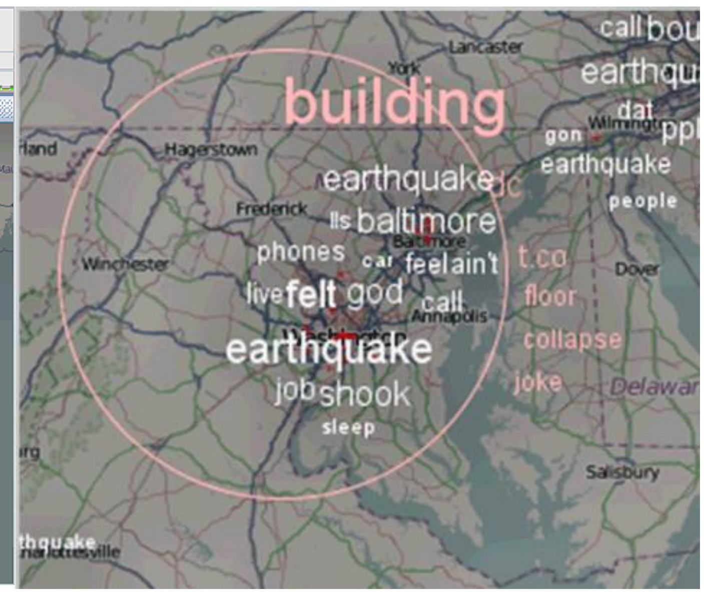
Example 1: detection of abnormal keyword frequencies

- D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl
  Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages
  Proc. **IEEE Pacific Visualization Symp. (PacificVis)**, pp.41–48, 2012

DBC is applied to Twitter events based on their spatial positions, times, and terms used. The analysis is done in real time as the events appear.



ST concentrations of term occurrences are represented by text clouds on a map.



An interactive "lens" shows more details in a selected area.

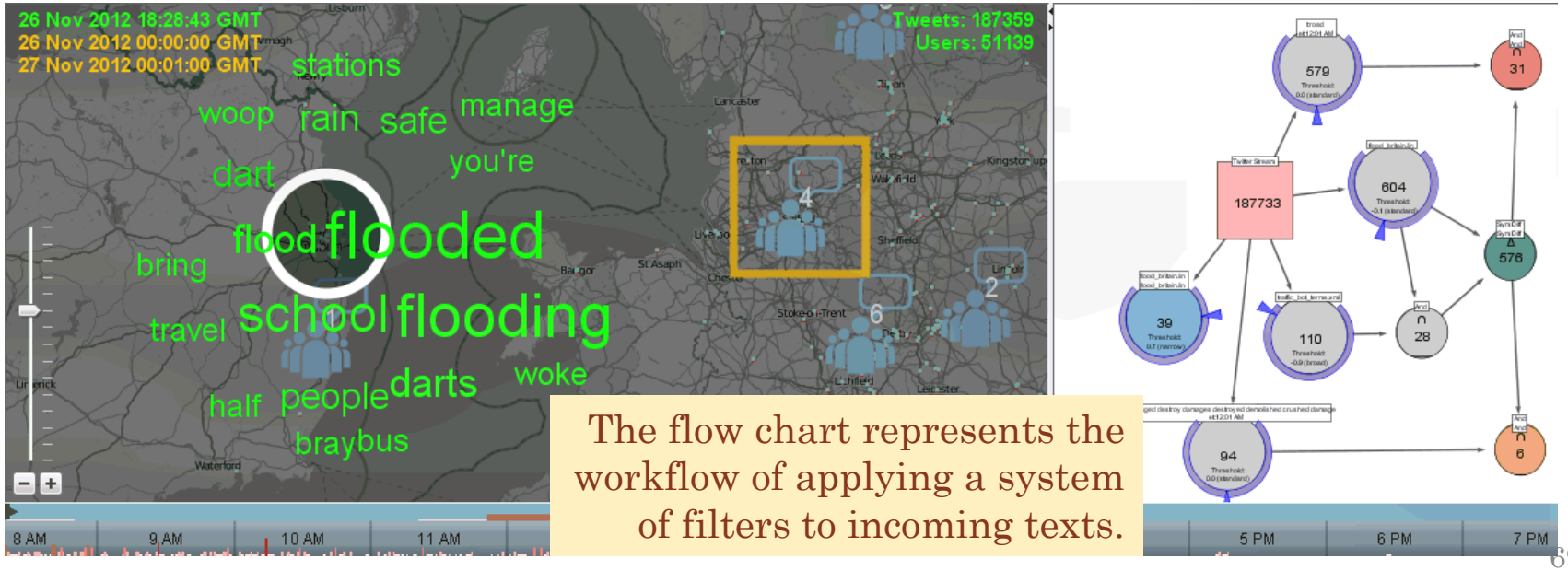
Analysis based on keyword extraction

Example 2: use of interactively built filters and classifiers

- H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl

  ScatterBlogs2: Real-Time Monitoring of Microblog Messages Through User-Guided Filtering

  IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, pp. 2022-2031, 2013

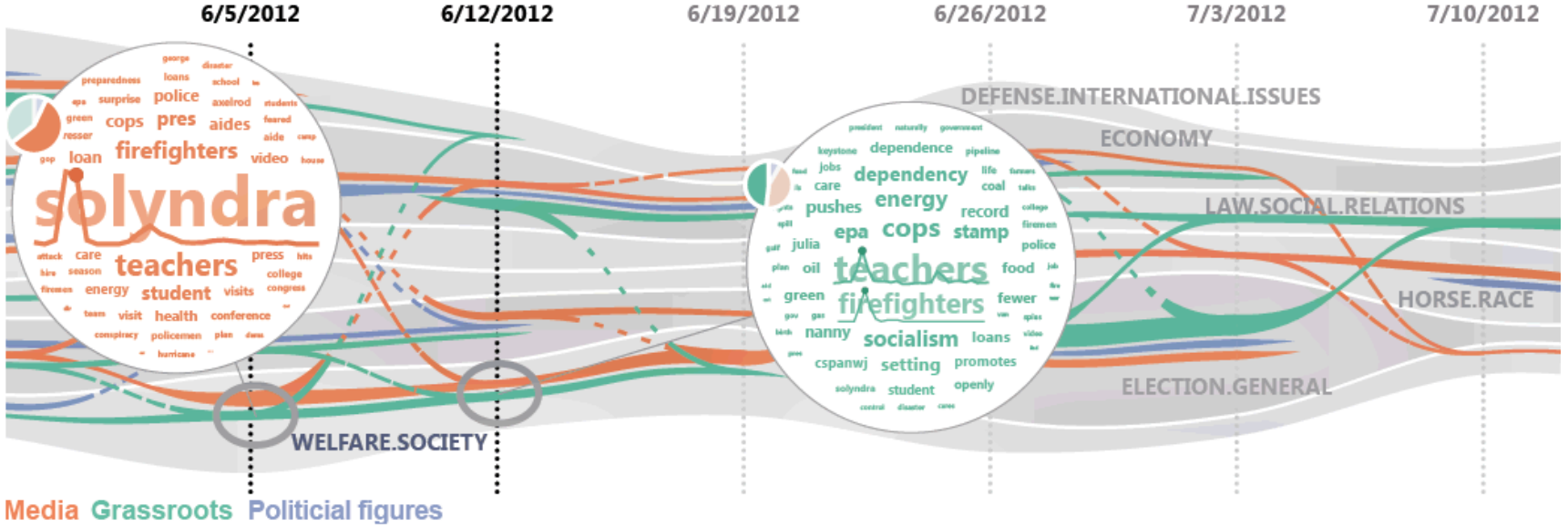


The flow chart represents the workflow of applying a system of filters to incoming texts.

Analysis based on keyword extraction

Example 3: analysis of temporal evolution of a text stream

- P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J.J.H. Zhu, and H. Qu
  Visual Analysis of Topic Competition on Social Media
  IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, pp. 2012-2021, 2013



Topics and relevant keywords are specified by experts. The system enables the analysis of the popularity of different topics over time, the trending keywords, and the contribution of different classes of text producers.

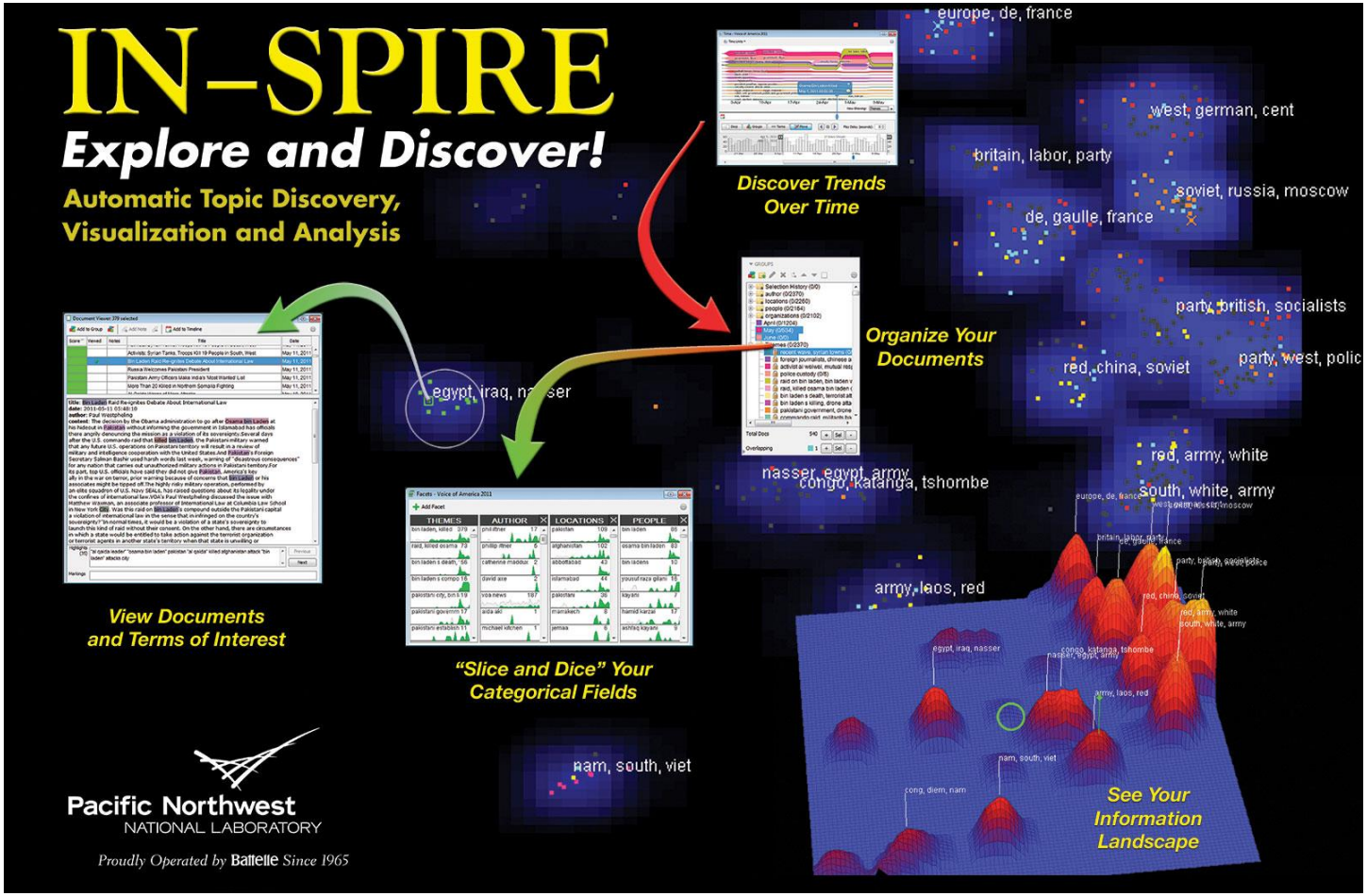
Probabilistic topic modelling

- Topic modelling is a class of statistical techniques that analyse the words in texts to discover the themes that run through them.
- A topic is represented by a set of significant words that are considered as related (due to repeated close occurrences in texts).
  - E.g., “dog”, “bone”, “tail”, “bark”, …
  - The words are associated with weights expressing their significance in regard to the given topic.
- Input of a topic modelling algorithm: set of pre-processed texts, each text represented by a set (bag) of words with their frequencies.
  - “Stop words” are previously removed.
- Output:
  1. Set of extracted topics (i.e., combinations of weighted keywords)
  2. For each text, the probability of each topic.

Analysis based on topic modelling

Example 1: IN-SPIRE™ *(commercially offered)*

- http://in-spire.pnnl.gov/ : descriptions, pictures, videos, FAQ, etc.



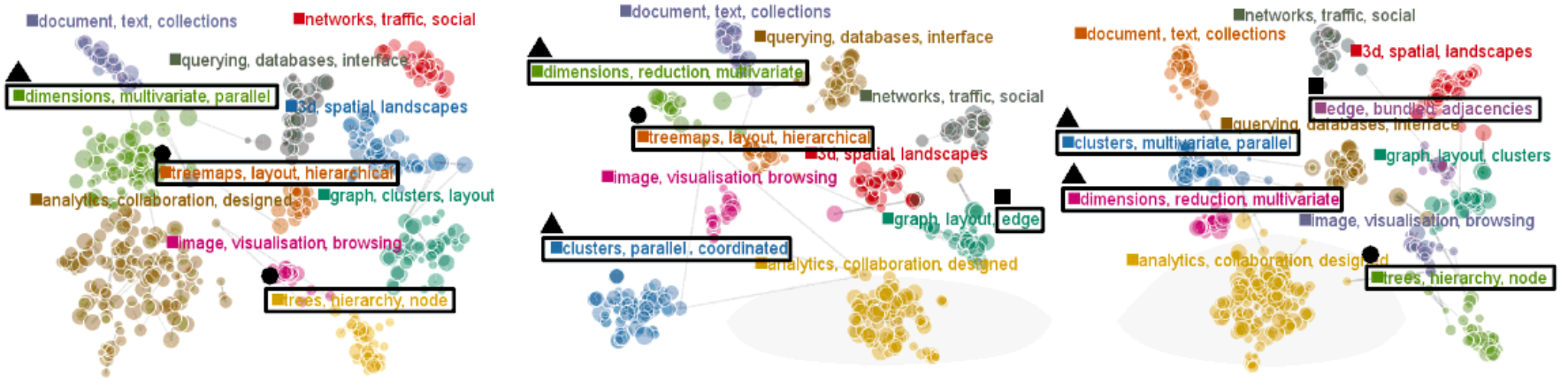
Analysis based on topic modelling

Example 2: involvement of the human analyst in TM

- J. Choo, C. Lee, C.K. Reddy, and H. Park

  UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization

  IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, pp. 1992-2001, 2013



(a) The initial visualization (b) The visualization after *topic keyword refinement* (c) The visualization after *keyword-induced topic creation* and *topic splitting*

Employs a semi-supervised learning method to incorporate user's feedback. Allows the user to refine keywords, split topics, create new topics, etc.

Natural Language Processing (NLP)

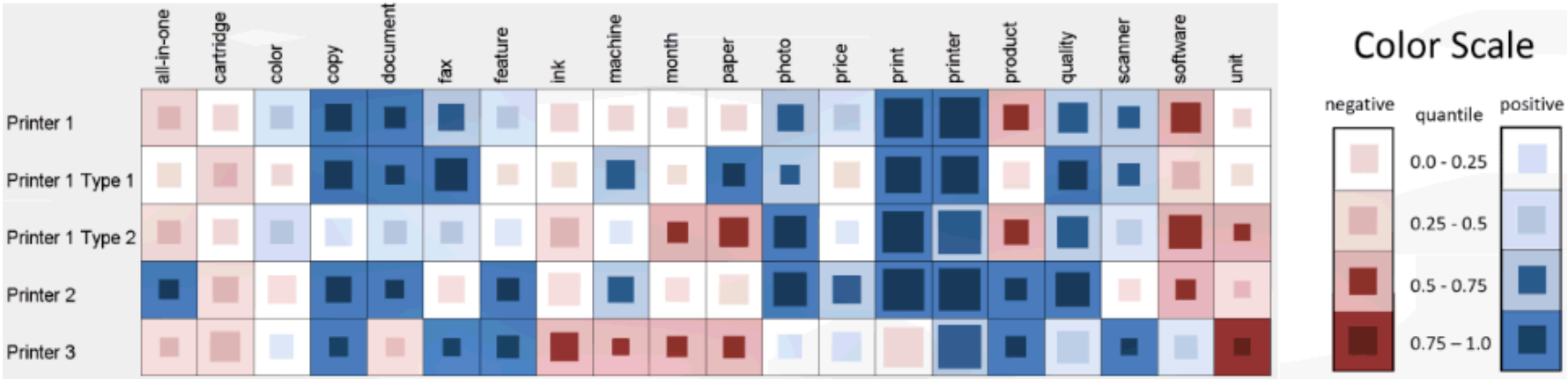
- A large area of research and technology, from which two classes of techniques are most frequently used in text data analysis:
  - Named Entity Recognition (NER), a.k.a. named entity identification
  - Sentiment analysis
- Named entity recognition techniques classify elements in texts into predefined categories:
  - names of persons, organisations, places;
  - expressions of times, quantities, monetary values, percentages; …
- Sentiment analysis techniques extract subjective information, i.e., positive or negative emotions and opinions

An example of using sentiment analysis

- D. Oelke, M. C. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L.-E. Haug and H. Janetzko

  Visual Opinion Analysis of Customer Feedback Data

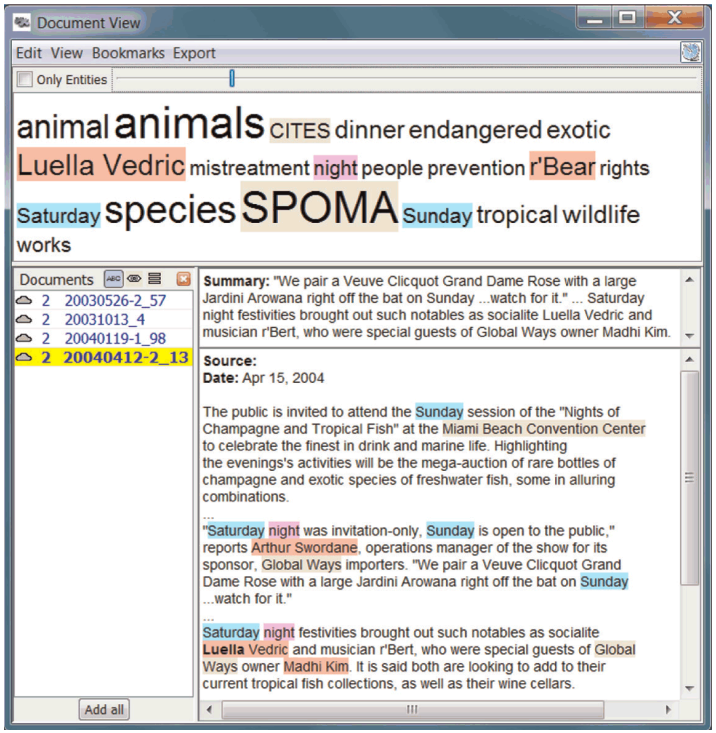
  Proc. **IEEE Symposium on Visual Analytics Science and Technology (VAST '09)**, IEEE, pp. 187-194, 2009



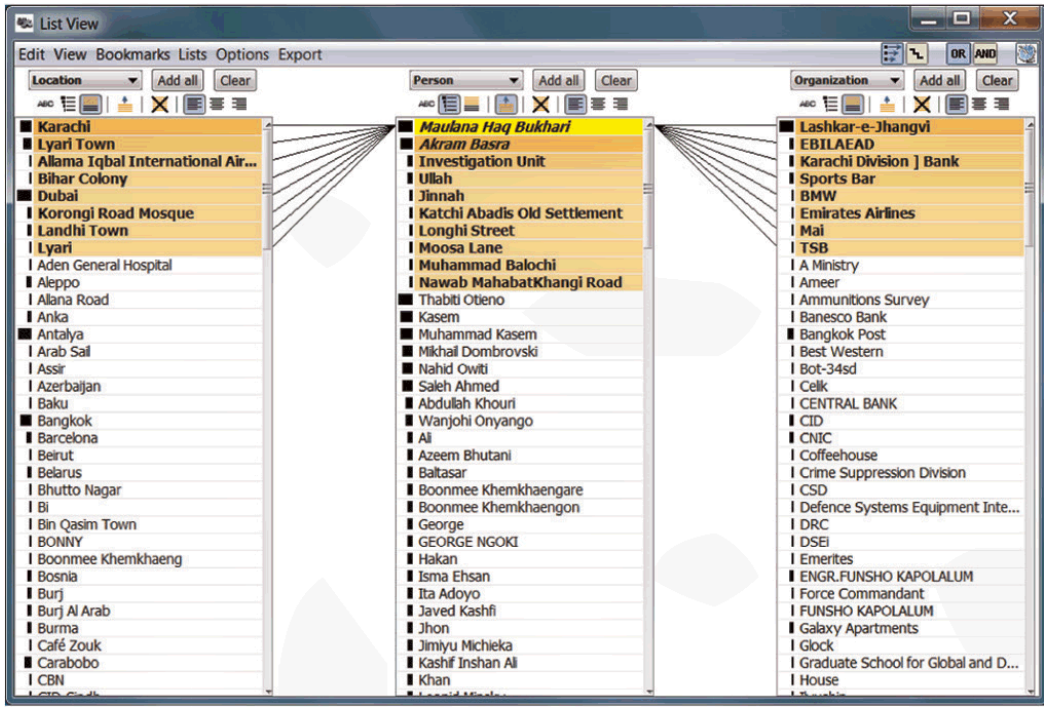
Automatically constructs visual summary reports from a large set of customer comments and ratings. Allows further analysis for detection of groups of customers with similar opinions and finding correlations between different aspects.

An example of using NER: Jigsaw

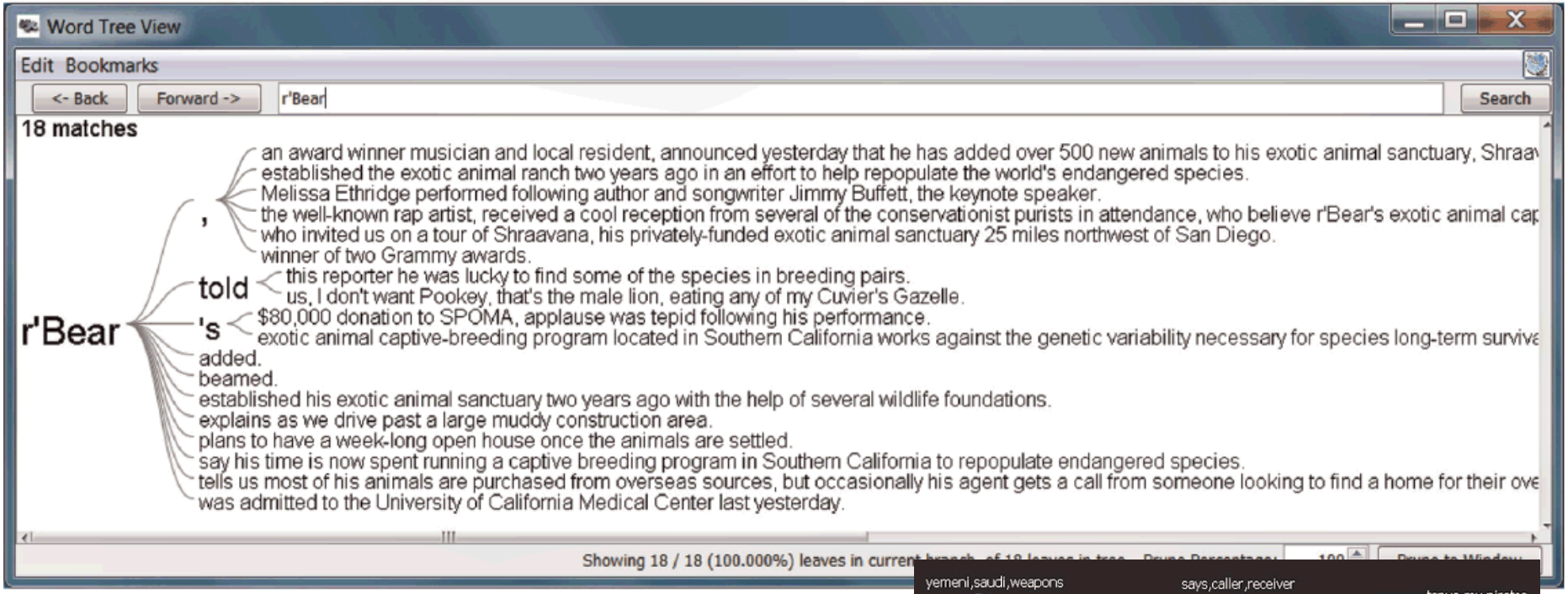
- http://www.cc.gatech.edu/gvu/ii/jigsaw/
  Descriptions, illustrations, tutorial video, other videos, papers
- The system is freely downloadable
- Primary purpose: support investigative analysis of a collection of text documents
  - Such analyses involve establishing connections between facts contained in different documents
  - Accordingly, system's primary focus is displaying connections between entities across the documents



Identified entities are highlighted in a text. Different colours correspond to different categories of entities.



This view shows lists of different categories of entities (locations, persons, and organisations). For a selected entity, connected entities (occurring in the same documents) are shown.



For a selected entity, a Word Tree summarises all sentences where it occurs (over the whole collection of documents).

A Document Cluster view shows clusters of related documents, indicating which documents have been already viewed. Documents mentioning a selected entity are specially marked.

Visual analytics of diverse data types

A summary

- Visual analytics methods and tools are developed not only for data representable in table form but also for other types of data.
  - Some data types have been primarily meant for human perception: texts, images, and videos.
    - Computers are much weaker than humans in processing such information.
  - However, humans cannot (efficiently) cope with large and/or numerous pieces of data of these types.
  - Particular focus of VA: support to analysis of large data.
- VA methods and tools employ specialised computational techniques oriented to these data types: graph analysis, image processing, video processing, text processing.
- Specialised visualisations are developed for these data types.
  - When computational processing techniques derive numeric attributes from these data, visualisations suitable for numeric data are (also) used.

Questions?

Visual analytics of diverse data types

Visual analytics systems

Commercial systems

- **Tableau Desktop** http://www.tableau.com/products/desktop
- **Spotfire** http://spotfire.tibco.com/
- **Palantir** https://www.palantir.com/products/
- **GeoTime** http://geotime.com/
- **Miner3D** http://www.miner3d.com/

Common focus: business analytics

- **IN-SPIRE™** http://in-spire.pnnl.gov/ - text analysis

Educational systems

- **Mondrian** http://stats.math.uni-augsburg.de/mondrian/
  - Numeric, categorical, and geo-referenced data
- **Improvise** http://www.cs.ou.edu/~weaver/improvise/
  - Various types of data, numerous visualisations, powerful tools for coordination between multiple views
- **Tulip** http://tulip.labri.fr/TulipDrupal/
  - Relational data (networks, graphs)
- **Jigsaw** http://www.cc.gatech.edu/gvu/ii/jigsaw/
  - Texts
- **Weave** - Web-based Analysis and Visualization Environment http://oicweave.org/index.php
  - Geographically referenced data

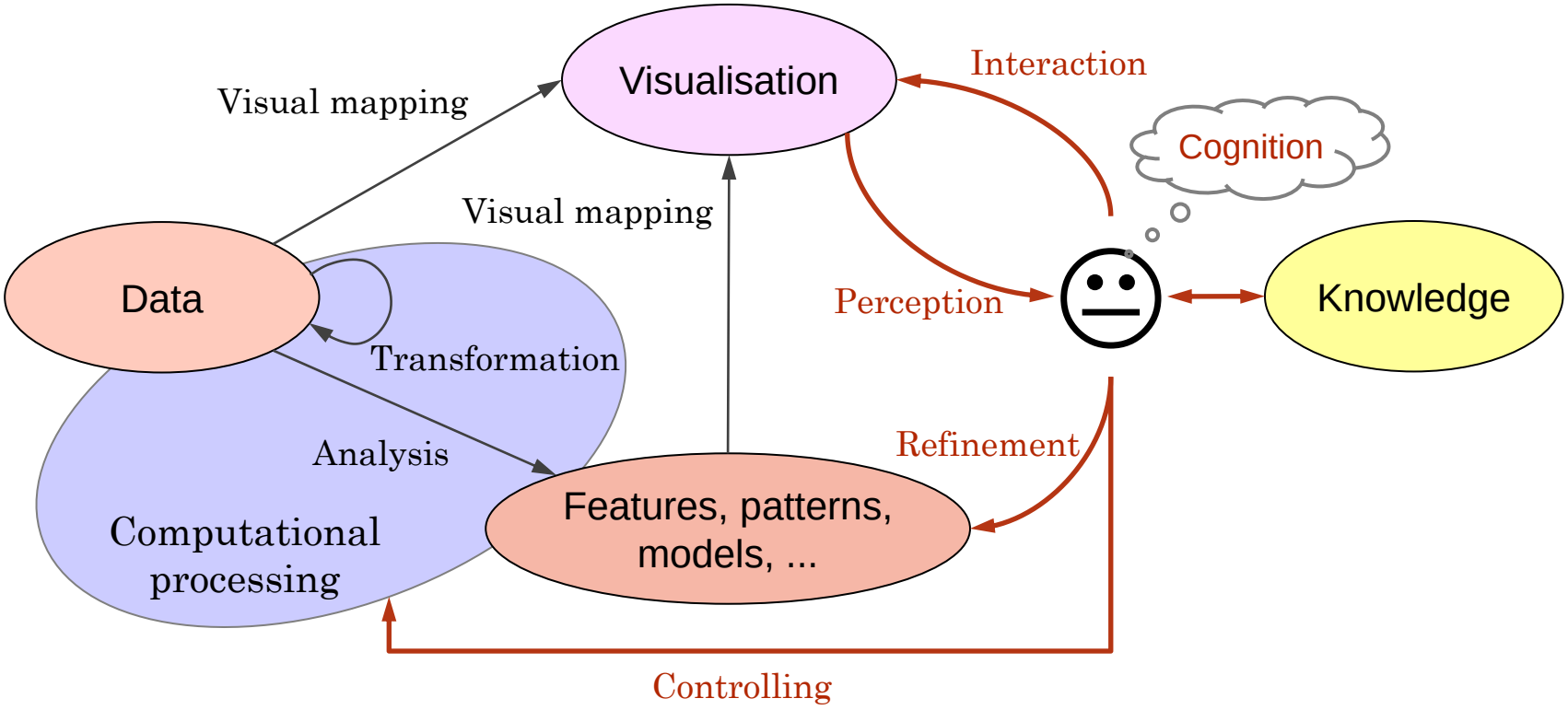
Wrap-up

What you (are supposed to) have learned

Specific knowledge and skills

- Fundamentals of visual representation of data
- Use of interactive operations
- Approaches to analysing data of different types
  - Multivariate (multi-attribute) data; spatially referenced data; time series; spatial time series; spatial events; OD moves and trajectories; spatial flows
- Data transformations
- Use of clustering
  - By example of clustering, the general way of using computational techniques in data analysis.

Visual Analytics approach



Highlighted in red are activities expected from the human analyst. Note that the human is not a passive recipient of computer outputs but an active and leading force in the analysis process.

What you (are supposed to) have learned

General knowledge and attitudes

- Visual analytics activity: data analysis and derivation of knowledge by humans employing their perceptual and cognitive capabilities.
- Visual analytics approach: combine computational processing with interactive visual interfaces enabling human cognition.
  - Important: the human interacts with the machine, not just views what it has produced!
    - Tries different parameter settings, different computational methods, different data transformations, feature selections, data divisions, etc.
    - Provides background knowledge and evaluation feedback when the algorithms are designed to accept these.
- Visual analytics as a behaviour and discipline of mind:
  - See and understand your data before trying to do something with it.
  - Do not blindly take what machine gives you; look, understand, think, experiment: what will happen if I change something?

Questions?