Gennady Andrienko,
Natalia Andrienko,
Salvatore Rinzivillo

Leveraging Spatial Abstraction in Traffic Analysis and Forecasting with Visual Analytics

Paper under review in
Information Systems (Elsevier)

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

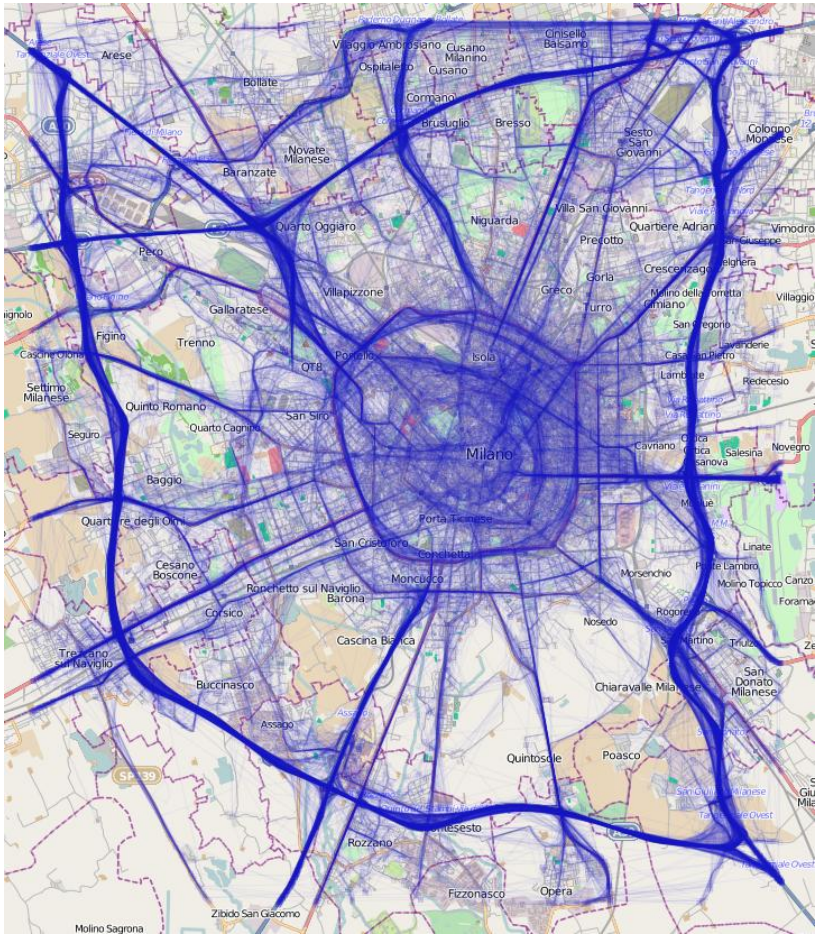
Predictive visual analytics by example

- 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

Example dataset: trajectories of cars in Milan

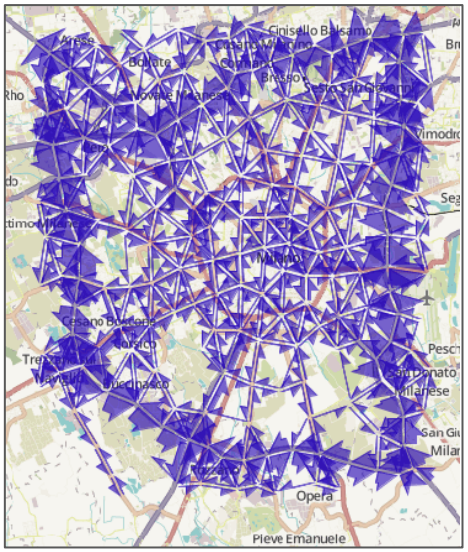
- GPS-tracks of 17,241 cars in Milan, Italy
- Time period: April 01-07, 2007 (Sunday to Saturday)
- Received from Octo Telematics www.octotelematics.com special thanks to Tina Martino
- Data structure:
  - Anonymised car identifier
  - Date and time
  - Geographic coordinates
  - Speed

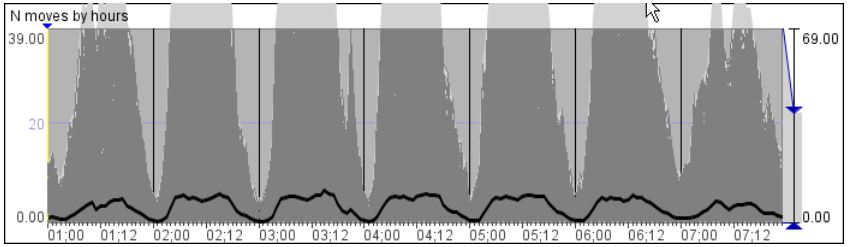


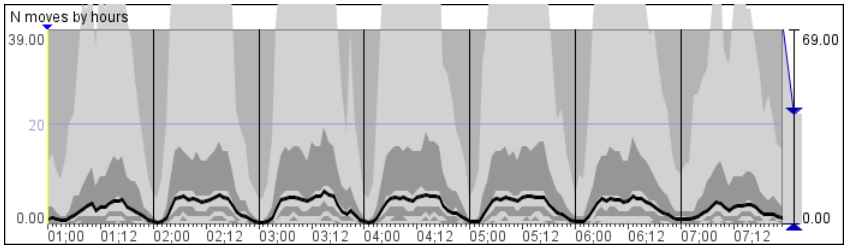
The trajectories from one day are drawn on a map with 5% opacity

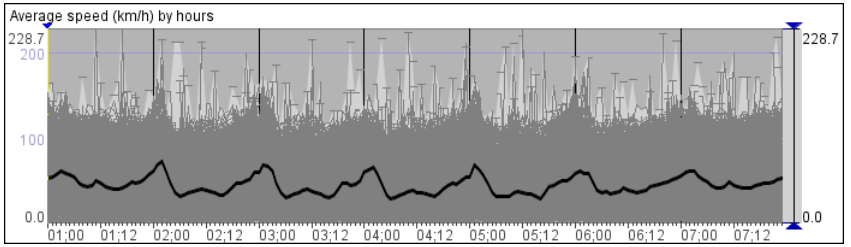
Data transformation: ST aggregation

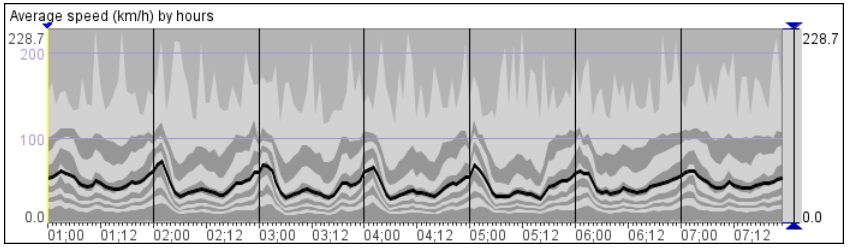
- Divide the territory into cells.
- Divide the time into hourly intervals.
- For each time interval and each ordered pair of neighbouring cells P $\rightarrow$ Q count the vehicles that moved from P to Q and compute their mean speed.









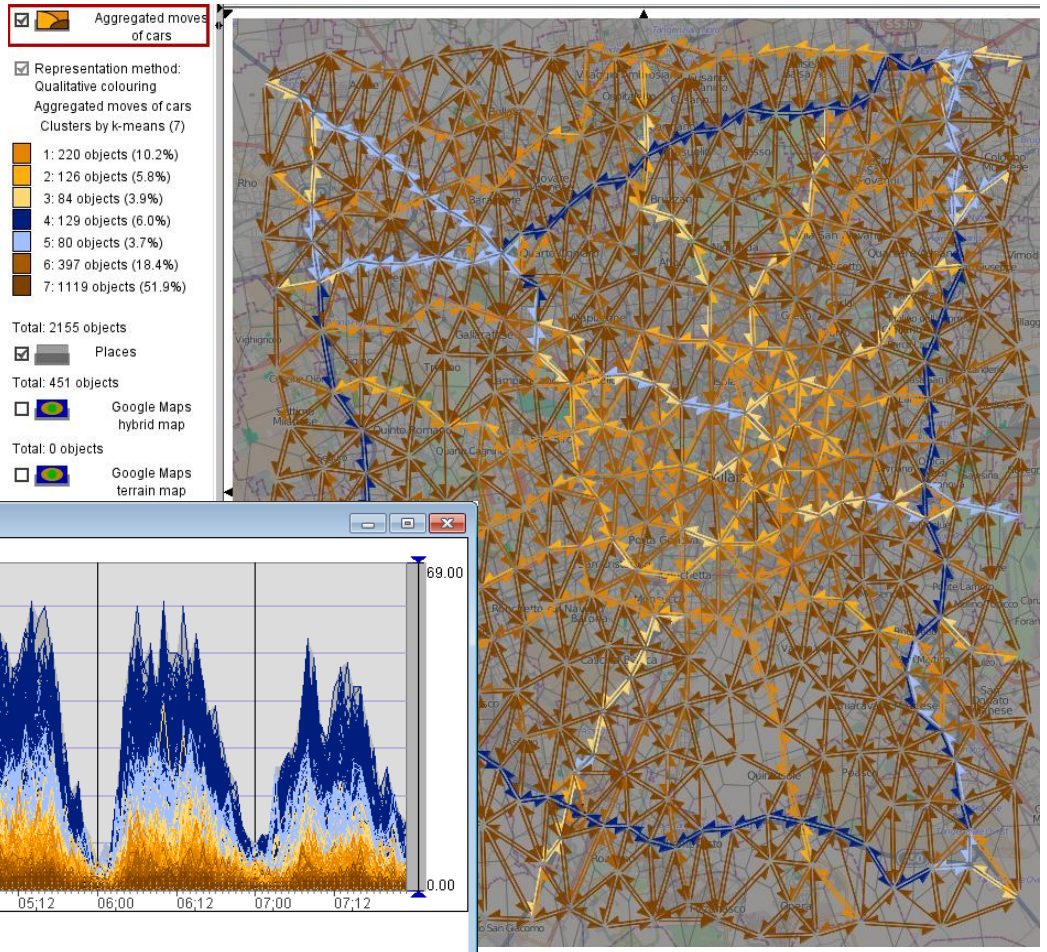


Part 1. Prediction of regular traffic flows

1) Partition-based clustering of the links

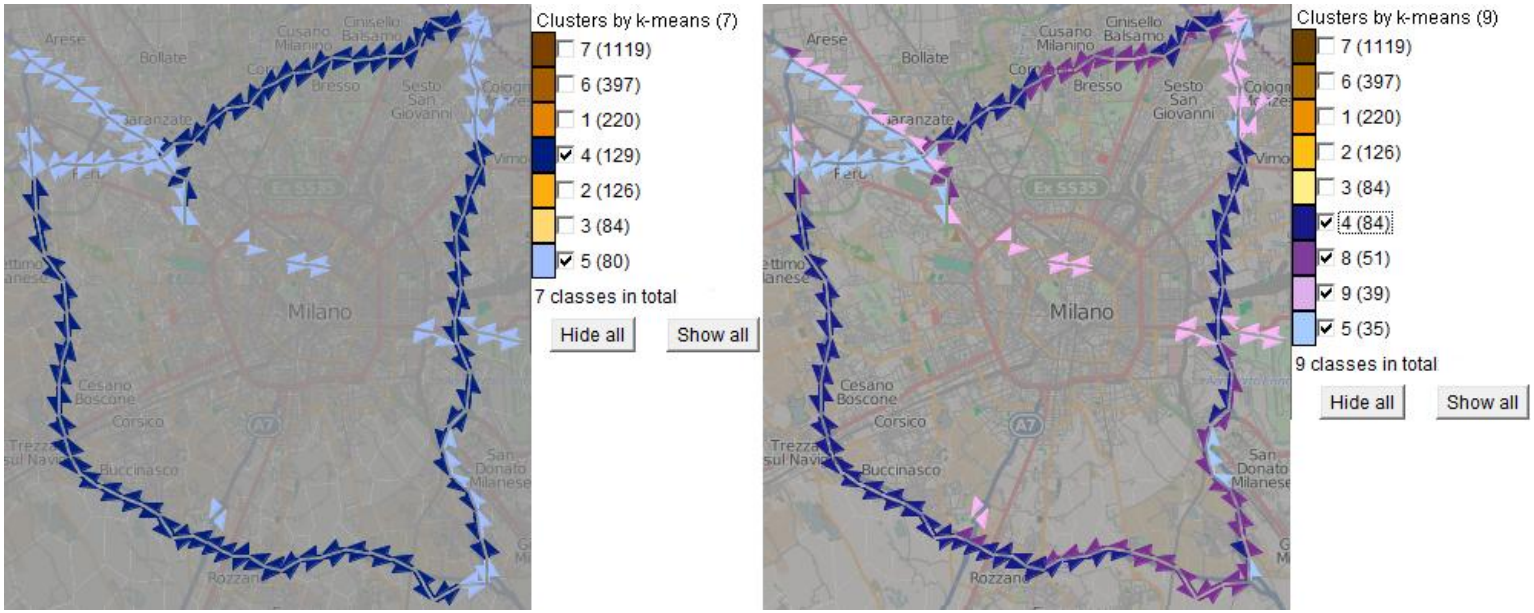
by similarity of the TS of the hourly move counts

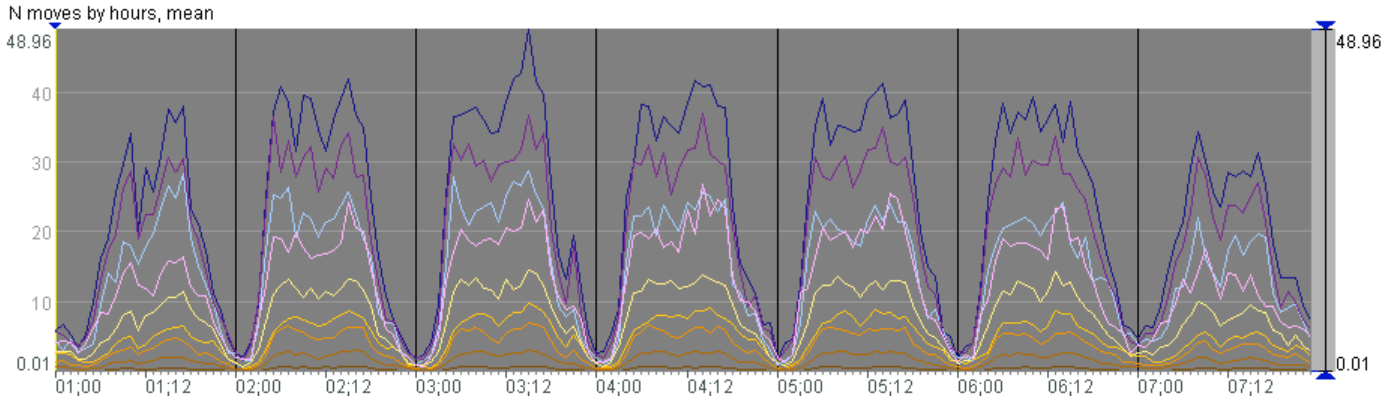
- Clustering method: k-means
- Tried different k from 5 to 15
- Immediate visual response facilitates choosing the most suitable k (i.e., giving interpretable and clear results).



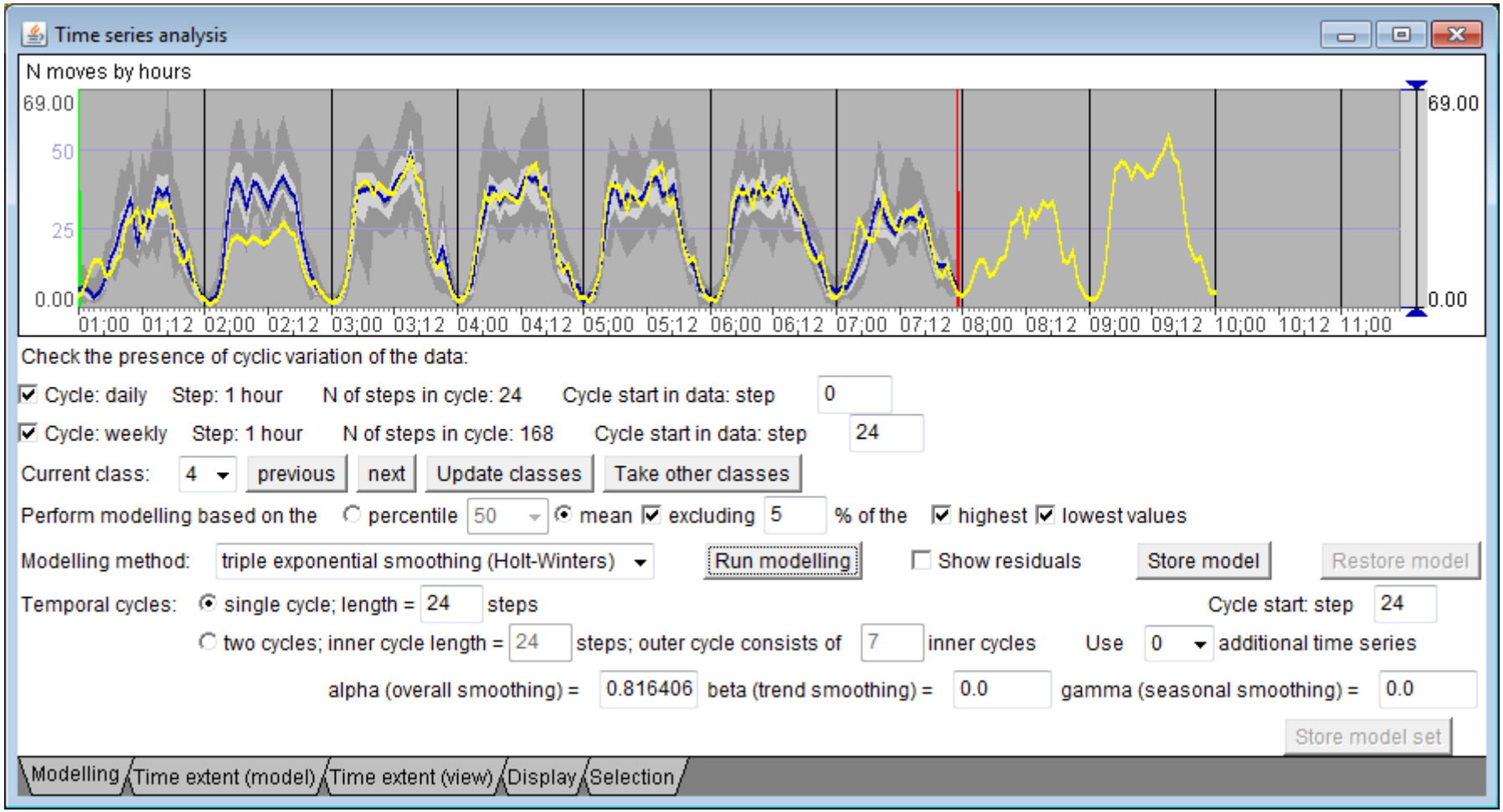
1.a) Re-grouping by progressive clustering

for reducing internal variation in clusters

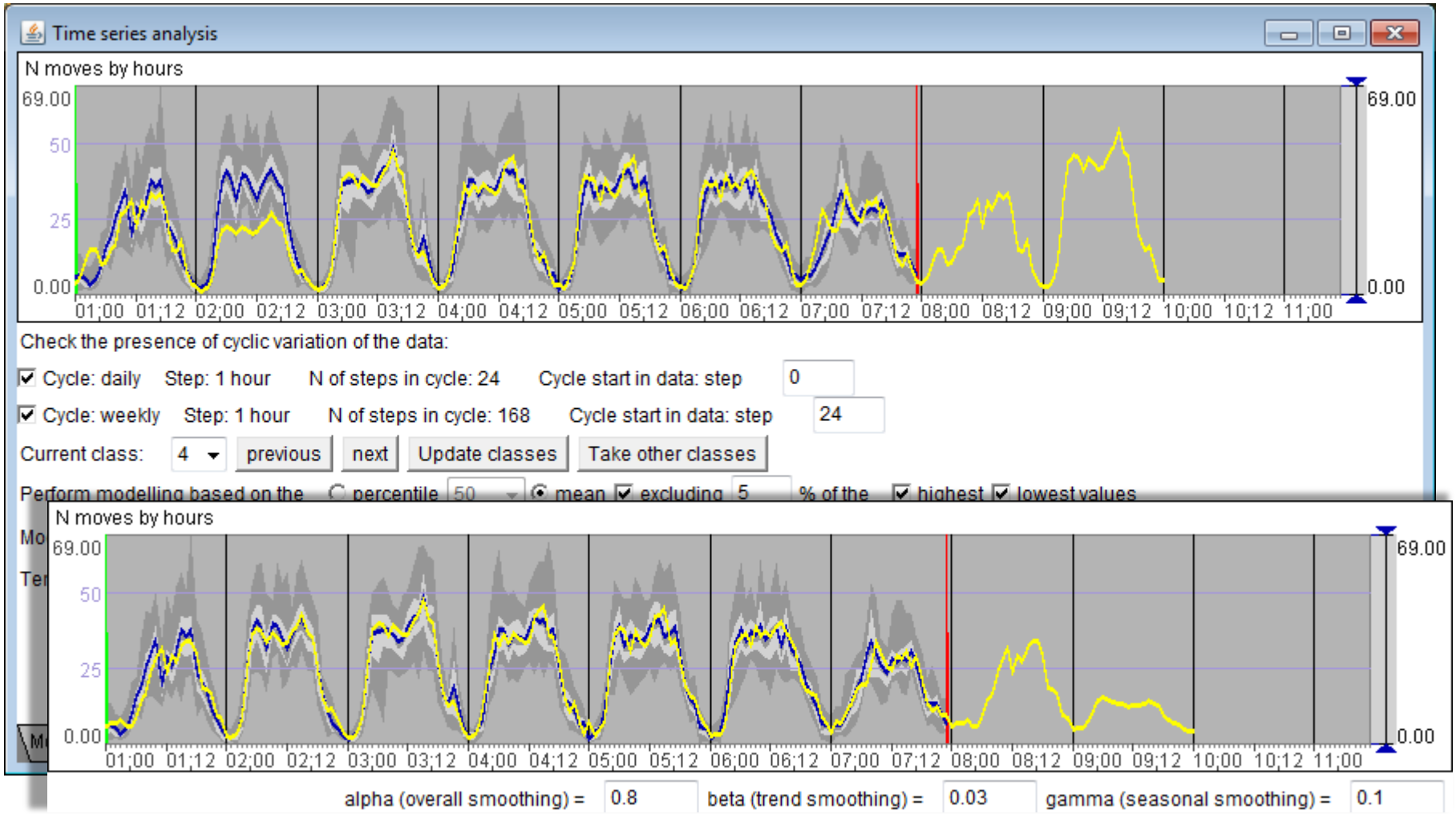




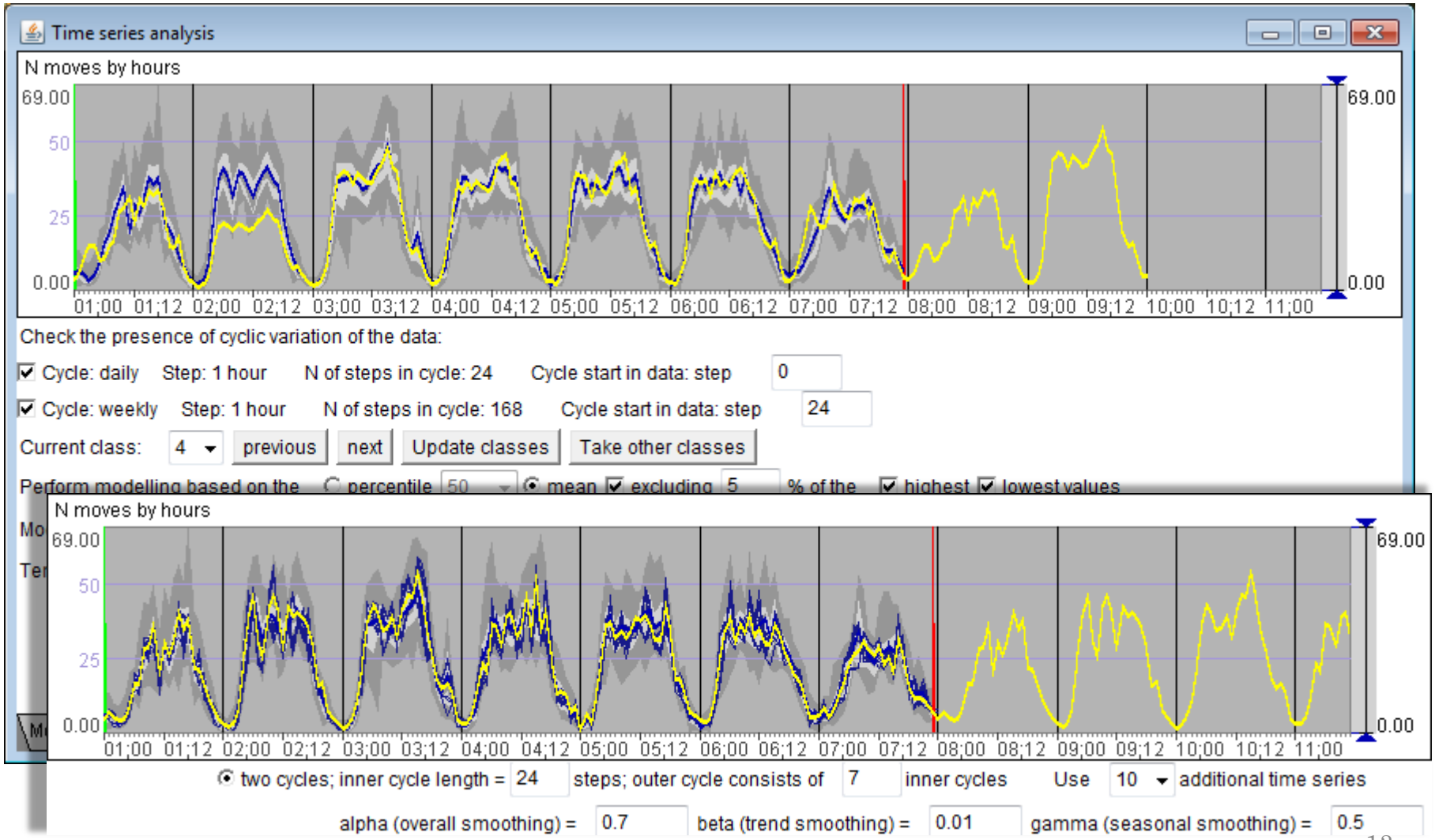
2) Cluster-wise time series modelling



2) Cluster-wise time series modelling



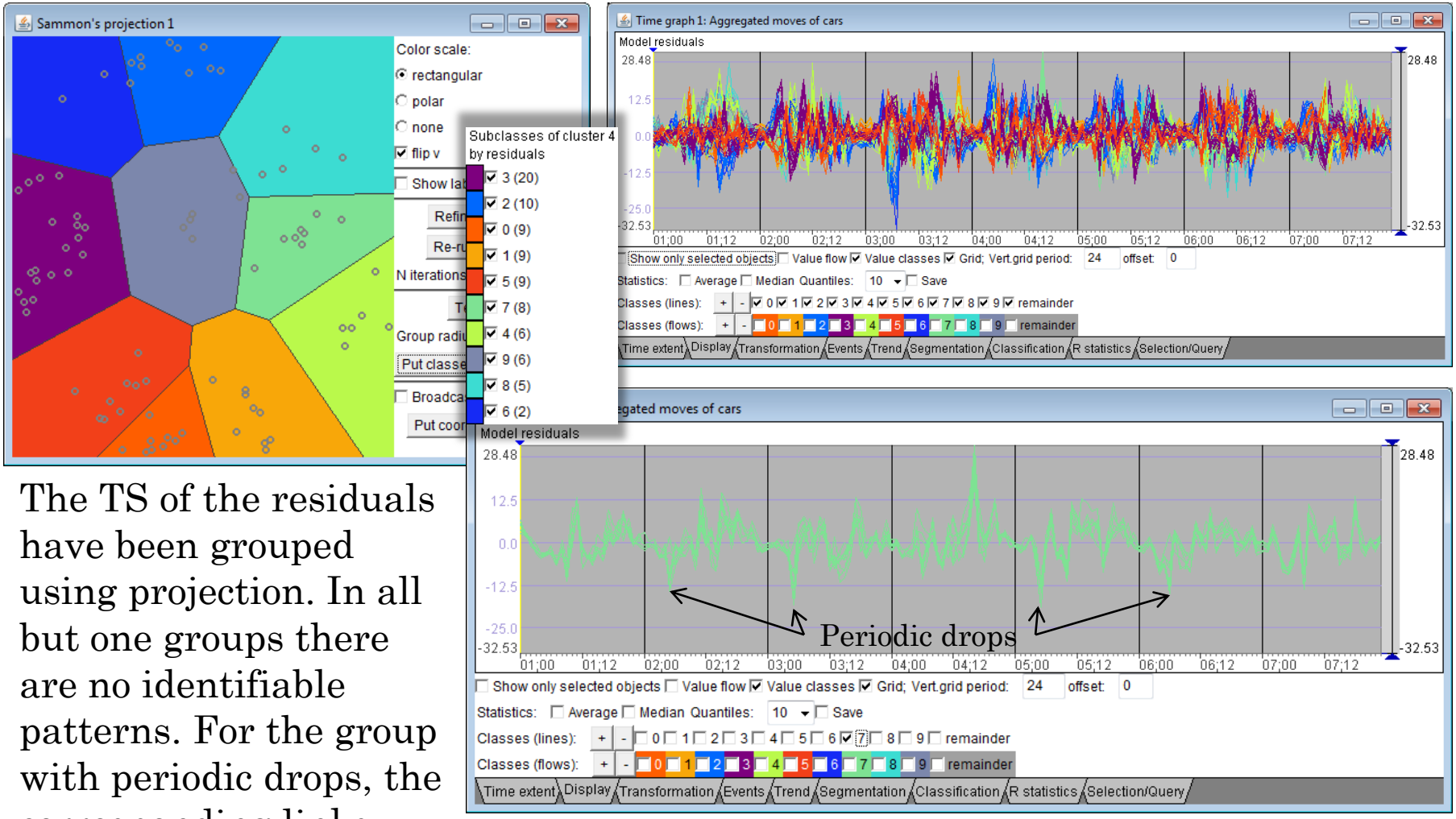
2) Cluster-wise time series modelling



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

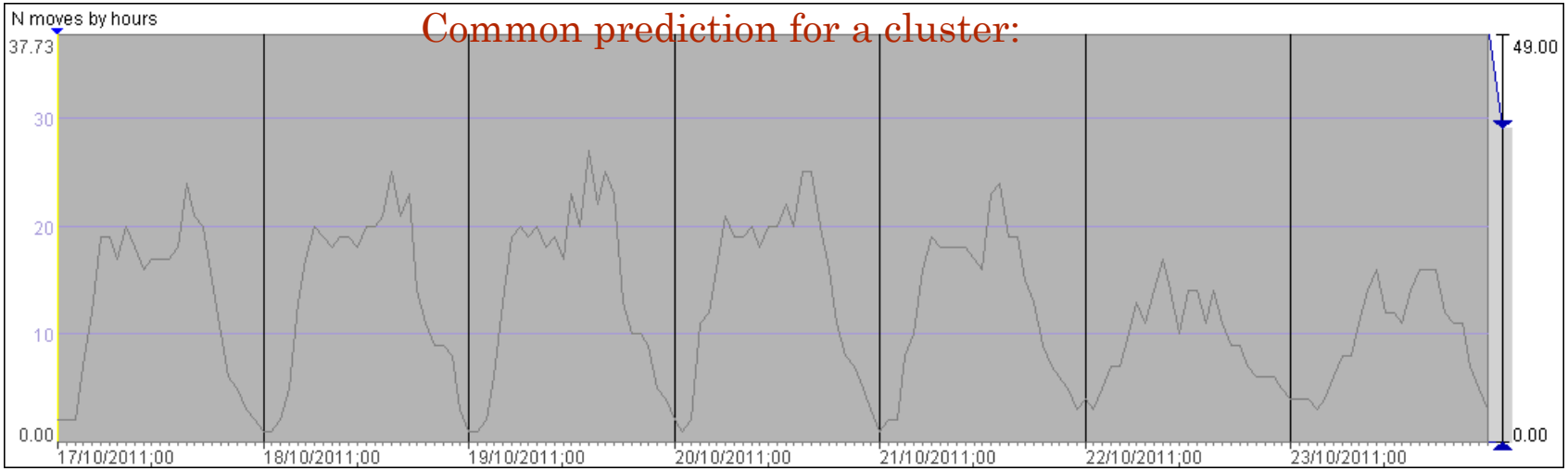
- For each link i, compute and store the basic statistics (quartiles) of the original values: $Q1_i$, M_i, $Q3_i$ (1st quartile, median, 3rd quartile)
- Compute the basic statistics of the model predictions for the whole cluster: Q1, M, Q3 (common for all cluster members)
- Shift *(level adjustment)*: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

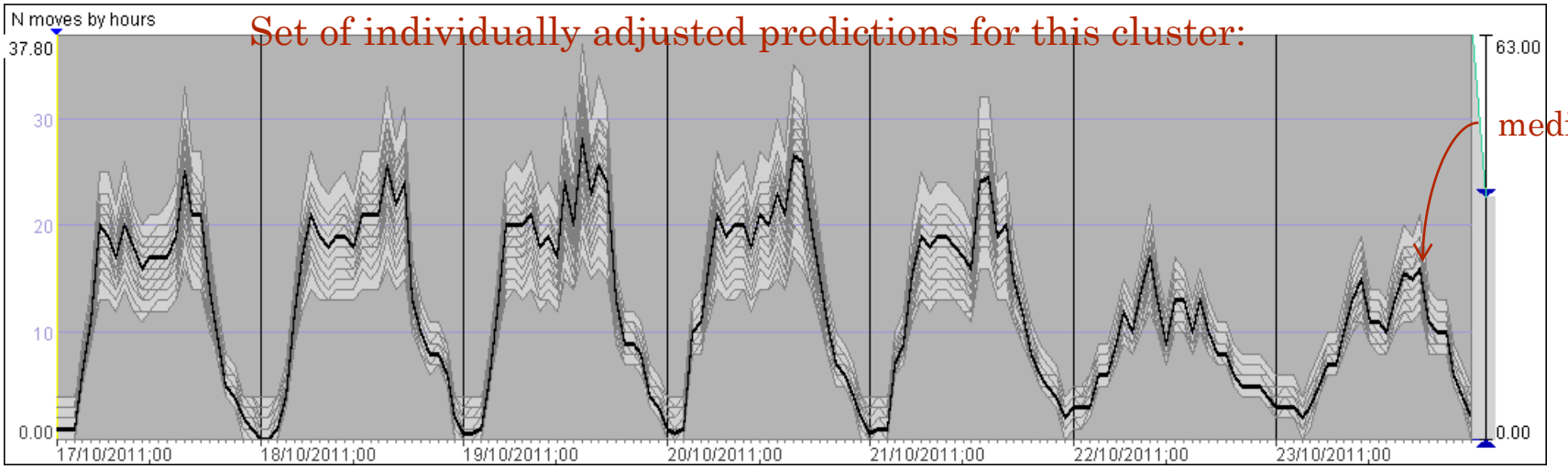
$$F_i^{low} = \frac{M_i - Q1_i}{M - Q1} \qquad F_i^{high} = \frac{Q3_i - M_i}{Q3 - M}$$

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

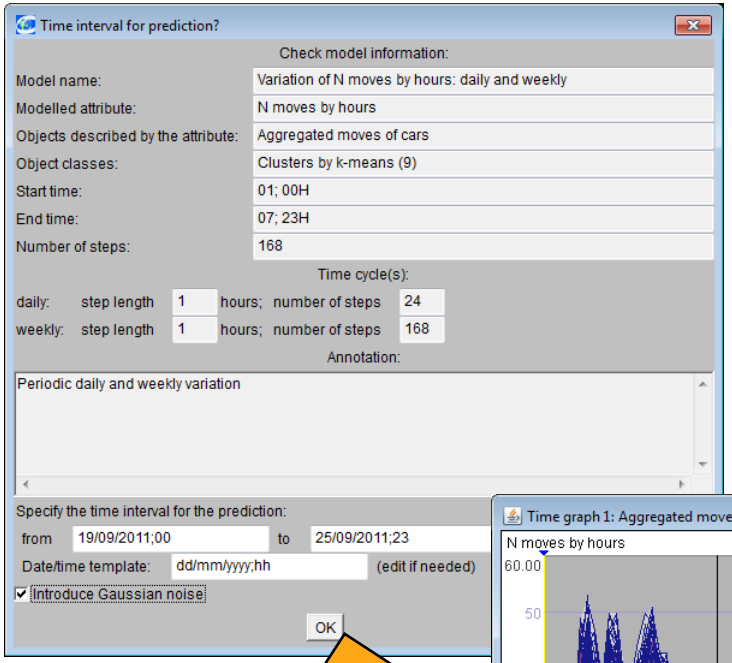
$$v^t_i = \begin{cases} M_i + F_i^{low} \cdot (v^t - M) + S_i, \text{ if } v^t < M \\ M_i + F_i^{high} \cdot (v^t - M) + S_i, \text{ otherwise} \end{cases}$$

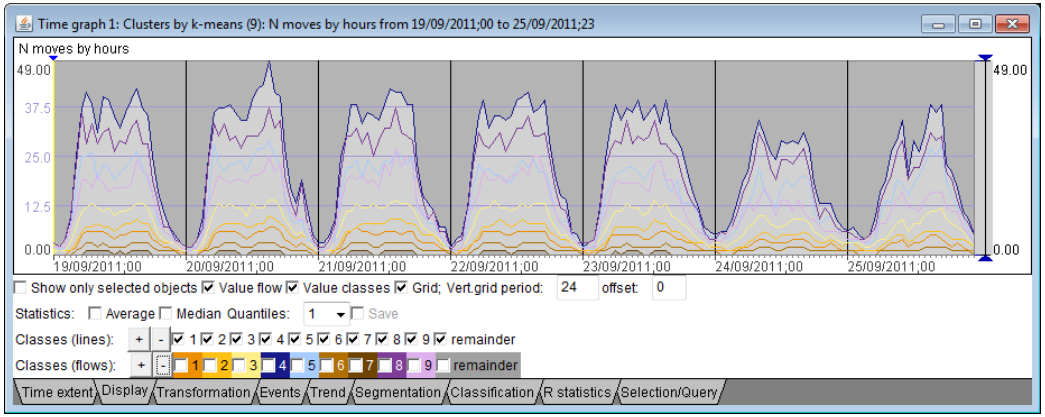
Example of individual adjustment

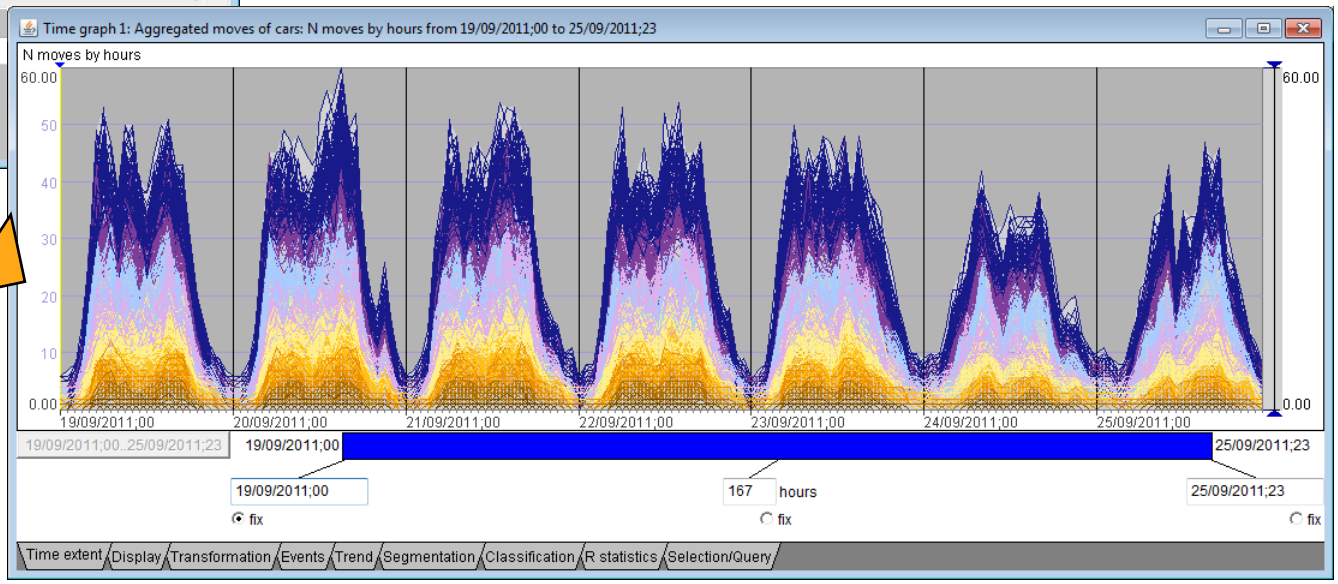




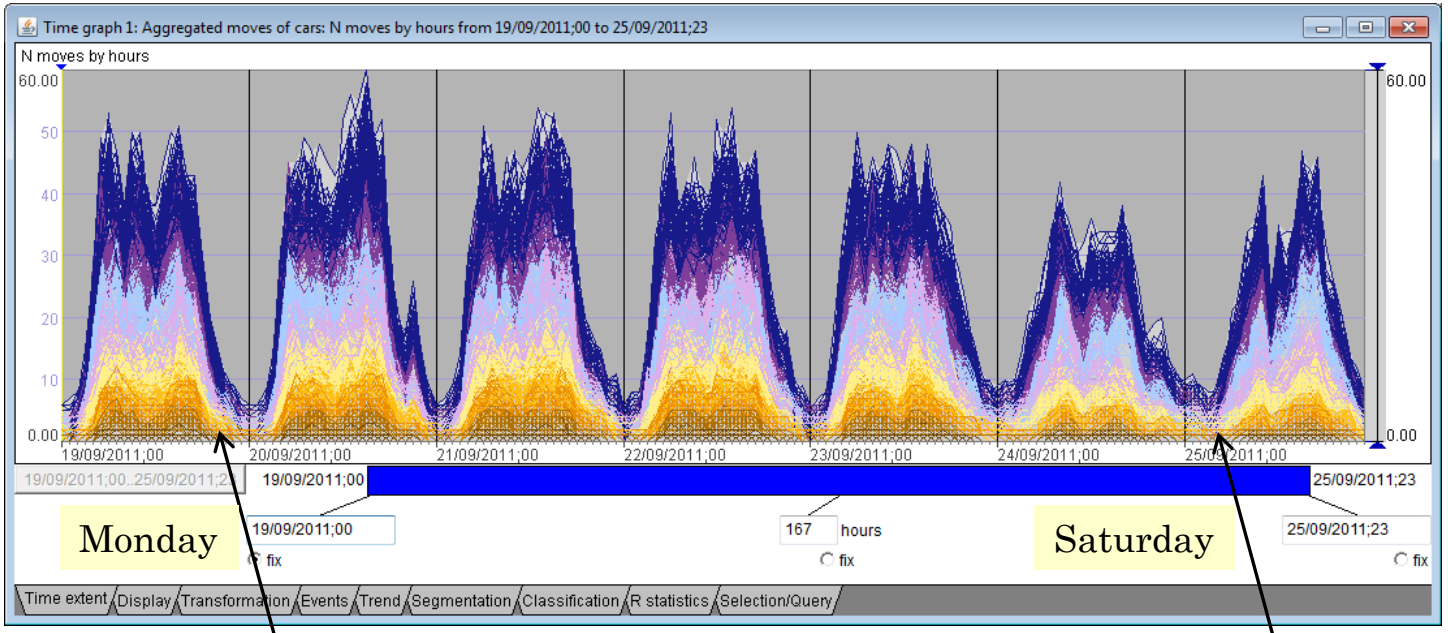
Example of prediction



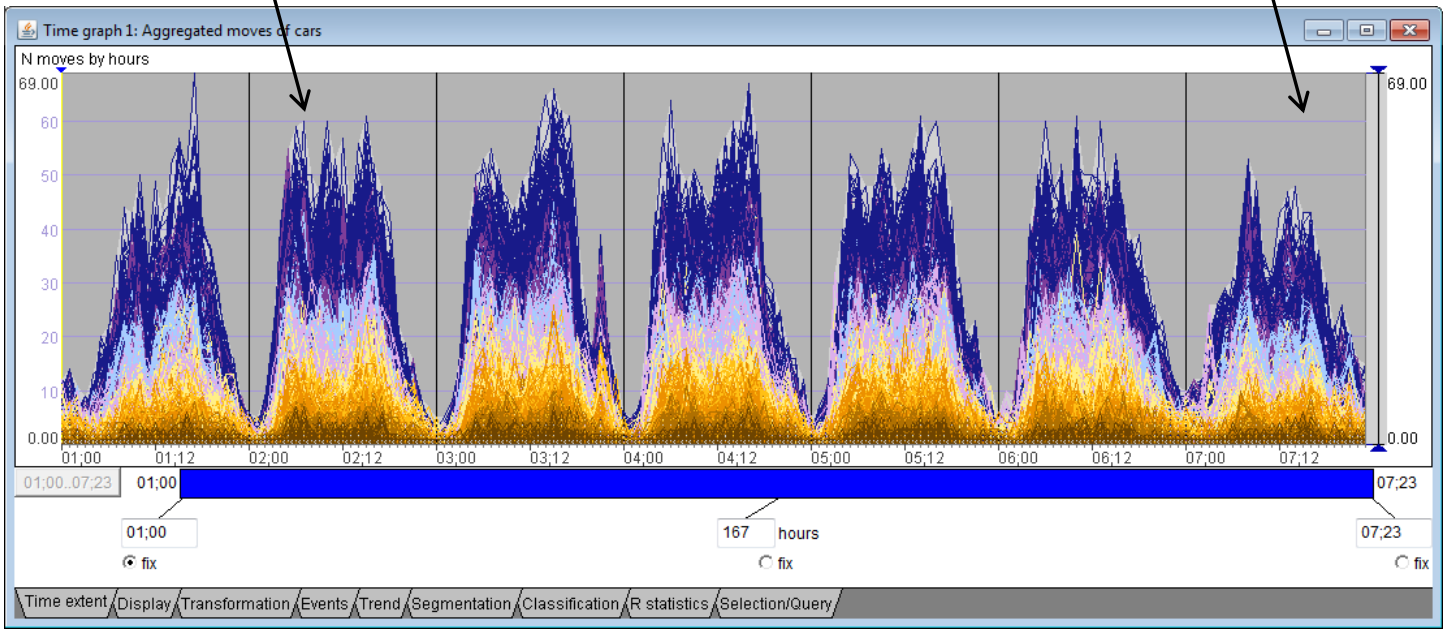




Predicted:



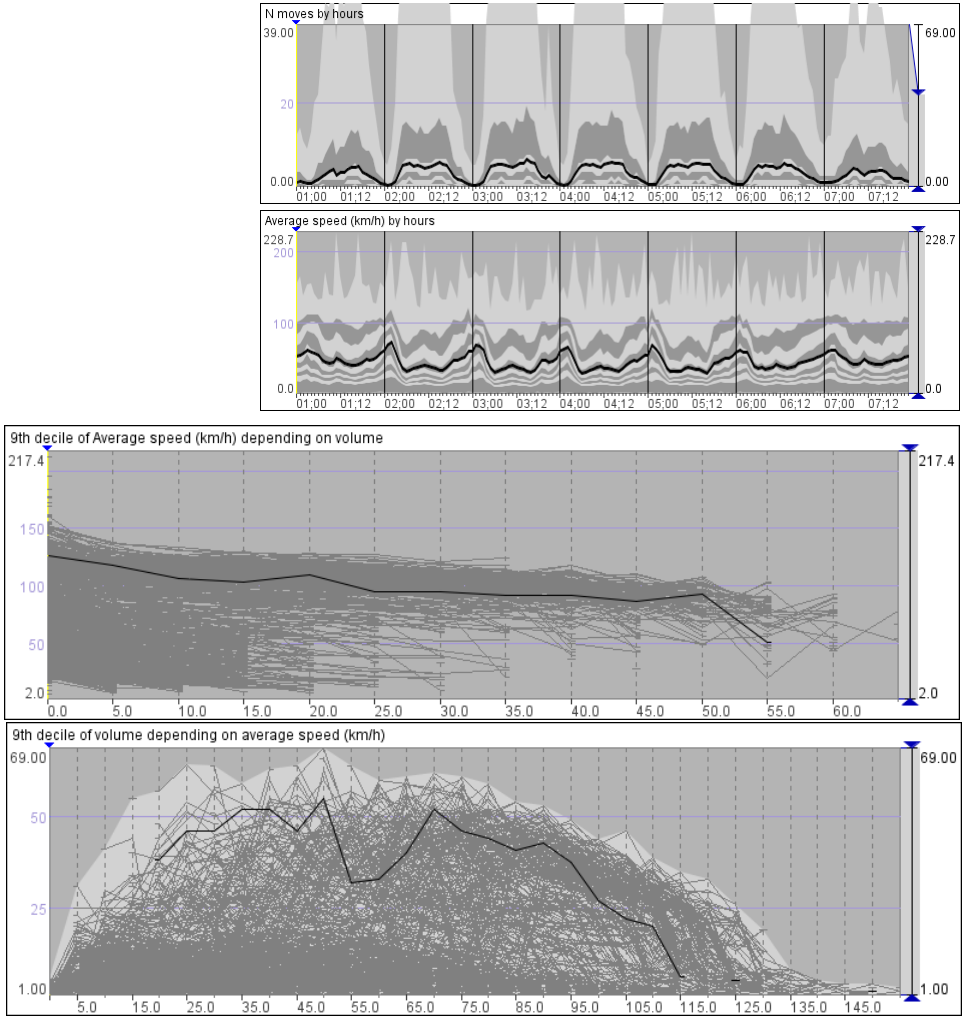
Original:



Prediction of extraordinary traffic flows

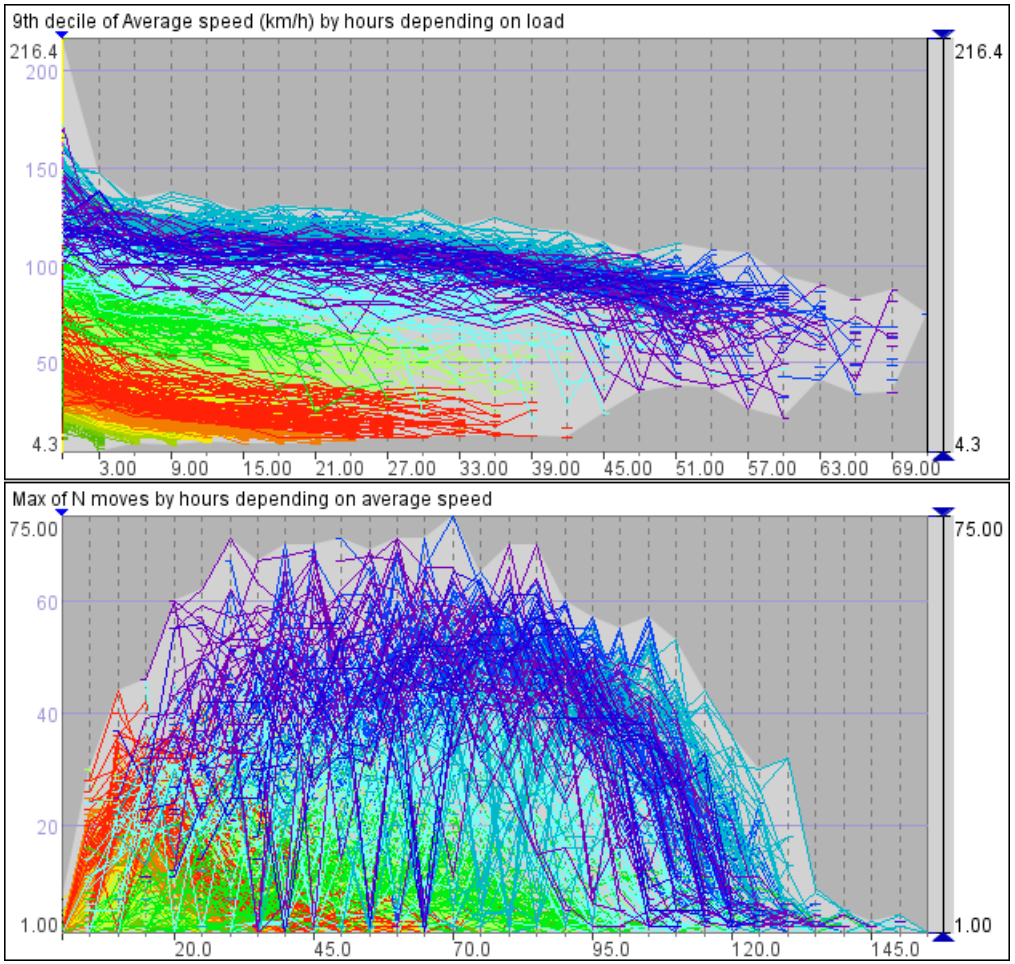
Volume-speed interdependencies

- 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.

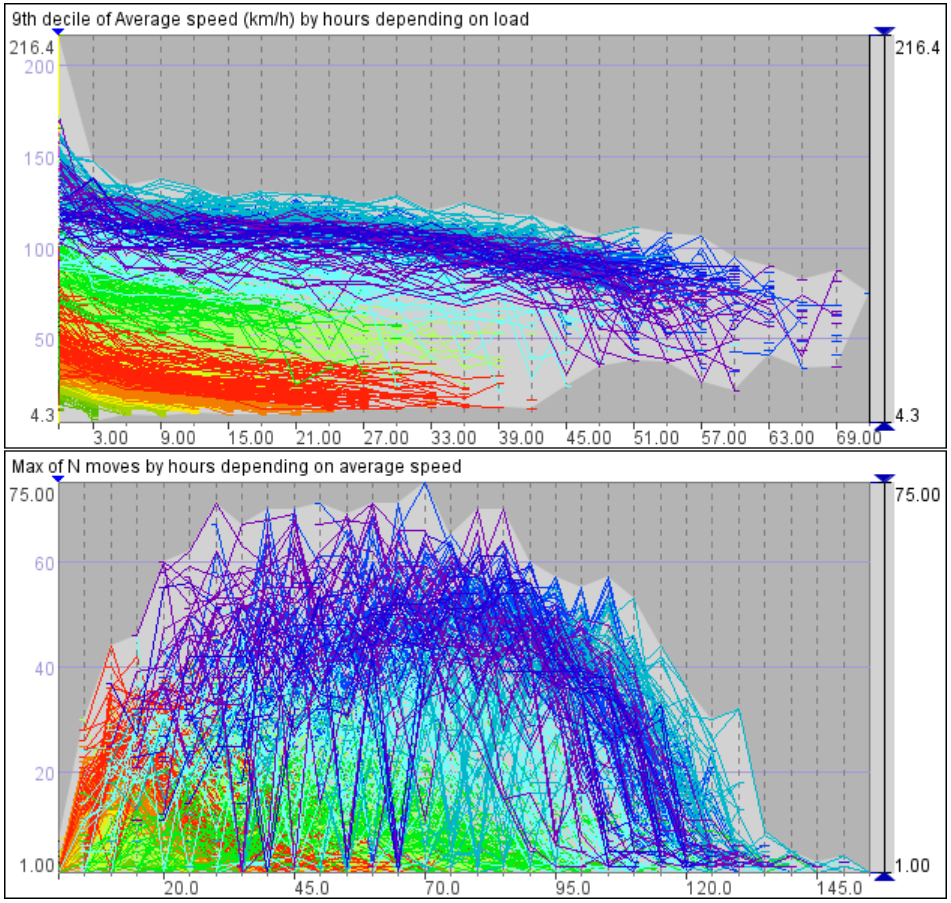


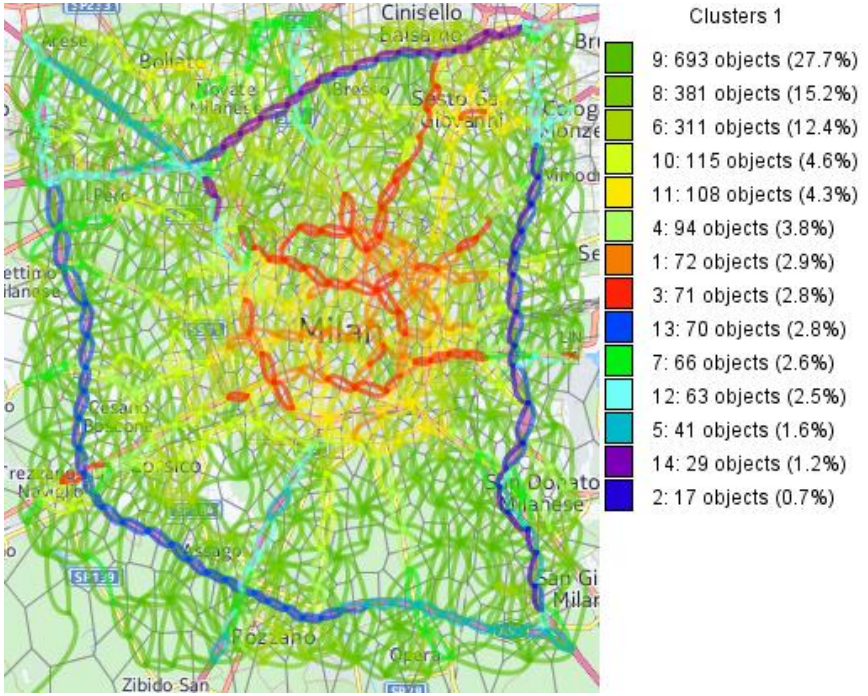
1) Data transformation

- 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)

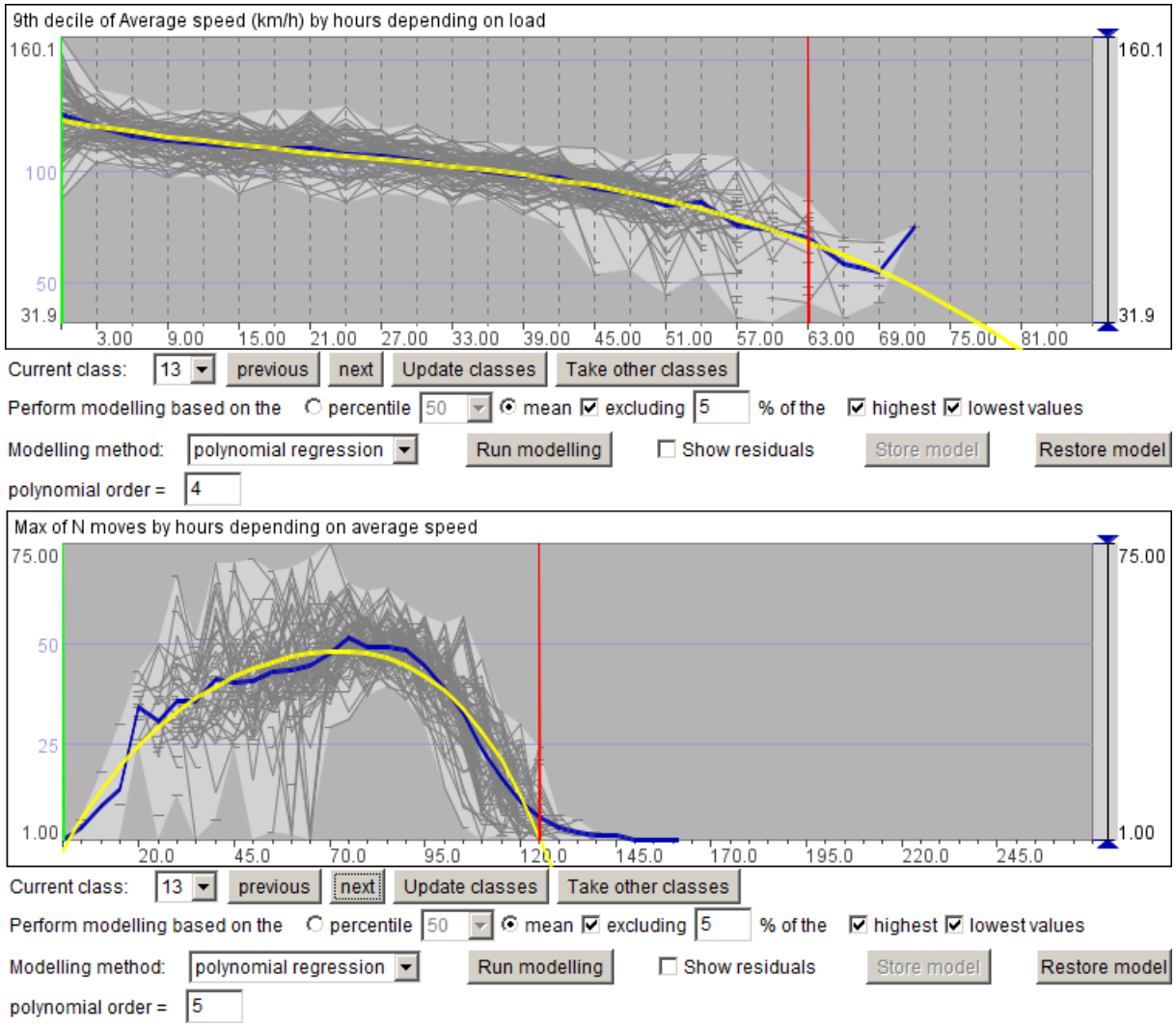


2) Partition-based clustering of the links by the similarity of the speed-volume dependencies



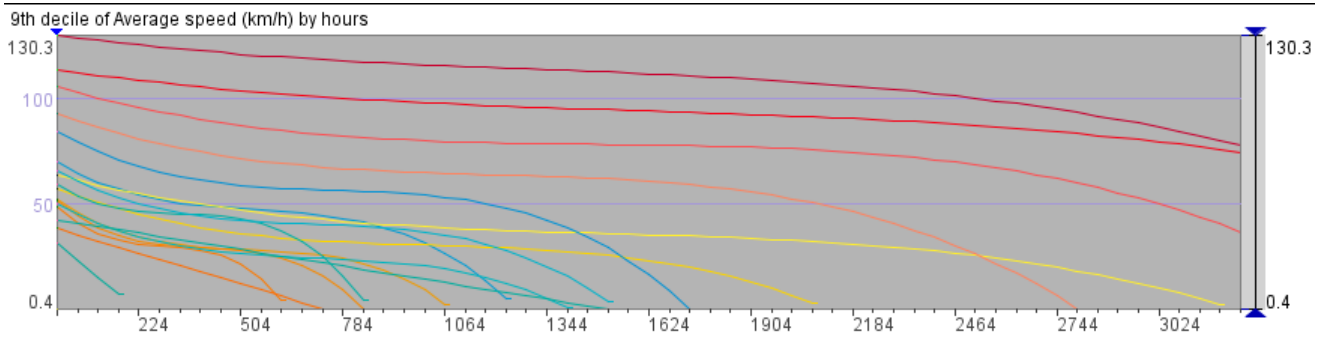


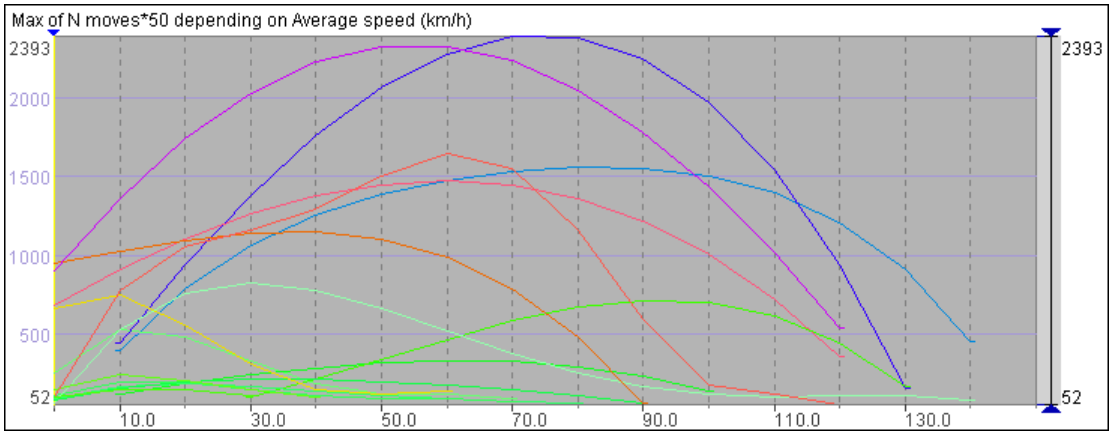
3) Representing the interdependencies by formal models



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.

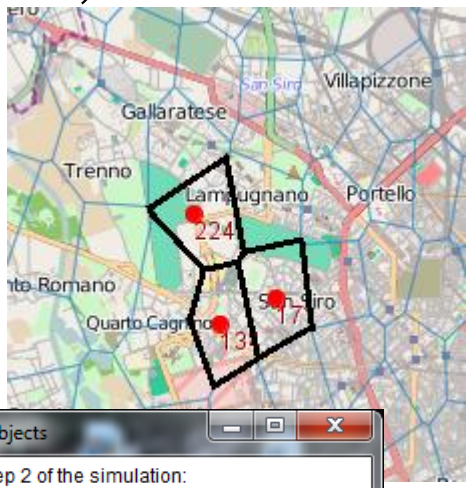
⇒ The aggregation of the original dataset does not give the true flow volumes for the links but about 2% of the true volumes. To obtain more realistic flow volumes, the computed volumes need to be multiplied by 50.**

** 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:

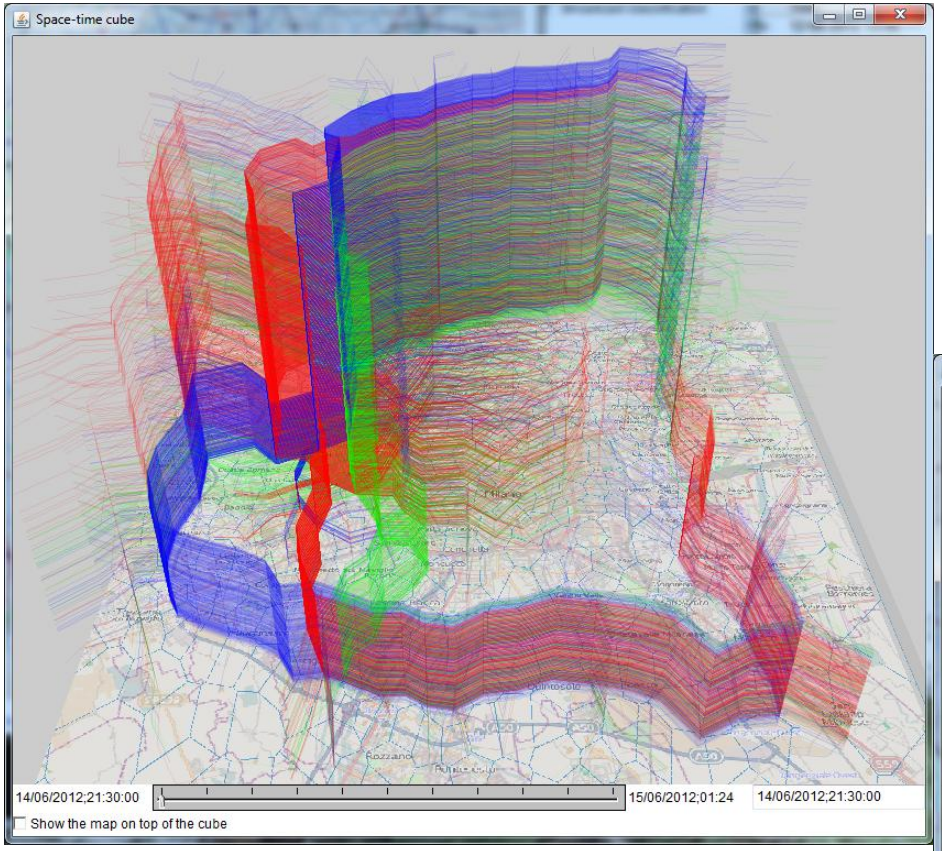
In place 171: 3000
In place 134: 4000
In place 224: 3000

Localize the places on map

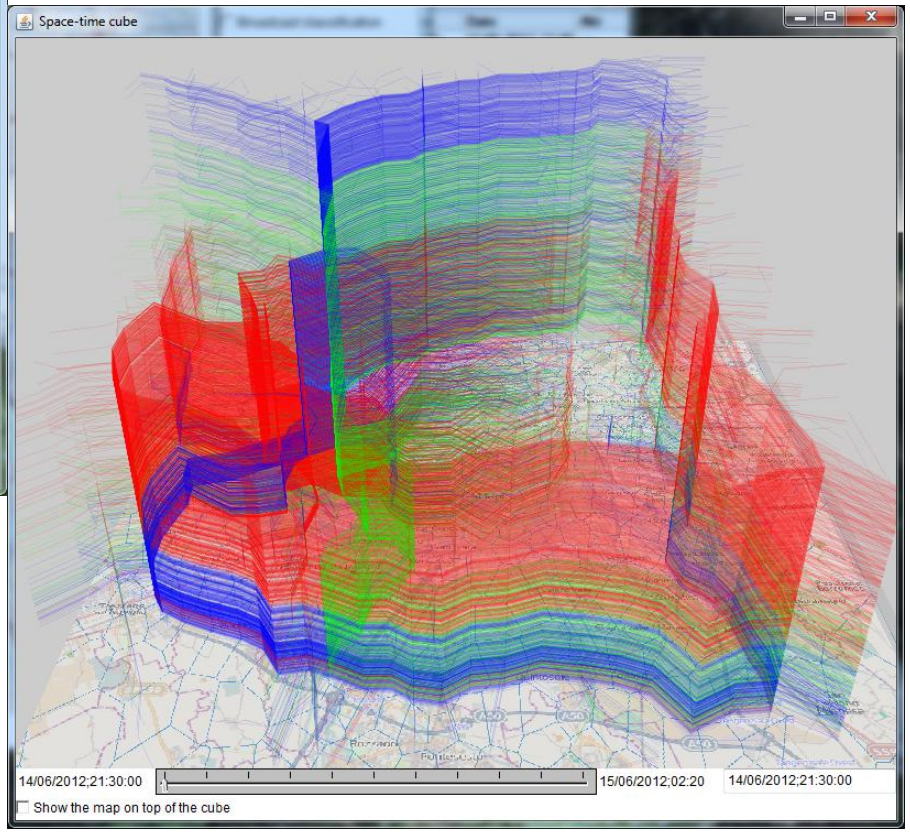
The given number of objects will be distributed among the 3 selected places of origin.

Continue Stop the process

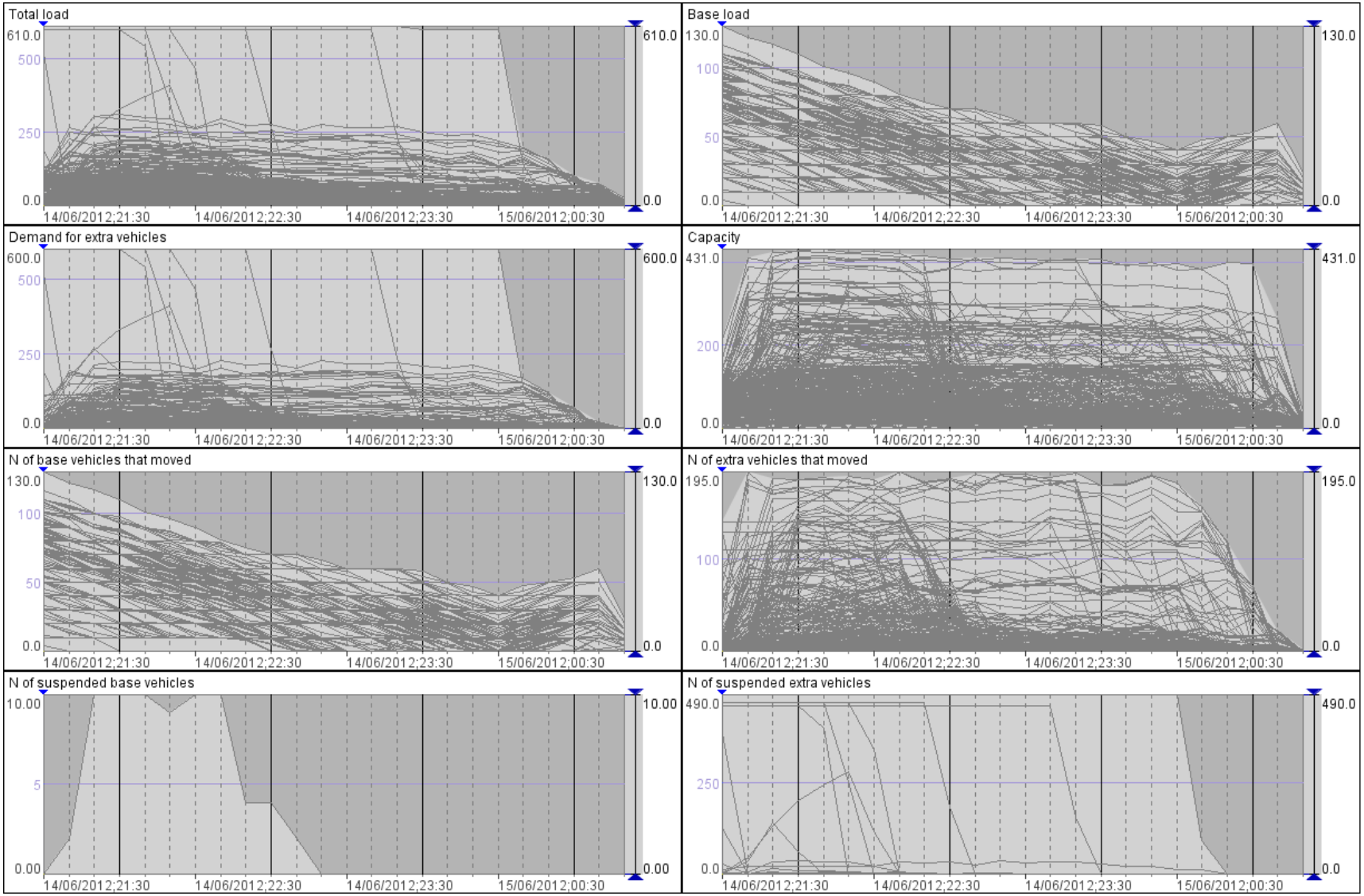
Simulated trajectories



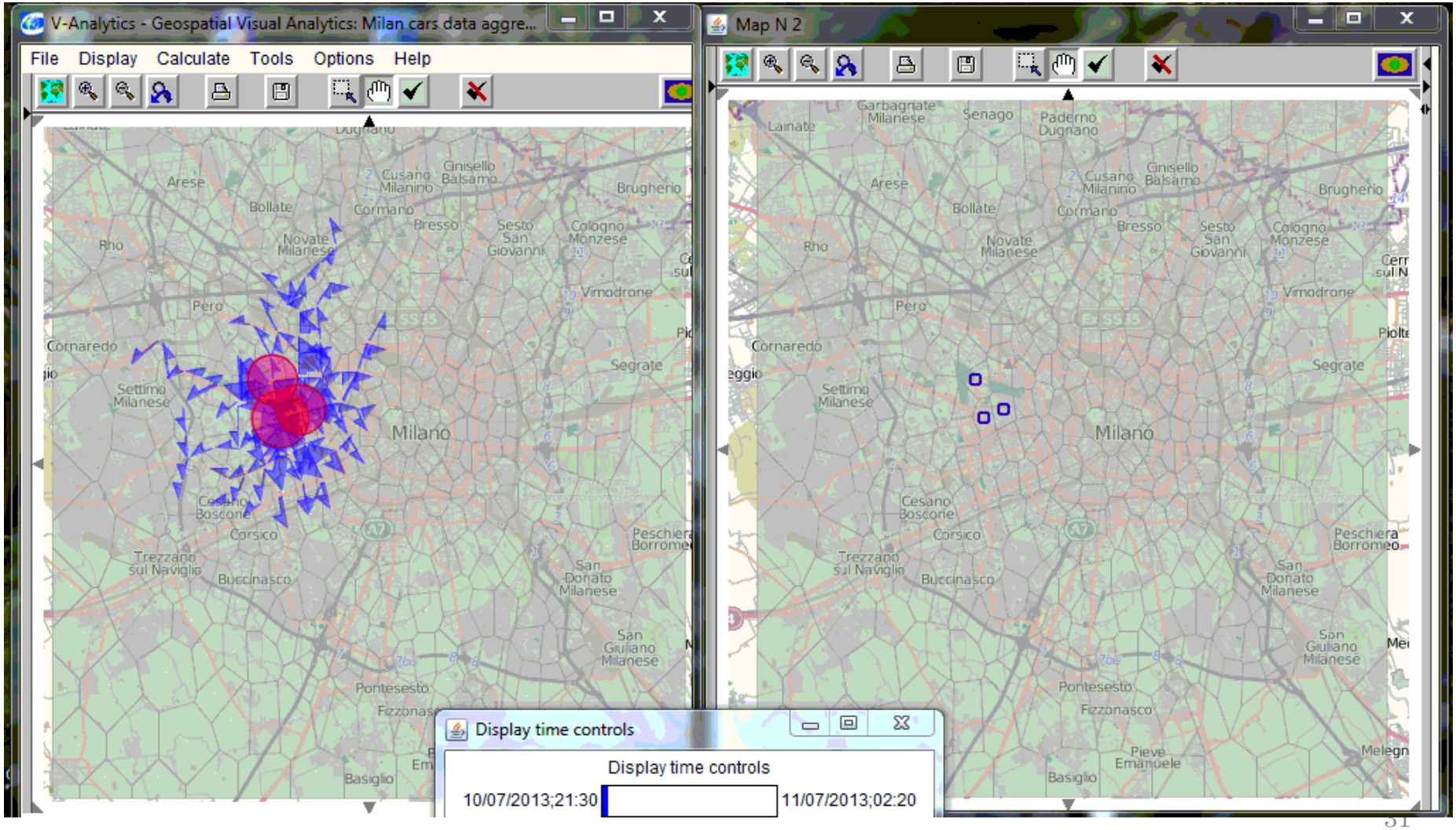
What will be the effect of re-routing a part of the traffic to the south?



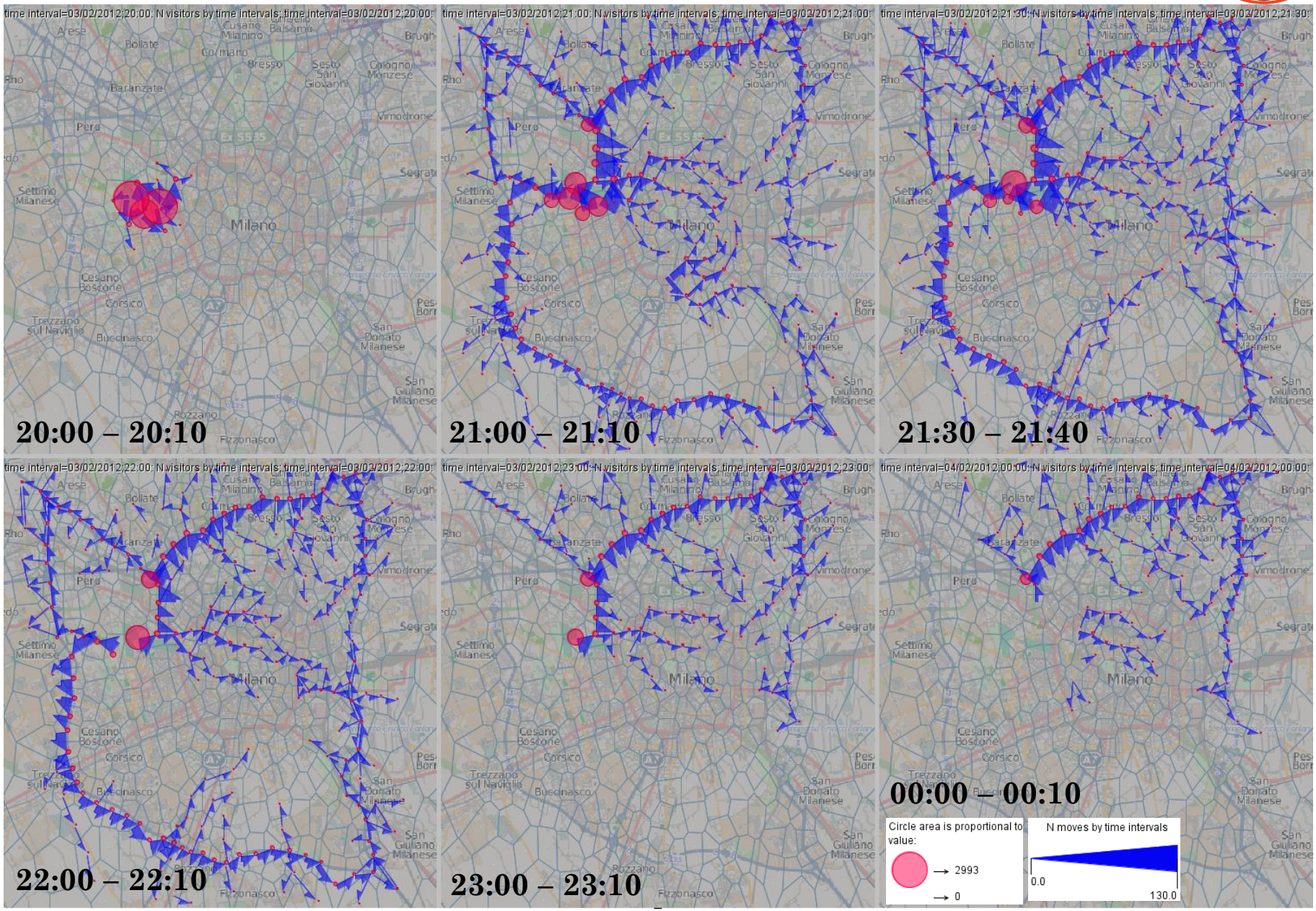
Aggregated representation of simulation results: time graphs



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

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