Extract space-time dynamics from sensor network to build urban traffic prediction model: a machine learning point of view

Serge Fenet and Yannick Perret and Julien Salotti

Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
serge.fenet@liris.cnrs.fr (corresponding author) — yannick.perret@liris.cnrs.fr — julien.salotti@insa-lyon.fr

1 Introduction and context

Urban transport, while being essential for citizens to perform their daily activities, also constitutes one of the major sources of urban pollution (global emissions, local air quality, noise) and physical accidents (the two thirds of them taking place in cities), directly affecting humans health. Moreover, the economic impact of traffic jams on goods transportation within cities reduces the efficiency of urban delivery process and impair its' economical aspect. In spite of great efforts made to favor modal report from car to other public transportation systems, the quest for an environmentally sustainable urban transport is a common and urgent challenge for all major cities in Europe. In these circumstances, and in order to (i) tackle the challenge of sustainable urban mobility, (ii) efficiently implement decision support tools, and (iii) exploit data allowing the assessment of policies and their resulting effects, urban planners need models.

Along the years, different types of urban models have been developed, ranging from the static and aggregate land use-transportation interaction (LUTI) models , to more recent bottom-up, activity and agent-based simulation. In the context of the Optimod'Lyon project, this work focuses on building a traffic prediction model of road traffic within the city of Lyon by using the Grand Lyon historical measures from more than 650 sensors between the year 2007 and now. This global model will be composed of a predictive model (gray box model that preserves dynamics) and an explanatory model (white box model that explicitly identify space-time features).

This paper aims at presenting the ongoing process of exploration of these data and the building of these models. Each section of this abstract highlight a potential barrier, and quickly presents the possible tracks a modeller could use to overcome them. The main conclusion we wish to convey is that, as almost every question is a research field in itself, important choices have to be made at every step, and there can be no universal model. The space available being short in this abstract, we focus here on textual description rather than graphical illustrations that could better highlight the arguments.

2 Data features

With 654 sensors gathering 240 measures of car count and lane occupation rate per day during more than 1700 day, a final amount of more than 540 million measures is both a blessing and a curse. These data contain vast amounts of information that we are still extracting, but also exhibit both noise and missing values as sensors can be faulty or suffer failure. Depending on the size of missing values blocks, such problems can be solved with interpolation (on short scale), multiple imputation (on larger scale), or by using clustering algorithms and partial deletion in order to reduce the available data to a dataset having no missing values. Artificial noisiness of data is more treacherous than missing values as it is much harder to identify: absurd values can be easily detected by analyzing anomalous deviation from a dynamically learnt profile, but abnormal plausible measures are much harder to identify.

3 Space distribution

Traffic data are the results of the dynamics of the city as a complex unknowable system, and thus exhibit both space and time components. These two components can be used as main exploration direction by using different tools.

3.1 Sensor clustering

Road sensors are distributed among road segments all through the city, and pick up different signals: some of them are located on wide circulating segments, others are situated on narrow street. They can be situated near important crossing points, or along continuous avenues. This heterogeneity leads to different data dynamics, that can be used to cluster the sensors. The idea is to group road segments sharing the same dynamics (same occurrences of the variation of a measure) independently of the measured value. To do so, we first convert time domain data to frequency domain data by computing the Discrete Fourier Transform of the time series of each sensor, then feed this frequency representation to standard clustering methods. In this case, the optimal number of cluster K is firstly computed by using the Mean Shift algorithm, then the algorithm K-Means is used with the optimal value of K (that can be proved by computing standard clustering quality indices like the 'within cluster sum of square'). It is also possible to sweep different value of K and use expert knowledge to choose the value that matches the most with the expert perception.

3.2 Space influence on data dynamics

An important distinguishing feature of our data is the ever-present dependency between road segments: the current state of a given segment will not only be influenced by it's own past, but also by the past of neighbouring segments within a horizon that will grow with the passing time. This space-time dependency is an important trigger of the non-linear behaviour of road traffic, and the evaluation of the dependency of each sensor to the others is thus an important component of the model. Moreover, this information can help us to identify unnecessary sensors –those which signal can be fully reconstructed from distant sensors–, or allow us to build 'virtual sensors' providing economical alternatives to unpractical or costly measurements.

A way to discover such cross-correlations is to use Dynamic Bayesian Networks. This probabilistic directed acyclic graphical model, whose edges represent conditional dependencies, will identify the minimum set of sensors necessary to explain the data and their probabilistic relationship. This graph can then be compared with the real distance graph with matrices correlation test (like Mantel test) to explore the influence of spatial city structure on the traffic.

4 Time distribution

The signal of each sensor is generated by non linear dynamic equations, leading to time-changing variance, asymmetric cycles, varying thresholds and higher-moment structures. However, each signal exhibit a strong low-frequency structure along with higher frequency random noise which color still has to be determined. It is then possible to extract long term low-frequency patterns on the one hand, and a description of noise on the other. As colored noise is indistinguishable from chaos, we can use Discrete Wavelet Transform to extract high-frequency stochasticity description (particularly noise envelope) at different time scales from sensor data. Such a noise model can then be used to reinject high-frequency noise into low-frequency forecasts.

4.1 Non linear deterministic dynamical systems tools

Non linear times series analysis is an old field at the crossing of statistical physics and chaos theory. From a computer science point of view, these series can be represented in a multi-dimensional phase space rather than in the time or the frequency domain, expressing all properties in terms of local quantities and space neigbourhood relations. This paradigm shift pave the way to many tool allowing to explore stationarity (with recurrence plots and stationarity tests), graphical embedding in low-dimension representation (with time delay coordinates, for example), non linear prediction (with threshold autoregressive model), non linear noise reduction (with phase space projection), and the creation of surrogate data (with constraints randomisation) when we have to deal with missing values. All these tools can help us to, first, distinguish chaotic time series from purely random data, and secondly assess the fact that the data are the product of a deterministic system that can be modelled.

5 Our current baseline: simple N-weighted learning from the past

Building a 'naive' baseline prediction method is an important starting task: it allows to have a set of reference results to compare to more sophisticated models, and so to evaluate the effective gain that must stem from the increased complexity. We implemented a N-weighted learning from the past method based on the idea that traffic behavior at a given time is often similar to traffic that occurred in the past and should then lead to a similar evolution in the near future. This method defines a distance measure on near past events based on a N-weighted calculus. It searches in the past the situations most similar to the current situation. It then takes the M most-fitting situations and uses the corresponding 'futures' to build prediction values for the current time. This set of situations is also used to compute an estimation of the variance of past data, allowing to also predict an expected variance around our predictions. This method has several parameters that need to be tuned: values for N and M, weights for the distance, weights for the combination of the M situations to create the prediction... Traffic can have very different dynamics depending on the day of the week, the current time in the day, the weather, etc. So, we trained our system to learn the best set of parameters for each of these different influences, giving us a large set of dedicated parameters to use in different situations. As explained previously, traffic data has both low-frequency (tendancy) and high-frequency (variability) component. That's why our method works on low-pass filtered data: it predicts the tendencies and gives an estimation of the noisy component, allowing to check if current data is within or outside the predictable variations. This latter case can thus be used as an indicator for an unexpected situation (road accident, etc.).

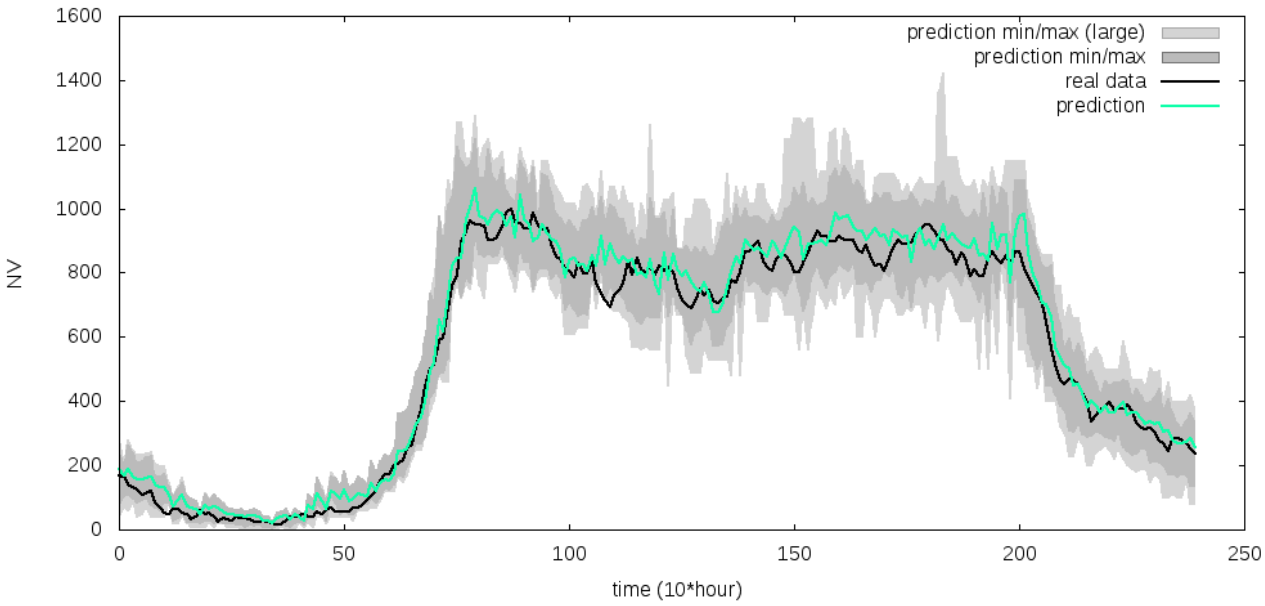

Fig. 1. Expressive power of naive baseline 'N-weighted learning from the past'

6 Conclusion and future work

This abstract has very briefly presented the directions we are currently exploring in order to build models describing and explaining the dynamics and stochasticity of the Lyon road traffic by focusing on different features of the available data. The firsts model are already being evaluated, and we plan to combine their multiple predictions by using machine learning methods like bagging (averaging predictions or majority voting from each model), boosting (developing multiple models in sequence by assigning higher weights for the training cases that are difficult to classify), or random forests. In order to evaluate their expressive power they will be compared with a naive N-weighted learning algorithm that we also presented.