A Data-Mining Approach to Travel Price Forecasting

Till Wohlfarth
Telecom Paristech / liligo.com
Paris, France
till@liligo.com

Stéphan Clémençon
Telecom Paristech
Paris, France
stephan.clemencon@enst.fr

François Roueff
Telecom Paristech
Paris, France
francois.roueff@enst.fr

Xavier Casellato
liligo.com
Paris, France
xavier.casellato@liligo.com

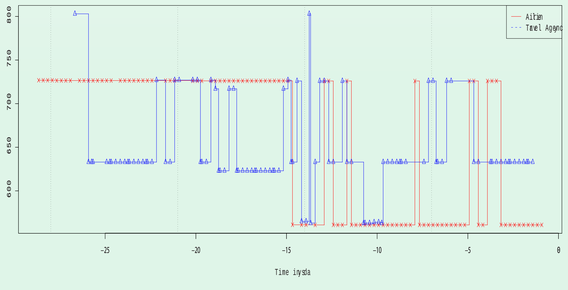
OBJECTIVE

Liligo.com is a real-time travel search engine that finds among more than 250 travel sites, agencies and tour operators all tickets that are available on the web market place for a particular travel search. The goal of our project is to consider the design of decision-making tools in the context of varying travel prices from the customer's perspective.
Based on vast streams of heterogeneous historical data collected through the internet, we describe here two approaches to forecasting travel price changes at a given horizon, taking as input variables a list of descriptive characteristics of the flight, together with possible features of the past evolution of the related price series. Though heterogeneous in many respects (e.g. sampling, scale), the collection of historical prices series is here represented in a unified manner, by marked point processes (MPP).

DATABASE

The (sampled) price series are observed through liligo.com's historical data repository, where all user search results are stored. Focus is here on 6 routes collected on 9 flight tickets providers. We consider trips of 3, 7 or 14 days to cover the most common length of stays based on liligo.com statistics. The routes selected are typical examples of the European flight travel market.

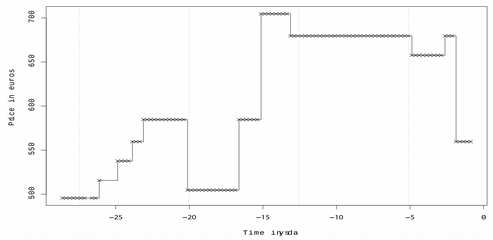
A database item contains a set of features either read from the historical data (departure station, arrival station, departure date, return date, supplier, ...) or computed from additional information such as past searches. For each database item, prices are sampled every 6 hours over a month before departure which leads to time series with more than one hundred sampling points.



We temporarily exclude travel agencies time series because of a second yield management phenomenon which brings too much noise.

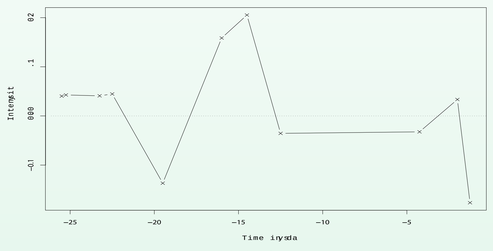
REPRESENTATION

As an illustration, consider the time series displayed in Figure (a). It contains the collected prices of a Paris-Marrakesh flight provided by a French low-cost, for a 3 days trip leaving on the 24th of October 2010. The observed trajectory is piecewise constant.



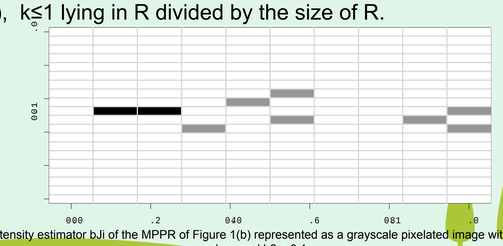
(a) Collected prices of 3 days long Paris-Marrakesh trip, '*' corresponds to observed points, plain lines represent the interpolated trajectory. Vertical dotted lines are displayed every 7 days from the departure date.

Piecewise constant trajectory can be entirely reconstructed from the initial price at time T_{0-28} and the sequence of change points in the series. It is in fact sufficient to have the sequence of change points instants with the corresponding relative jumps (also called the price returns)(Figure (b)). As a consequence each item i in the data base is assigned an initial price $P^{(i)}$ and a marked point process of returns (MPPR) $N^{(i)}$ that allow to construct the whole interpolated price trajectory with a minimum set of values.



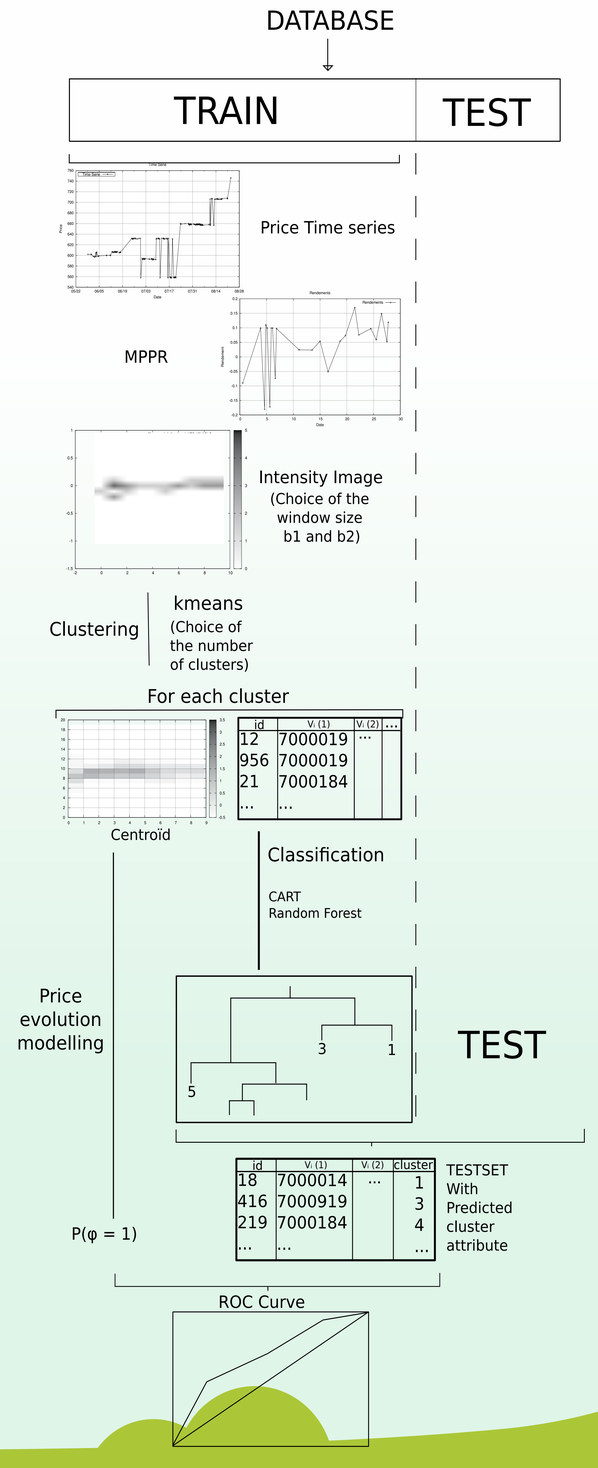
(b) Marked point process of returns (MPPR) of the time series displayed in Figure 1(a).

We will finally model the MPPR's $N^{(i)}$ as an inhomogeneous Poisson point process. Such processes are parameterized by a (non-negative) intensity function J_i on the plane that can be interpreted as follows: the probability of having a jump in the interval [t, t + dt] with return value in [s, s + ds] is approximatively given by $J_i(s, t)dtds$. A simplified way to parameterize $N^{(i)}$is to use a finite dimensional representation of Ji using a pixelated image of the rectangle $[T_0-28, T_0]\times[-1, 1]$. Each pixel corresponds to a constant value of Ji on a rectangle cell/pixel R of size $b_1 \times b_2$. This value can be estimated as the number of jump points $(T_k^{(i)}, s_k)$, $k \leq 1$ lying in R divided by the size of R.

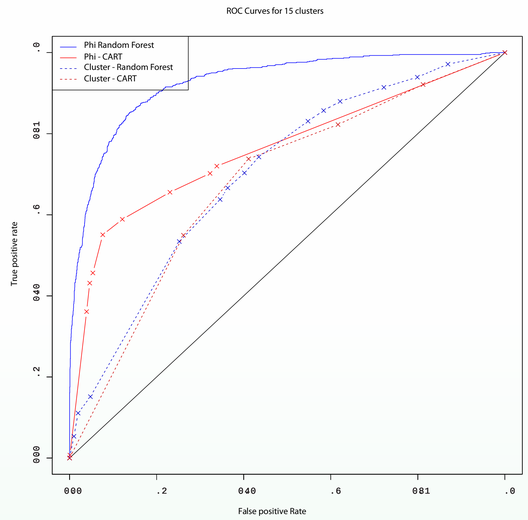


(c) Intensity estimator bJi of the MPPR of Figure 1(b) represented as a grayscale pixelated image with b1 = 3 days and b2 = 0.1

PREDICTION



RESULTS



The ROC curves obtained from the model free method (Phi) is displayed with the cluster based approach. As expected the direct method performs much better as its goal is concentrates on a specific event prediction task rather than a complete model of price time series. We believe that the direct method should be used for pre-registered purchase periods to give a first coarse advice to the customer. The model based one can then been used to provide a more specific, yet less reliable advice adapted to the needs of the customer.

With the random forest algorithm, we can see the importance of each attributes in the classification process. We will be able to give meaningful interpretation of the discriminant attributes. The most important attributes are the the temporal ones (departure day, return day, day of the year, day of the week...).
As we add the first evolutions information, we notice the importance of the first cells of the grayscale in the cluster classification. Obviously the majority of the cells (always equal to zero) will not be useful, but the low variation cells have a significant importance that will give us better result on the attribute based classification.

Now we have to improve the cluster based prediction and confront it to complete new flights, including the ones from travel agencies.

References

S. Daudel and G. Vialle, Yield management : applications to air transport and other service industries. P.I.T.A., 1994.
T. Hastie, R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning, corrected ed. Springer, July 2003
D. J. Daley and D. Vere-Jones, An introduction to the theory of point processes. Vol. I, 2nd ed., ser. Probability and its Applications (NewYork). Springer-Verlag, 2003