The Canadian Journal of Statistics
Vol. 31, No. ?, 2003, Pages ???-???
La revue canadienne de statistique

Road trafficking description and short term travel time forecasting, with a classification method

Jean-Michel LOUBES, Elie MAZA and Marc LAVIELLE

Key words and phrases: forecasting method; functional classification; learning theory; mixture model. .
MSC 2000: Primary 62H30; secondary 62P30.

Abstract: The purpose of this work is, on the one hand, to study how to forecast road trafficking on highway networks and, on the other hand, to describe future traffic events. Here, road trafficking is measured by the vehicle velocities. We propose two methodologies, the first one is based on an empirical classification method, and the second one, on a probability mixture model. We use an SAEM type algorithm (a Stochastic Approximation of the EM algorithm) to select the densities of the mixture model. Then, we test the validity of our methodologies by forecasting short term travel times.

Title in French: Description de trafic routier et prévision de temps de parcours à court terme, avec une méthode de classification

Résumé: Les objectifs de l'étude exposée ici sont, d'une part, la mise en place d'une méthode de prévision de temps de parcours sur le réseau routier de l'agglomération parisienne, et d'autre part, la description des comportements futurs du trafic. Ici, le trafic routier est mesuré par la vitesse des véhicules. Les auteurs proposent deux méthodologies, l'une est basée sur une méthode de classification automatique, et la seconde, sur un modèle de mélange. Afin d'estimer les paramètres du modèle de mélange, nous utilisons l'algorithme SAEM (une Approximation Stochastique de l'algorithme EM). Enfin, nous testons et comparons les méthodes proposées en effectuant des prévisions sur un échantillon de test.

1. INTRODUCTION

Even if long term road traffic forecasting was developed a long time ago (see http://www.bison-fute.equipement.gouv.fr), short term forecasting appeared only recently. Indeed, new technologies enable us to obtain precise data, not only a qualitative variable describing the state of a car stream: moving, blocking or stopped, as it is done by Couton, Danech-Pajouh & Broniatowsky (1996), but also a quantitative variable (with measures of speed, flow and occupancy rate), necessary for such a study.

In this work, our main purpose is to forecast travel time on the Parisian highway network. More precisely, we want to forecast, at time H, the time needed at time $H + h$, $h \geq 0$, to travel from one point to another. Contrary to previous works, see for example Van Grol, Danech-Pajouh, Manfredi & Whitakker (1998) or Danech-Pajouh & Aron (1994), where forecasts were made only at a specific point of the network, we aim at forecasting travel time, which implies estimating

speeds at all the points of the observation grid, i.e. all the measurement stations of the network. Moreover, our methodology builds archetypes of road trafficking whose interpretation is of great interest to study the different roadtrafficking behaviours. We point out that we did not consider time series, as done by Belomestny, Jentsch & Schreckenberg (2003), since the structure of our data prevents a wide use of these technics, as discussed later in the paper.

Our study relies on two commonly accepted assumptions. First, short term road trafficking mostly depends on what just happened. Second, there are a fixed number of traffic patterns, and every new observation day can be compared to them. The data used for this study have confirmed these statements. As a consequence, the issue of travel time forecasting can be divided into two steps. First, we estimate the representative behaviours or patterns of road trafficking. Then, we compare the incoming observations to these archetypes, and choose to which cluster this observation belongs to.

Functional data analysis methods are well suited to forecast outcomes made of functions and have been widely investigated over the last few years. Such techniques enable to fit a nonlinear model to the data, and then use this model to predict the forecoming values. For general reference, we refer for example to the following papers: Preda & Saporta (2004); Bosq (2003); Besse & Cardot (1996); Núñez-Antón, Rodríguez-Póo & Vieu (1999); Ferraty & Vieu (2003) or Ferraty & Vieu (2004). In this paper, we try to release too strong assumptions over the data and, for this, we focus on functional classification methods.

In our work, we compare two different ways of finding the features of road traffic. On the one hand, we aim at modelling road traffic with a mixture model, assuming that the daily evolution of the vehicle speed is drawn from a mixture of probability. So, it is necessary to estimate the components of the mixture, as well as the optimal number of components (see for example, Chen 1995; Lindsay & Lesperance 1995 or Cheng & Liu 2001). The components of the mixture are thus the archetypes we are looking for. Such method has already been used but only for a qualitative study of road traffic, by Dochy (1995) or Couton, Danech-Pajouh & Broniatowsky (1996).
On the other hand, standard classification methods enable us to allocate data into representative sets: see, for more general references, Gordon (1999); Celeux (1988); Breiman, Friedman, Olshen & Stone (1984) or Jambu (1978). Indeed, classification methods aim at gathering individuals into a restricted number of representative classes. Representative classes are such that two individuals taken inside the same class are *similar* (homogeneous class), and such that two individuals taken in two different classes are *distinct* (heterogeneous classes). The introduction of an appropriate distance index for speed curves (distance, dissimilarity index, variation, ultra-metric variation, etc.) will enable us to quantify the qualitative terms *similar* and *distinct*. So, the major part of the work here consists in finding a suitable distance and the optimal number of clusters used to summarize the information conveyed by the data. Then we extract the main feature from each cluster to obtain the archetypes.

The article falls into 7 main parts. Section 1 is the introduction. In Section 2, we describe the data used for this work as well as the preliminary treatments to detect and eliminate outliers. Then, we present the forecasting methodology. Section 3 provides a model for the vehicle speed change by considering a mixture setting. An SAEM type algorithm is used to estimate the different components of the mixture. Section 4 is devoted to the study of an empirical classification to construct significant clusters, and archetypes of each traffic behaviour are given in Section 5. In Section 6, we compare the two different approaches by forecasting travel times with the patterns obtained by the two methodologies. Finally, Section 7 is devoted to the conclusion.

2. DATA AND METHODOLOGY

2.1 Description

On the main roads around Paris, counting stations can be found, approximately at every 500 meters along main road axes. Such sensors provide the following observations, see Cohen (1990): the flow, the occupancy rate, and the speed, defined by the mean of vehicle speeds over a period of 6 minutes. Througout all the paper, we will use the following notations:

- Let C_s, $s = 1, \ldots, S$, be a counting station, where S stands for the number of stations on the network (actually $S \approx 2000$),
- Let J_n, $n = 1, \ldots, N$, be an observation day, where N is the number of days considered in the study.

The database used in the paper, was provided by the SIER (*Service Interdépartemental d'Exploitation Routière*) and is composed of the daily evolution of the vehicles speed over $N = 709$ days. For each station C_s and each day J_n, we observe $Y_n^s(t)$, $t = 1, \ldots, T = 180$, corresponding to the average speed over a period of 6 minutes, ranging from 5 AM to 11 PM, given 180 daily speed measurements per station, see Figure 1 for an example of such a speed curve.

Our study is carried out on a representative axis of Paris highway network (named A4W) where it is difficult to forecast travel times and which is known to be representative of Parisian road traffic behaviour. This road section is 21.82 kilometers long and has 38 counting stations.

2.2 Data quality

Rough data of the 2 years database cannot be used directly since aberrant and missing data are too numerous, due to the defaults of the counting stations. Hence, we provide a two step filtering and completion algorithm. The first step detects aberrant data, and was elaborated in collaboration with the SIER managers. The second step deals with the completion of missing data.

1. Aberrant data detection is based on the three following points:
   (a) detection of excessive speed measures, higher than 160 km/h
   (b) detection of too low speed measures, lower than 5 km/h during more 3.6 hours,
   (c) detection of constant speed measurements, constant for more than 0.5 hour,
2. Missing data completion is carried out by calculating a space and time average with non missing data: replace a missing data by

$$Y_n^s(t) = \frac{Y_n^{s-1}(t) + Y_n^{s+1}(t) + Y_n^s(t-1) + Y_n^s(t+1)}{4},$$

   or by the average of the non missing values if one is also missing. Obviously, if all the measurements are missing, there is no completion. This step is repeated until 80% of the data is completed.

The three errors described in step 1 are well known by road traffic managers. For example, the constant speed measurements are due to stations that have not been re-initialized after a measure and that automatically repeat this same measure over several consecutive periods.

After performing this algorithm, the number of days used for the study is reduced since we only use the curves Y_n^s without missing data. We can notice that some counting stations have no complete days before the protocol. In many counting stations, missing data are not MAR (i.e.

Missing At Random data), and too numerous, preventing the use of the EM algorithm, used in Section 3, to complete the data.

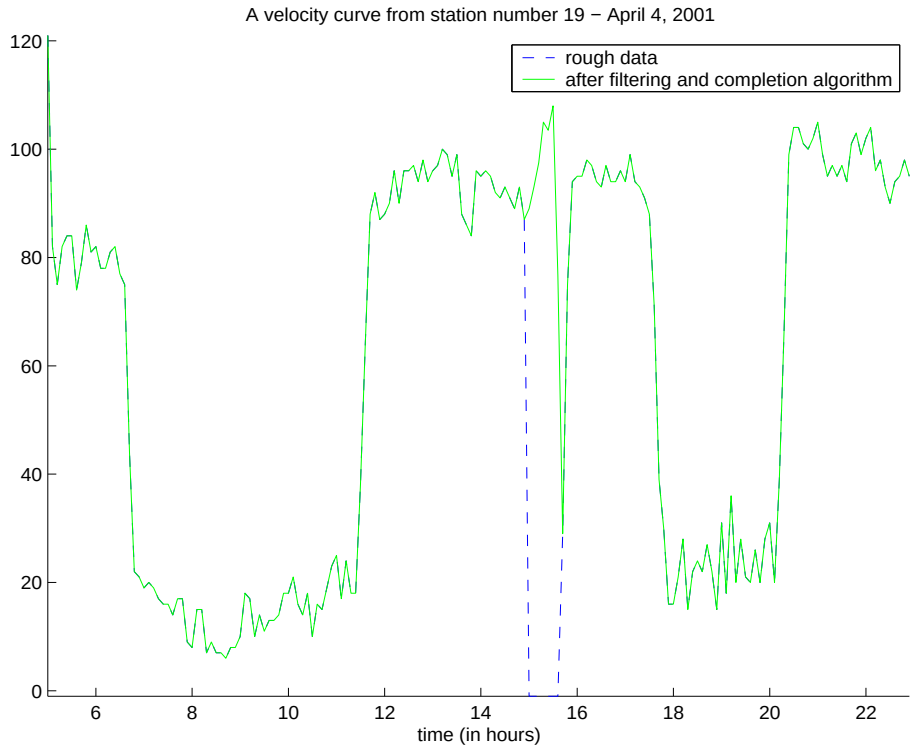

Figure 1: Plot of a speed curve from the counting station 19, before the protocol (dotted line) and after the protocol (solid line). In this example, all missing values (noted -1) have been completed.

2.3 Forecasting method

For a new day J_{n_0}, at time $H \in \left[\frac{t_0+49}{10}, \frac{t_0+50}{10}\right[$, corresponding to period t_0, we observe the speed measurements $Y^s_{n_0}(t)$, $\forall t < t_0$, $\forall s = 1, \ldots, S$. Hence we want to forecast $Y^s_{n_0}(t)$, $\forall t \geq t_0$, $\forall s = 1, \ldots, S$, in order to forecast travel time on a given itinerary. Indeed, once obtained all the values $Y^s_{n_0}(t)$, for all $t \geq t_0$ and for all counting stations C_s, $s = 1, \ldots, S$, we can compute the travel time from one point to another at time $H + h$, for any $h \geq 0$.

We assume that for each counting station C_s, there is a number m_s of representative behaviours of road trafficking, noted $f_1, \ldots, f_{m_s}$. Hence, the forecasting method can be divided into two main parts:

1. Estimate the standard profiles, $f_1, \ldots, f_{m_s}$, for each counting station C_s, $s = 1, \ldots, S$,
2. Matching the incoming observations to these archetypes and hence estimating speed at all counting stations to forecast travel time.

For sake of simplicity, since the study is carried out for each counting station, we will drop the s index and so we will note m the number of behaviours and $f_1, \ldots, f_m$, the associated standard

profiles.

To estimate the standard profiles $f_1, \ldots, f_m$, we use two different methods: a mixture model in Section 3, and an empirical classification method in Section 4. Also, in order to make any comparisons between travel time forecasts using these two methods, we get a *test sample* of $N_T = 19$ days. This test sample will be used in Section 5 to forecast simulated travel times.

3. MIXTURE MODEL

3.1 Model description

Consider C_s a counting station. For this chosen station, each day J_n, $n = 1, \ldots, N$, we observe the vehicle speed at discrete times $t = 1, \ldots, T$, with $T = 180$. Set, for each $n = 1, \ldots, N$, $y_n = {}^t(y_n(1), \ldots, y_n(T)) \in \mathbb{R}^T$ the vector of daily observed speeds, and $Y_n = {}^t(Y_n(1), \ldots, Y_n(T))$ the corresponding random vector. We assume, as quoted in Section 2, that there are m different archetypes, $f_1, \ldots, f_m$, where for all $j = 1, \ldots, m$, $f_j = {}^t(f_j(1), \ldots, f_j(T)) \in \mathbb{R}^T$. The assumption behind this modelling is that highway traffic phenomena do not depend on the traffic of the previous days, and that there are exogenous variables determining to which pattern the observed data belong to. So, the measure of the vehicle velocity at one point can be written as follows

$$Y_n = \sum_{j=1}^{m} \mathbf{1}_j(X_n) f_j + \epsilon_n, \; n = 1, \ldots, N, \text{ where} \tag{1}$$

- X_n, $n = 1, \ldots, N$, are i.i.d. non observable variables, taking values in the discrete set $\{1, \ldots, m\}$,
- $\epsilon_n \in \mathbb{R}^T$, $n = 1, \ldots, N$, is a Gaussian vector, independent from the observations, with variance $\sigma^2 I_T$, with I_T the $T \times T$ identity matrix. The observations come from counting stations which are all the same, satisfying to the quality controls. Hence the variance is taken constant equal to σ^2.

The unknown parameters are: the number of components m; the archetypes f_j, $j = 1, \ldots, m$; the noise variance σ^2; as well as the parameters of the law of X_n. The discrete law of X_n is entirely characterized by the probabilities $\pi_j = \mathbf{P}(X_n = j)$, $j = 1, \ldots, m$.

In a first approach, we assume that m is known. Selecting the right number of models is the topic of Section 3.2. Hence, the parameters to be estimated are:

$$\Psi = {}^t(\pi_1, \ldots, \pi_m, f_1, \ldots, f_m, \sigma).$$

Set $y = (y_1, \ldots, y_N)$ the observed values of the random sample $Y = (Y_1, \ldots, Y_N)$. Set also $x = (x_1, \ldots, x_N)$ the non observed values of the random sample $X = (X_1, \ldots, X_N)$.

To estimate Ψ, consider the maximum likelihood estimator. The log-likelihood of the model can be written in the following form

$$L(y; \Psi) = \sum_{n=1}^{N} \log \left(\sum_{j=1}^{m} \pi_j \phi(y_n; f_j, \sigma) \right),$$

where $\phi(\cdot; \mu, \sigma)$ is the density of a Gaussian vector with mean $\mu \in \mathbb{R}^T$ and variance $\sigma^2 I_T$. The log-likelihood estimator of Ψ is a root of the equation

$$\nabla_\Psi L(y; \Psi) = 0,$$

where $\nabla_\Psi L(y;\Psi)$ is the gradient of L with respect to the unknown parameters of Ψ.

In a mixture model, analogous to the one studied by McLeish & Small (1986), the solution of the previous equation can be computed efficiently with an EM algorithm, as it is stated in the work of Basford & McLachlan (1985) or Lachlan (1982). The EM algorithm was created by Dempster, Laird & Rubin (1977) to maximize the log-likelihood with missing data. It enables, with a recursive method, to change the problem of maximizing the log-likelihood into the problem of maximizing the completed log-likelihood of the model:

$$L_C(y,x;\Psi) = \sum_{n=1}^{N}\sum_{j=1}^{m} \mathbf{1}_j(x_n)\log\left(\pi_j\phi(y_n;f_j,\sigma)\right).$$

Set $Z_n = (Z_{nj})_{j=1,\dots,m} = (\mathbf{1}_j(X_n))_{j=1,\dots,m}$. This variable completes the model since it points out which class the random vector Y_n belongs to. This variable follows a multinomial distribution with unknown parameter $\pi = {}^t(\pi_1,\dots,\pi_m)$.

Let describe the $p+1$ step of the EM algorithm. Set

$$Q\left(\Psi,\Psi^{(p)}\right) = \mathbf{E}\left[L_C(Y,X;\Psi)|Y=y;\Psi^{(p)}\right],$$

the expectancy of the log-likelihood of the complete data, conditionally to the observed data, and with respect to the value of the parameter computed at step p, written $\Psi^{(p)}$. Then we obtain

$$Q\left(\Psi,\Psi^{(p)}\right) = \sum_{n=1}^{N}\sum_{j=1}^{m}\mathbf{E}\left[Z_{nj}|Y_n=y_n;\Psi^{(p)}\right]\log\left(\pi_j\phi\left(y_n;f_j,\sigma\right)\right).$$

Hence, the step $p+1$ of the EM algorithm is divided into two stages: the expectation stage (E) and the maximization stage (M):

(E) In this stage, the random variable Z_{nj} is replaced by its expectancy, conditionally to the observed data, and with respect to the current value of the parameter:

$$\tau_k^{(p)}(y_n) = \mathbf{E}\left[Z_{nk}|Y_n=y_n;\Psi^{(p)}\right] = \mathbf{P}\left(Z_{nk}=1|Y_n=y_n;\Psi^{(p)}\right) = \frac{\pi_k^{(p)}\phi\left(y_n;f_k^{(p)},\sigma^{(p)}\right)}{\sum_{j=1}^{m}\pi_j^{(p)}\phi\left(y_n;f_j^{(p)},\sigma^{(p)}\right)}.$$

(M) In this stage, the maximization is conducted by choosing the value of the parameter Ψ that maximizes $Q(\Psi,\Psi^{(p)})$. It will be written $\Psi^{(p+1)}$. The estimators are the followings:

$$\begin{aligned}
\pi_j^{(p+1)} &= \frac{1}{N}\sum_{n=1}^{N}\tau_j^{(p)}(y_n),\\
f_j^{(p+1)} &= \frac{\sum_{n=1}^{N}\tau_j^{(p)}(y_n)y_n}{\sum_{n=1}^{N}\tau_j^{(p)}(y_n)},\\
\sigma^{(p+1)} &= \left(\frac{1}{NT}\sum_{n=1}^{N}\sum_{j=1}^{m}\sum_{t=1}^{T}\tau_j^{(p)}(y_n)\left(y_n(t)-f_j^{(p)}(t)\right)^2\right)^{1/2}.
\end{aligned}$$

The model we use, undergoes the assumptions over the EM algorithm, which ensures its convergence.

In order to avoid local minima, we have used a Stochastic Approximation of the EM algorithm, the SAEM algorithm. Such algorithm has been developed, and its convergence has been proved, by Delyon, Lavielle & Moulines (1999). The main advantage for using SAEM algorithm rather than EM algorithm is that the former is less sensitive to the choice of the starting point in the algorithm. For a good choice of the initialization parameters, the outcome of the two algorithms are quite the same, while, for a bad choice, the estimates given by successive applications of EM algorithm can be far the one from the others. On the contrary, SAEM provides the same results. For more about the comparison between stochastic versions of the EM algorithm, we refer to Broniatowsky, Celeux & Diebolt (1983); Celeux & Diebolt (1992) or Celeux, Chauveau and Diebol (1995).

The step $p+1$ of the SAEM algorithm comes from the step $p+1$ of the EM algorithm in the following way:

- The E stage is replaced by a simulation stage. In this stage, we draw $K(p+1)$ realizations of the multinomial variable Z_{nj}, written z_{nj}^k, $k = 1, \ldots, K(p+1)$, according to the distribution given by the values of the parameters at step p, $\Psi^{(p)}$. The log-likelihood is then modified in the following way:

$$\widehat{Q}_p(\Psi) = \widehat{Q}_{p-1}(\Psi) + \gamma_{p+1}\left(\frac{1}{K(p+1)}\sum_{k=1}^{K(p+1)}\sum_{n=1}^{N}\sum_{j=1}^{m} z_{nj}^k \log\left(\pi_j^{(p)}\phi\left(y_n; f_j^{(p)}, \sigma^{(p)}\right)\right) - \widehat{Q}_{p-1}(\Psi)\right),$$

  where $(\gamma_p)_{p\geq 1}$ is a sequence of positive reals.

- The M stage of the algorithm takes place as previously.

We have used this modified algorithm in our work, with a numerical good choice of the sequences $(\gamma_p)_{p\geq 1}$ and $(K(p))_{p\geq 1}$, which leads to the results presented in Section 5.

3.2 Estimation of the number of components of the mixture

The aim of this study is to find the optimal number m of components of the mixture (1). For this, we use a methodology close to model selection approach. For a theoretical approach of these technic, we refer for instance to the work of Baraud (2000); Birgé & Massart (1998) or Barron, Birgé & Massart (1999).

For each value of $m \geq 1$, we consider the set $\mathcal{F}_m = \{g_1, \ldots, g_m,\ g_i \in \mathbb{R}^T,\ \pi_1, \ldots, \pi_m, \sigma\}$, and we write $\mathcal{F} = \cup_{m\geq 1}\mathcal{F}_m$ the collection of all the different models. For a fixed m, we have seen in Section 3.1 that it was possible to estimate the unknown parameters of the model, $\hat{\Psi}^{(m)}$. Hence, it is now possible to compute the estimated log-likelihood of the chosen model $L\left(\hat{\Psi}^{(m)}; y, m\right)$. The idea is given in the following remark.

REMARK 1. *The best choice for m, m^*, is the one, such that the function $m \mapsto L\left(\hat{\Psi}^{(m)}; y, m\right)$ does not increase in a significant way for values greater than m^*.*

So, set $J(\Psi, y) = -L(\Psi; y)$. We use the following notations:

$$\hat{\Psi}^{(m)} = \arg\min_{\psi\in\mathcal{F}_m} J(\Psi, y), \tag{2}$$

$$J_m = J\left(\hat{\Psi}^{(m)}, y\right).$$

For all $\beta > 0$ and for all $1 \leq m \leq M$, where M is an upper bound for the maximum number of components, set

$$\hat{m}(\beta) = \arg \min_{1 \leq m \leq M} (J_m + \beta m).$$

The following proposition is due to Lavielle (2002).

PROPOSITION 1. *There is a sequence* $m_1 = 1 < m_2 < \ldots$, *and a sequence* $\beta_0 = +\infty > \beta_1 > \ldots$, *with*

$$\forall i \geq 1,\ \beta_i = \frac{J_{m_i} - J_{m_{i+1}}}{m_{i+1} - m_i},$$

such that

$$\forall \beta \in]\beta_i, \beta_{i-1}[,\ \hat{m}(\beta) = m_i.$$

As a consequence, the estimation procedure of the optimal number of components of the mixture is given by:

- For $m = 1, \ldots, M$, compute $\hat{\Psi}^{(m)}$ and J_m,
- Then compute the sequence $(\beta_i)_{i=1,\ldots,M}$, as well as l_i the length of the intervals $]\beta_i, \beta_{i-1}[$, for all $i = 1, \ldots, M$,
- Keep the largest value of the m_i such that $l_i >> l_j$, for all $j > i$.

Actually, the previous procedure is a model selection approach with a stability criterion that replaces the trade off between bias and variance, as it is quoted by Birgé & Massart (2001). This proposition provides an automatic criterium that mimics the main idea developped in Remark 1. Figure 2 present the result of this estimation procedure on the counting station 19.

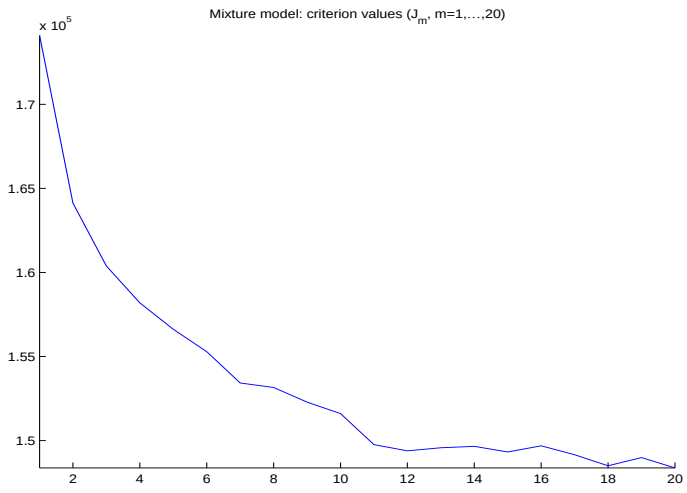

Figure 2: Estimation procedure of the optimal number of components of the mixture model for counting station 19.

This method is based on mixture model (1). An alternative approach is given by the method of automatic classification, in Section 4, which makes it possible to be closer to the data observed.

4. CLASSIFICATION METHOD

4.1 Hierarchical classification

The outcome of a hierarchical classification strongly depends on the choice of between-individuals and between-clusters distance. Classical distance, see for example, Gordon (1999) or Dazy & Le Barzic (1996), were not appropriate for this kind of temporal data. Indeed, the study of the road traffic implies taking into account the temporal aspect of our speed curves. For example, consider three simplified speed curves, X, Y and Z, obtained one from another by a translation. These three curves are characterized by a constant speed, 90 km/h, from 5 AM to 11 PM, except over a 2 hours period during which, the speed is reduced to 30 km/h, respectively at 8 AM, 11 AM and 2 PM. For the Euclidean distance d we get $d(X,Y) = d(Y,Z) = d(X,Z) = 389$. But, a suitable classification distance must make the difference between a deceleration which occurs at 8 AM, at 11 AM or at 2 PM. Thus, we build a distance, denoted Δ, taking into account this shift effect.

DEFINITION 1. *Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$. Set $\Delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ as*

$$\Delta(x,y) = \sqrt{{}^t(x-y)W(x-y)},$$

with W a $n \times n$ matrix defined by $W_{ij} = \frac{n-|i-j|}{n}$, $\forall i = 1,\ldots,n$, $\forall j = 1,\ldots,n$.

We point out that Δ is a distance on $\mathbb{R}^n$. For the preceding example, Δ gives the following results: $\Delta(X,Y) = \Delta(Y,Z) = 637$ and $\Delta(X,Z) = 967$. Thus, Δ enables to differentiate the translation speed curves. This property, on such curves, can be proved by straightforward calculations. So, take Δ as the between-individuals distance, and define the between-clusters distance index as the distance index of the maximum variation, noted D. Hence, for two clusters A and B, we get

$$D(A,B) = \max_{x \in A,\, y \in B} \Delta(x,y).$$

Choosing the criterion of the maximum variation enables us to obtain homogeneous classes, loosing between classes heterogeneity.

The hierarchical classification is carried out with the Johnson agglomerative algorithm, which gathers, at each step, the closest clusters.

4.2 Choice of the optimal number of clusters

Once the hierarchical classification is carried out, we aim at keeping only a small number m^* of significant classes, for each counting station C_s. This implies cutting the classification tree at a given height, which depends on the accuracy of the description of the data we want to keep. Here, this level will be data driven and chosen in order to minimize the forecasting error over an observation sample. Hence, our classification method can be viewed as a learning theory methodology. Each station database is divided into two samples:

- A *model sample*, used to estimate standard profiles, with N_M complete days (80% of the data),
- A *learning sample*, used to estimate the optimal number of standard profiles, with N_L complete days (20% of the data).

Hence, we forecast travel time on the learning sample with $m = 1, \ldots, M$, with M a fixed big enough integer, and then, choose the number of clusters minimizing the forecasting error criterion. The optimal number of clusters, m^*, of the counting station number s, C_s, minimizes the absolute forecasting error over two hours. Hence, we write

$$m^* = \arg \min_{m=1,\ldots,M} \sum_{n=1}^{N_L} \sum_{t=11}^{161} \sum_{p=t}^{p+19} \left| Y_n(p) - f_{m,j(n,t)}(p) \right|,$$

with

$$f_{m,j(n,t)} = \arg \min_{f_{m,\tilde{m}},\, \tilde{m}=1,\ldots,m} \Delta' \left((Y_n)_1^{t-1}, (f_{m,\tilde{m}})_1^{t-1} \right),$$

where $f_{m,1}, \ldots, f_{m,m}$, are the standard profiles obtained for m clusters. Hence, $f_{m,j(n,t)}$ is the closest profile to Y_n in the sense defined by the distance Δ' defined in Section 5, at period t, when we choose m standard profiles.

Figure 3 shows the absolute errors calculated station 19, for $m = 1, \ldots, 20$. The forecasting error first decreases while m increases, while there is an over fitting phenomenon when the number of profiles increases after a certain value. So it is possible to estimate an optimal number of classes. We also point out that most of the counting stations have the same behaviour.

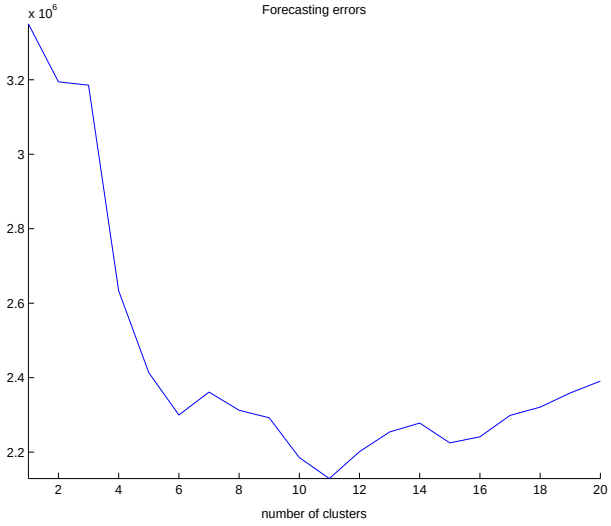

Figure 3: Absolute forecasting error for the station 19, with m standard profiles, $m = 1, \ldots, 20$. The optimal value is reached for $m^* = 11$.

5. ARCHETYPES FOR ROADTRAFFICKING BEHAVIOUR

First of all, we point out that for both models, The number of chosen archetypes or clusters depends on the counting station we consider. In this study, the optimal numbers of representatives lies between 5 and 15 for 80% of the counting stations.
For the mixture model ,using the criterion presented in Section 3.2, for counting station 19, the values of $J\left(\hat{\Psi}^{(m)}, y\right)$ do not decrease in a significant way for values greater than $m^* = 11$. The behaviour of the loglikelihood is presented in Figure 2. Figure 4, on the bottom, represents the 11 archetypes of the station 19. Then, for all C_s, $s = 1, \ldots, S$, we select the optimal number of representatives using this stopping criterion.

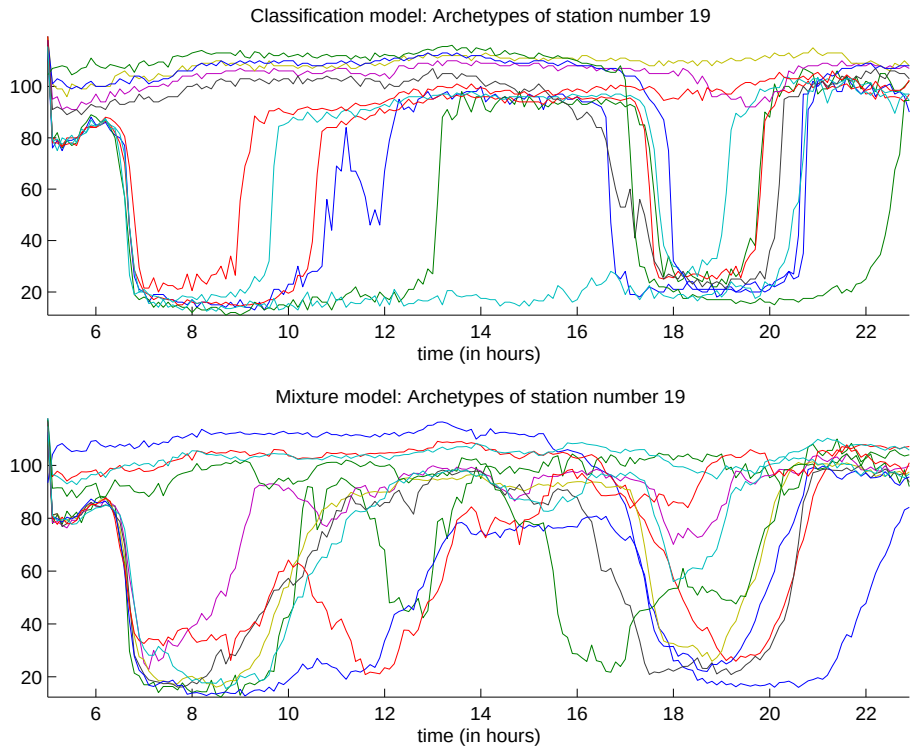

Figure 4: Standard profiles of the counting station 19. On the top, with classification model. On the bottom, with mixture model.

For the classification model, the learning process explained in Section 4 provides representative clusters. But, after splitting the data into m^* representative classes, for each station C_s, we now have to extract the standard profiles, $\hat{f}_1, \ldots, \hat{f}_{m^*}$, for each cluster, hence obtaining a representative curve of the speed behaviour in each class. For this, we use a robust estimator: the median of the speed curves of each class. Figure 4, on the top, presents the $m^* = 11$ standard profiles obtained for the station 19.

The two methods for this particular station give the same number of archetypes. In general, the optimal number of representative functions selected by the mixture model is slightly smaller than the number given by the classification method. Thus, some known behaviours appear for both models, like traffic jams at peak hours and traffic keeping moving otherwise. Nevertheless, we can see an important difference between the different models. The hierarchical classification let appear some curves that seem to be outliers or rare events. Hence, for station 19, in Figure 4, there are curves with larger and deeper traffic jams on the top figure than on the bottom one. Indeed the EM-type algorithm over-smooths the curves and do not take into account rare events, which play an important role in roadtrafficking description.

6. TRAVEL TIME FORECASTING

In the two previous parts, we have constructed, for each observation station C_s, $s = 1, \ldots, S$, and using two different methods, the standard profiles $f_j^{(i)} \in \mathcal{F}^i$, $i \in \{1, 2\}$, $j = 1, \ldots, m_i$. These two sets, $\mathcal{F}^1$ and $\mathcal{F}^2$, represent the archetypes of the daily vehicle speed resulting, respectively, from

the mixture model ($i = 1$) and from the empirical classification ($i = 2$). Our aim is now to use these profiles to forecast, for a given itinerary, a customer travel time, at $H+h$, with h (in minutes) in the set $\{18, 30, 48, 60, 78, 90, 108\}$.

Let J_{n_0} be the observation day, and t_0 such that $H \in \left[\frac{t_0+49}{10}, \frac{t_0+50}{10}\right[$. In order to forecast, we estimate, for all the stations of the itinerary, the speeds $f^s(t)$, $\forall s \in \mathcal{S}$, $\forall t \geq t_0$, where $\mathcal{S}$ is the set of all stations of the chosen itinerary. Once speed evolutions are known, the estimation of the travel time is easy. As a consequence, the main issue is, for each station, the estimation of the traffic velocity. For this, we compare the incoming data of the day J_{n_0} before time H, i.e. $Y^s_{n_0}(t)$, $\forall s \in \mathcal{S}$, $\forall t < t_0$, with all the curves of $\mathcal{F}^1$ or $\mathcal{F}^2$, by choosing the nearest curve. For this, define for all $i \in \{1, 2\}$, for all $g \in \mathcal{F}^i$, $g_1^{t_0-1} = {}^t(g(1), \ldots, g(t_0-1))$ and $Y_1^{t_0-1} = {}^t(Y^s_{n_0}(1), \ldots, Y^s_{n_0}(t_0-1))$, and Δ', a modified restriction of Δ to the subset $\mathbb{R}^{t_0-1} \times \mathbb{R}^{t_0-1}$, defined as follows.
Let $Y_1^{t_0-1}$ be the observed data at J_{n_0} the observation day and on the station C_s, before time H, i.e. for $t < t_0$ (for sake of simplicity we have omitted station and day indexes). Define the distance between $Y_1^{t_0-1}$ and the different archetypes of the station C_s, restricted to the values of $t < t_0$, and written $(f_j)_1^{t_0-1}$, with $f_j \in \mathcal{F}$, $\forall j = 1, \ldots, m$, by

$$\Delta'\left(Y_1^{t_0-1}, (f_j)_1^{t_0-1}\right) = \frac{\Delta\left(PY_1^{t_0-1}, P\,(f_j)_1^{t_0-1}\right)}{\sqrt{\pi_j}},$$

where π_j, $j = 1, \ldots, m$, is the size of the cluster number j, and P is a $t_0 - 1 \times t_0 - 1$ matrix, defined by $P_{ij} = \begin{cases} \frac{1}{t_0-i} & \text{if } i = j \text{ and } i \geq t_0 - 10, \\ 0 & \text{in any other case.} \end{cases}$

Hence, after having chosen one of the two models, $\mathcal{F} = \mathcal{F}^1$ or $\mathcal{F}^2$, the estimator will be for all $t \geq t_0$, $\hat{f}(t)$, with

$$\hat{f} = \arg\min_{g \in \mathcal{F}} \Delta'\left(g_1^{t_0-1}, Y_1^{t_0-1}\right).$$

Then, we have used the test sample to forecast travel time on the A4W road section, using the standard profiles obtained with the two methodologies described in Section 3 and Section 4.

We compare the results with the estimations given by the stationary model, defined as the simplest model, which estimates the speed by the last observed speed, i.e. $Y^s_{n_0}(t_0-1)$. This model plays a key role since, on the one hand it is the only reference we have, and on the other hand, the forecasting results with such a model gives us an indicator of the traffic behaviour on the considered road section. Indeed, good forecasts point out that the traffic is moving freely. On the contrary, bad forecasts with the stationary model show that the itinerary is often congested, leading to numerous changes in the velocity that prevent the use of a stationnary model.

Table 1 and Table 2 present some characteristics (minima and maxima) of travel time errors obtained with 2000 simulated itineraries on the test sample, for the three models: the stationary model, the classification model and the mixture model. The error (in minutes) is defined by

$$\text{error} = \frac{\text{real travel time} - \text{estimated travel time}}{\text{real travel time}}.$$

Figure 5 shows the evolution of travel time forecasting errors (mean and standard deviation) from a 0 to 2 hours forecasting horizon: $h \in \{18, 30, 48, 60, 78, 90, 108\}$, h in minutes.

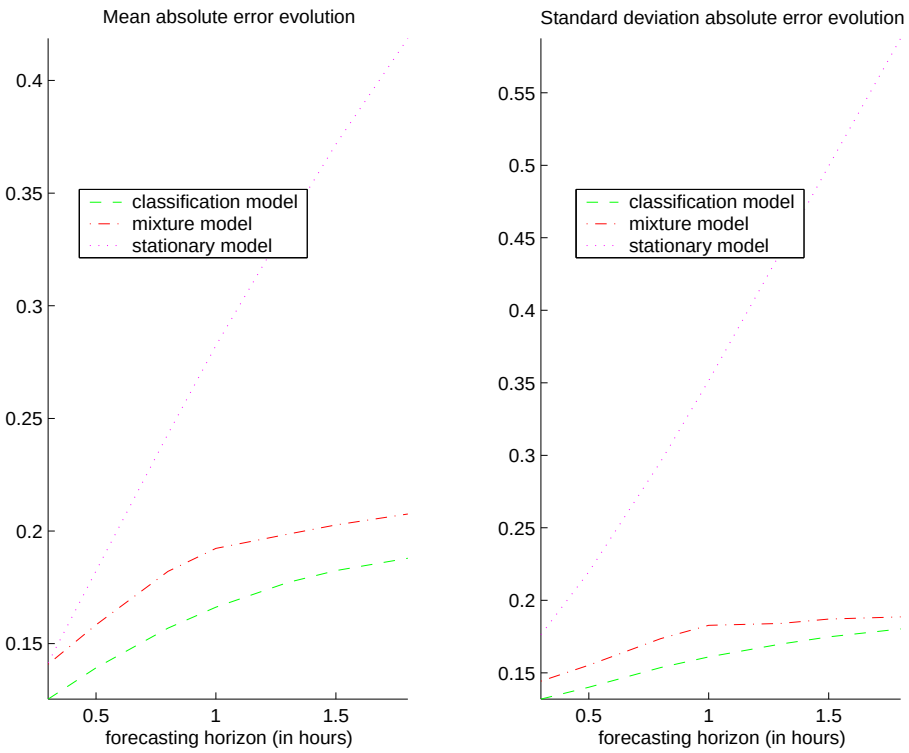

Figure 5: Evolution of travel time forecasting errors (mean and standard deviation) from a 0 to 2 hours forecasting horizon.

Figure 6 is an example of the evolution of the forecast errors throughout a test sample day (April 4, 2001) for $h = 60$ (1 hour horizon). More precisely, for each of these days, in our fixed itinerary, we consider a traveller with that itinerary beginning at each of the periods, and then we calculate actual and model predicted travel times.

These results enable us to compare the three procedures: the rough stationary model which is compared to the predictors model: the classification model with an optimal number of clusters chosen with a learning sample and the model of the loglikelihood estimator. Hence, we can draw the following conclusions.

First of all, both predictors improve the estimate provided by the stationary model, with smaller prediction variances (Figure 5) and small error ranges (Table 1 and Table 2). This inprovement depends on the road which is studied. The more different behaviours of road trafficking can be found on that road, the better is the gain. This is easy to explain since the stationary model provides a mean pattern which is far from the real feature when there are many. Yet, this number of representative patterns is a measure of the complexity of the road, standing for its variability, with respect to changes day after day.

We also point out that both models underestimate the real travel time (Figure 5), so for practical applications, this bias can be taken into account.

When comparing the performance of our estimators, we can see that the loglikelihood estimator

is slightly outperformed by the classification type estimator. Indeed, the mean of the errors with the classification model is closer to 0, and the variance is smaller (Figure 5). Moreover, error range is smaller for the classification model (Table 1 and Table 2). The reasons for this difference are the following:

- First, the distance chosen to evaluate the performance of the estimator, is the same that is used to classify the data in the methodology described in Section 4, since this distance best matches the prediction goals. But the optimal choice of models is achieved, via a learning process, by minimizing the prediction distance over a learning sample. Hence, this choice induces a bias in favor of the classification method.

- Outliers and rare events play also an important role in this study. On the one hand, the mixture model is very sensitive to outliers. Indeed, the loglikelihood estimator uses all the data with the same weights to build an average representative, while a classification method tends to isolate such outliers in special classes. Hence, the blurring effect of outliers is stronger for the mixture model since they add a deviation term to the estimates. On the other hand, rare events are more easily catched by the classification method. Indeed, we have point out in Section 5, that standards profiles given by the classification method contain more rare event profiles. Hence, like we can see, for example, in Figure 6, which is one of the worst case, the congested phenomena are slightly better detected: the travel times with traffic jams of the mornings are better estimated. Unlike theoretical study in model selection, we found the optimal number of models by minimizing the prediction error but without a penalty term. As a matter of fact, we did not want to discard the rare events that can be alone in a cluster, but represent a real behaviour in road trafficking. Hence, it enables the predictor based on the classification model to keep in mind some unlikely events, and to give an adequate response when the observations do not follow a typical pattern.

- Nevertheless, there are two advantages for using the loglikelihood model. First, it is a very efficient method on a computational point of view, which is much faster than the computations necessary to perform the learning process of the model selection method. Moreover, for small values of the number of models, loglikelihood estimators provide a better description of the data. But, increasing the number of models for the loglikelihood does not improve the estimation error. Indeed the additional selected patterns are redundant, since, as we have already said, the loglikelihood estimator does not put the stress on rare events, while the classification type predictor isolates such features in single classes.

Table 1: Minimum error evolution for different values of forecasting horizon.

For. Hor.	Classification	Mixture	Stationnary
18 min.	-0.62	-1.13	-2.21
30 min.	-0.67	-0.79	-2.53
48 min.	-0.79	-1.02	-3.37
60 min.	-0.88	-1.10	-3.82
78 min.	-0.96	-1.13	-5.26
90 min.	-0.95	-1.14	-6.17
108 min.	-1.18	-1.13	-7.30

Table 2: Maximum error evolution for different values of forecasting horizon.

For. Hor.	Classification	Mixture	Stationnary
18 min.	0.67	0.71	0.67
30 min.	0.67	0.72	0.68
48 min.	0.66	0.71	0.73
60 min.	0.66	0.72	0.77
78 min.	0.66	0.71	0.83
90 min.	0.67	0.71	0.90
108 min.	0.68	0.71	0.95

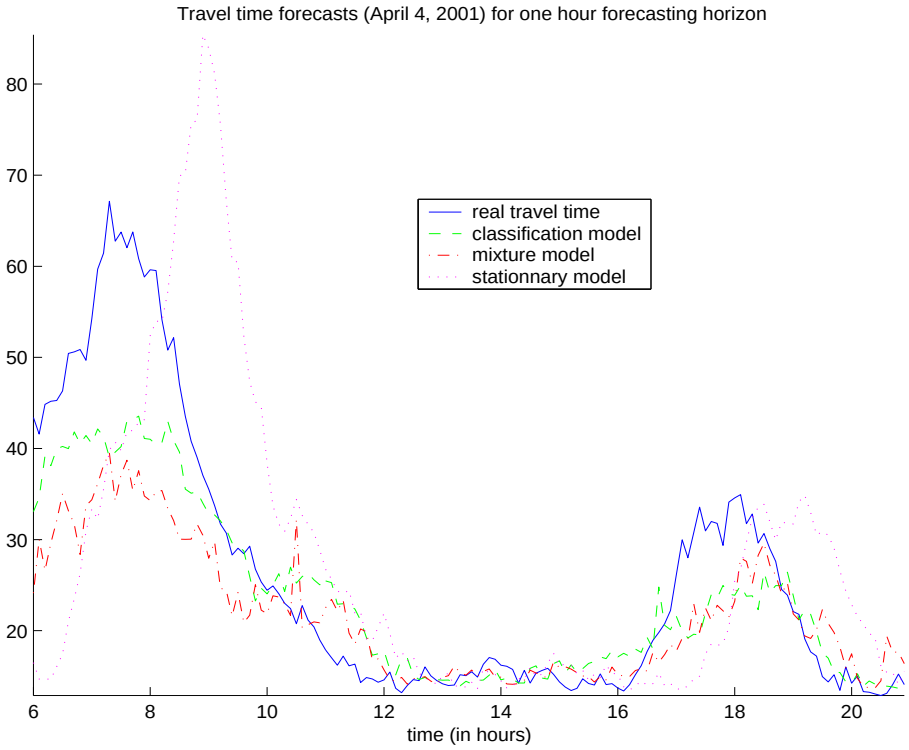

Figure 6: Forecasting travel times on a test sample day (April 4, 2001).

7. CONCLUSION

Our results are encouraging and are far better than the results given by the usual global forecasting methods (*Sytadin*, http://www.sytadin.tm.fr, or *Bison Futé* for instance), which only rely on the rough model. Both models are interesting: the mixture model by its simplicity and the good performance, and the classification model which is the more accurate but, at the same time, the more complicated on a computational point of view.

However, it is possible to improve the performances of the method, for example in the choice of the archetypes that stand for a whole cluster. Indeed, in a class, the functions are similar but they

may be translation shift between them. As a consequence, taking the median of all the functions for the representant of the class often leads to an over smoothing effect. Methods able of keeping the structure of the functions group, as it is done by Kneip & Gasser (1992); Kneip (1994) or Ramsay & Dazell (1991), are developed by Gamboa, Loubes & Maza (2003) in the setting of high dimensional data.

Moreover, other modelling attempts can be conducted. It seems rather natural to take into account the dependency of all the stations which are considered in this work as independent. A method using the spatial links between the observation cells is taken into account by the authors in a forthcoming work.

Finally aggregating the estimators should also improve the performance of the procedure. Indeed, in this work, we have considered separately the prediction given by each methodology. An alternative should be to use a linear combination of such predictors to improve our results. Such a work is still in progress.

ACKNOWLEDGEMENTS

We are grateful to J-M. Azais and F. Gamboa for their advice.

REFERENCES

Y. Baraud (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4), 467–493.

A. Barron, L. Birgé & P. Massart (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3), 301–413.

K. E. Basford & G. J. McLachlan (1985). Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association*, 80(390), 286–293.

D. Belomestny, V. Jentsch & M. Schreckenberg (2003). Completion and continuation of nonlinear traffic time series: a probabilistic approach. *J.Phys.A:Math. Gen.*, 36, 11369–11383.

P. C. Besse & H. Cardot (1996). Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *The Canadian Journal of Statistics*, 14(4), 467–487.

L. Birgé & P. Massart (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3), 329–375.

L. Birgé & P. Massart (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3), 203–268.

D. Bosq (2003). Processus linéaires vectoriels et prédiction. *Comptes Rendus Mathématique, Académie des Sciences, Paris*, 337(2), 115–118.

L. Breiman, J. Friedman, R. Olshen & Charles J. Stone (1984). *Classification and regression trees.* Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software. Belmont, CA.

M. Broniatowski, G. Celeux & J. Diebolt (1983). Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data Analysis and Informatics*, 359–374. North Holland.

G. Celeux (1988). Classification et modèles. *Revue de Statistique Appliquée*, 36(4), 43–57.

G. Celeux, D. Chauveau & G. Diebol (1995). On stochastic versions of the EM algorithm. *Rapport de recherche INRIA*, 2514.

G. Celeux & J. Diebolt (1992). A stochastic approximation type EM algorithm for the mixture problems. *Stochastics Reports*, 41, 119–134.

J. Chen (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1), 221–233.

R. C. H. Cheng & W. B. Liu (2001). The consistency of estimators in finite mixture models. *Scandinavian Journal of Statistics*, 28(4), 603–616.

S. Cohen (1990). *Ingénierie du trafic routier.* Presses de l'Ecole Nationale des Ponts et Chaussées. INRETS, France.

F. Couton, M. Danech-Pajouh & M. Broniatowski (1996). Application des mélanges de lois de probabilité à la reconnaissance de régimes de trafic routier. *Recherche Transports Sécurité*, 53, 49–58.

M. Danech-Pajouh & M. Aron (1994). *ATHENA: Prévision à court terme du trafic sur une section de route.* INRETS, France.

F. Dazy & J.-F. Le Barzic (1996). *L'analyse de données évolutives.* Editions Technip.

B. Delyon, M. Lavielle & E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1), 94–128.

A. P. Dempster, N. M. Laird & D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1), 1–38.

T. Dochy (1995). *Arbres de régression et Réseaux de neurones appliqués à la prévision de trafic routier.* Thèse de l'université Paris Dauphine.

F. Ferraty & P. Vieu (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2), 161–173.

F. Ferraty & P. Vieu (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *Journal of Nonparametric Statistics*, 16(1-2), 111–125.

F. Gamboa, J.-M. Loubes & E. Maza (2005). Structural estimation for high dimensional data. *forthcoming paper.*

A. D. Gordon (1999). *Classification – 2nd Edition.* CHAPMAN & HALL/CRC. University of St. Andrews, UK.

M. Jambu (1978). *Classification automatique pour l'analyse des données. I.* Dunod, Paris.

A. Kneip (1994). Nonparametric estimation of common regressors for similar curve data. *The Annals of Statistics*, 22(3), 1386–1427.

A. Kneip & T. Gasser (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3), 1266–1305.

M. Lavielle (2002). On the use of penalized contrasts for solving inverse problems. *Preprint.*

B. Lindsay & M. Lesperance (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47(1-2), 29–39.

G. J. McLachlan (1982). On the bias and variance of some proportion estimators. *Communications in Statistics (B)*, 11(6), 715–726.

D. L. McLeish & C. G. Small (1986). Likelihood methods for the discrimination problem. *Biometrika*, 73(2), 397–403.

V. Núñez-Antón, J. M. Rodríguez-Póo & P. Vieu (1999). Longitudinal data with nonstationary errors: a nonparametric three-stage approach. *Test*, 8(1), 201–231.

C. Preda & G. Saporta (2004). PLS approach for clusterwise linear regression on functional data. *Classification, clustering, and data mining applications*, 167–176. Springer, Berlin.

J. O. Ramsay & C. J. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society (B)*, 53(3), 539–572.

H. J. M. Van Grol, M. Danech-Pajouh, S. Manfredi & J. Whittaker (1998). DACCORD: on-line travel time prediction. In 8^{th} WCTR 1998, 2.

Received ???
Accepted ???

Jean-Michel LOUBES: `Jean-Michel.Loubes@math.univ-montp2.fr`
Laboratoire de probabilités et statistique de l'Université Montpellier 2
Montpellier, 34000
France

Elie MAZA: `Elie.Maza@math.ups-tlse.fr`
Laboratoire de Statistique et Probabilités de l'université Toulouse 3
Toulouse, 31000
France

Marc LAVIELLE: `Marc.Lavielle@math.u-psud.fr`
Laboratoire de Mathématiques de l'université Paris Sud
Orsay, 94125
France