ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

*Appl. Statist.* (2020)
**69**, *Part* 3, *pp.* 623–640

# A novel regularized approach for functional data clustering: an application to milking kinetics in dairy goats

C. Denis,

*AgroParisTech, Institut National de la Recherche Agronomique, Paris, Université Paris-Saclay, Paris, and Université Paris-Est, Champs-sur-Marne, France*

and E. Lebarbier, C. Lévy-Leduc, O. Martin and L. Sansonnet

*AgroParisTech, Institut National de la Recherche Agronomique, Paris, and Université Paris-Saclay, Paris, France*

[Received July 2019. Revised January 2020]

**Summary.** Motivated by an application to the clustering of milking kinetics of dairy goats, we propose a novel approach for functional data clustering. This issue is of growing interest in precision livestock farming, which is largely based on the development of data acquisition automation and on the development of interpretative tools to capitalize on high throughput raw data and to generate benchmarks for phenotypic traits. The method that we propose in the paper falls in this context. Our methodology relies on a piecewise linear estimation of curves based on a novel regularized change-point-estimation method and on the $k$-means algorithm applied to a vector of coefficients summarizing the curves. The statistical performance of our method is assessed through numerical experiments and is thoroughly compared with existing experiments. Our technique is finally applied to milk emission kinetics data with the aim of a better characterization of interanimal variability and towards a better understanding of the lactation process.

*Keywords*: Change point; Functional data clustering; Regularized methods

## 1. Introduction

Precision livestock farming is a blooming field grounded in the development of sensors providing high throughput data and thus potentially increasing access to valuable information on biological processes. Therefore, developing methods for data analysis and interpretation has become a challenging issue in animal science. Economic performance of dairy goat farming systems is primarily based on milk production and a large amount of farmers' working time is spent milking animals; see Marnet *et al.* (2005). Moreover, with the increasing size of goat herds and the rapid growth of the dairy goat industry, more detailed information on individual milking performance is necessary. In this context, a better understanding of the variability in milk flow kinetics could for instance help in refining selection criteria for breeding programmes, simplifying milking workload or controlling udder health. Milk emission kinetics recorded during milking of dairy goats are classically described and classified through synthetic parameters such as milking time, maximum and average milk flow rates, and the time to reach 500 g $\text{min}^{-1}$ milk flow; see Romero *et al.* (2017). In this paper, we explore the possibility of considering milk

*Address for correspondence*: C. Lévy-Leduc, AgroParisTech, 16 Rue Claude Bernard, Paris 75231, France.
E-mail: celine.levy-leduc@agroparistech.fr



0035–9254/20/69623

emission kinetics as a whole function, opening new perspectives to study interanimal variability.

From a statistical point of view, this issue belongs to the general field of functional data analysis; see Ramsay and Silverman (2005) for a survey on this subject. In the specific functional data clustering framework, several approaches have been proposed by Abraham *et al.* (2003), Jacques and Preda (2013) and Bouveyron *et al.* (2015) among others. For a review on this subject, we refer the reader to Jacques and Preda (2014a) and the references therein. The main idea underlying these approaches consists in applying unsupervised clustering methods to the vector of the coefficients corresponding to the projection of the curves onto an appropriate basis, the *B*-splines basis being one of the most popular. This kind of approach was extended to deal with multivariate functional data by Jacques and Preda (2014b), who proposed the first model-based clustering algorithm in the multivariate context, and more recently by Schmutz *et al.* (2018).

To deal with the functional clustering of the milking kinetics of goats, which correspond to the cumulative amount of milk (in millilitres) retrieved from the goat udder during machine milking at the experimental station, some specific features must be taken into account; see Fig. 1 for some examples of such kinetics. We can see from Fig. 1 that these curves are non-decreasing and can be split into two parts, namely an increasing linear part and an almost constant part. Inspired by Abraham *et al.* (2003), we propose in this paper a dimension reduction approach based on a continuous piecewise linear function fit to each curve which boils down to a change-point-detection issue which will be crucial in our method.

The problem of detecting change points in the mean of a signal is largely addressed in the

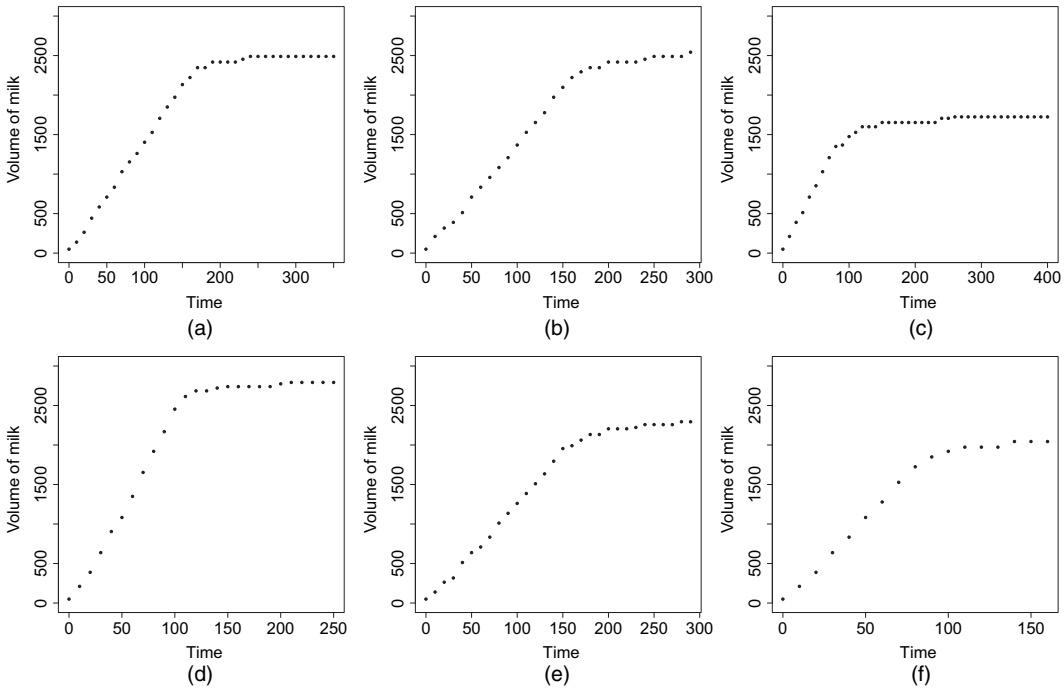


**Fig. 1.** Some examples of milking kinetics of goats: (a) goat 250053006212047; (b) goat 250053006212054; (c) goat 250053006212064; (d) goat 250053006212075; (e) goat 250053006212084; (f) goat 250053006213039

literature. In particular, it is now well known that in (penalized) maximum likelihood frameworks the dynamic programming (DP) algorithm (Bellman, 1961; Auger and Lawrence, 1989) and its recent pruned versions (Killick *et al.*, 2012; Rigaill, 2015; Maidstone *et al.*, 2016) are the only algorithms that retrieve the exact solution very quickly. However, DP can be used only if the contrast to be optimized is additive with respect to the segments; see for example Bai and Perron (2003), Picard *et al.* (2005) and Lavielle (2005). When detecting changes in the slope with a continuity condition, the segments will unavoidably be linked and therefore the additivity condition is not satisfied. This partly explained why this change-point-detection problem has not been thoroughly investigated in the literature compared with the simpler detection in the mean problem. Recently, Fearnhead *et al.* (2019) proposed to extend the pruned exact linear time algorithm (Killick *et al.*, 2012) to this problem. Their idea is to include the penalty in the DP algorithm with a pruning strategy. The penalty that they proposed is proportional to the number of change points up to a penalty constant. However, this penalty constant needs to be chosen in advance, which is not easy in practical situations.

In this paper, we first propose a novel method to estimate the change points in the slope combining the trend filtering that was proposed by Tibshirani (2014) with a (penalized) maximum likelihood approach which is useful for removing the spurious change points that may be proposed by trend filtering. These change point estimators are then used for devising a new dimension reduction approach: each curve is summarized by a vector containing the coefficients of its projection onto an order 2 $B$-spline basis having for knots the change points obtained and also the change point locations. Including the change points both in the features characterizing the curves and in the $B$-spline knots is the main novelty compared with classical approaches reviewed in Jacques and Preda (2014a).

The paper is organized as follows. The methodology that we propose is decribed in Section 2. The performance of our approach is investigated in Section 3 through numerical experiments. Finally, in Section 4, we apply our method to the data that motivated this study.

## 2. Methodology

In this section, we describe our novel functional data clustering approach which consists of two steps which can be summarized as follows:

*step 1*, piecewise linear estimation of the curves by using a novel change-point-estimation method based on the trend filtering approach and $B$-splines;
*step 2*, applying the $k$-means algorithm to a vector of coefficients summarizing the curves that are obtained in the first step.

These two steps are further described in what follows.

### *2.1. First step: piecewise linear estimation of the curves based on a change-point-estimation method*

In what follows, we assume that the observations of a given curve $\mathbf{Y} = (Y_1, \ldots, Y_n)$ correspond to a noisy function evaluated at the input points $\mathbf{x} = (x_1, \ldots, x_n)$. In this step, we aim to estimate each curve by a piecewise linear function by using a two-stage approach described below.

#### *2.1.1. First stage: trend filtering for change point estimation*

We use the trend filtering approach that was proposed by Tibshirani (2014) which consists in fitting to the observations $\mathbf{Y}$ the vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)$ by using a regularized method. More precisely, we use

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n}(\|\mathbf{Y}-\boldsymbol{\beta}\|_2^2 + \lambda\|D^{(2)}\boldsymbol{\beta}\|_1),$$

where $\|y\|_2^2 = \Sigma_{i=1}^n y_i^2$, $\|y\|_1 = \Sigma_{i=1}^n |y_i|$, for $y = (y_1, \ldots, y_n)$, $\lambda$ is a positive constant which must be tuned and $D^{(2)}$ is the discrete difference operator of order 2 defined by

$$D^{(2)} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \end{pmatrix}.$$

The estimated piecewise linear function is obtained with $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ where $\hat{\lambda}$ must be properly tuned. Usually, this parameter is chosen by using resampling approaches such as cross-validation or stability selection; see Meinshausen and Bühlmann (2010). From $\hat{\boldsymbol{\beta}}(\hat{\lambda})$, we define a set of potential change point indices as the co-ordinates where the vector $D^{(2)}\hat{\boldsymbol{\beta}}(\hat{\lambda})$ is not equal to 0. However, in change-point-estimation frameworks, the performance of such methods may be altered since some change points may be omitted by subsampling. Moreover, it is well known that such regularization approaches lead to oversegmentation phenomena. Usually, in these cases, a DP algorithm is then used on the set of potential change points to remove the irrelevant change points; see for instance Harchaoui and Lévy-Leduc (2008, 2010).

We make a proposal following this strategy: to avoid the use of a resampling method, we choose a sufficiently small $\lambda$ to obtain a sufficiently large a set of potential change points. More precisely, we set a maximal number of change points denoted by $K_{\max}$ and choose $\lambda$ such that, among the $\lambda$s leading to $K_{\max}$ change points, $\hat{\lambda}$ is the value minimizing $\|\mathbf{Y} - \hat{\boldsymbol{\beta}}(\lambda)\|_2^2$.

Let $(\hat{n}_1, \ldots, \hat{n}_{K_{\max}})$ be the resulting change point indices and the associated change point positions $(\hat{t}_1, \ldots, \hat{t}_{K_{\max}}) = (x_{\hat{n}_1}, \ldots, x_{\hat{n}_{K_{\max}}})$. For each $K$ in $\{1, \ldots, K_{\max}\}$, we use the DP algorithm to retrieve the $K$ most relevant change point indices among $\hat{n}_1, \ldots, \hat{n}_{K_{\max}}$. DP is thus applied to the restricted set $Y_{\hat{n}_1}, \ldots, Y_{\hat{n}_{K_{\max}}}$. Note that a slight modification of the algorithm is considered to make the piecewise linear fit to data continuous. The optimal number of change points $\hat{K}$ is then chosen by using the criterion that was proposed by Lavielle (2005).

### *2.1.2. Second stage: projection onto the B-spline basis having as knots the change points obtained*

Each curve will then be summarized by a few coefficients corresponding to the coefficients of its projection onto the $B$-spline basis $(B_{i,2})_{1\leqslant i\leqslant \hat{K}+2}$ defined as follows; see Hastie *et al.* (2009), page 206, for a review on the subject. Let $\hat{t}_0 = x_1$ and $\hat{t}_{\hat{K}+1} = x_n$. Also we define the augmented knot sequence $\tau$ such that

$$\begin{aligned} \tau_1 = \tau_2 = \hat{t}_0 = x_1, \\ \tau_{j+2} = \hat{t}_j, \qquad j = 1, \ldots, \hat{K}, \\ \tau_{\hat{K}+3} = \tau_{\hat{K}+4} = \hat{t}_{\hat{K}+1} = x_n, \end{aligned}$$

namely

$$(\tau_1, \ldots, \tau_{\hat{K}+4}) = (x_1, x_1, \hat{t}_1, \ldots, \hat{t}_{\hat{K}}, x_n, x_n).$$

The $i$th $B$-spline function $B_{i,2}$ having $\tau$ for knot sequence satisfies

$$B_{i,2}(u) = \frac{u-\tau_i}{\tau_{i+1}-\tau_i} B_{i,1}(u) + \frac{\tau_{i+2}-u}{\tau_{i+2}-\tau_{i+1}} B_{i+1,1}(u),$$

with $i \in \{1, \ldots, \hat{K}+2\}$ and the convention $0/0 = 0$, where

$$B_{i,1}(u) = \begin{cases} 1, & \text{if } \tau_i \leqslant u < \tau_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, each curve is estimated by $\hat{f}$ defined by

$$\hat{f}(u) = \sum_{i=1}^{\hat{K}+2} \hat{\theta}_i B_{i,2}(u), \tag{1}$$

where the $\hat{\theta}_i$s are obtained by using a least square criterion. Hence, the coefficients summarizing each curve are

$$(\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{K}+2}, \hat{t}_1, \ldots, \hat{t}_{\hat{K}}). \tag{2}$$

Note that the coefficients $\hat{\theta}_k$ can be interpreted as the values of the estimated piecewise linear curves at the change points.

### *2.2. Second step: clustering using the k-means algorithm*

We propose to cluster the curves (milking kinetics) by using the summarized coefficients obtained in expression (2). However, the number of change points $\hat{K}$ may change from one curve to another, leading to different lengths of the summarized coefficient vectors. Thus, we consider summarized coefficients of length $2\hat{K}_M + 2$ where $\hat{K}_M$ corresponds to the largest value of $\hat{K}$. For curves having a number of change points that is smaller than $\hat{K}_M$, we replace the missing $\hat{t}_k$ and the missing $\hat{\theta}_k$ by 0 that are added after the $\hat{t}_k$s and after the $\hat{\theta}_k$s respectively. Other values could have been considered such as $x_n$ for the missing $\hat{t}_k$s and $Y_n$ for the missing $\hat{\theta}_k$s. We decided to choose 0 for the two following reasons. Firstly, the choice of $x_n$ was not adapted for the application that we had in view since it was not relevant to cluster the curves by using their length. Secondly, the choices $x_n$ and $Y_n$ have been tested through some simulations (adapted from model 1 of the simulation study) and we could observe that they altered the performance of the clustering compared with our strategy. The results are not displayed here but they are available on request.

To ensure that all the coefficients belong to the same range, the coefficients are centred and normalized among the different individuals to guarantee that the empirical mean is equal to 0 and the empirical variance is equal to 1. Then, we use the $k$-means algorithm of Hartigan and Wong (1979). The number $k$ of clusters is chosen by using the strategy that was proposed by Charrad *et al.* (2014) which consists in using the majority rule, i.e. taking for $k$ the value that is chosen by the largest number of criteria among the four following indices: the Krzanowski and Lai (1988) index, Hartigan index, SDindex and Ptbiserial index. Further details on these indices can be found in Charrad *et al.* (2014).

## 3. Numerical experiments

In this section, we investigate the statistical performance of our procedure. The simulation scheme that we used for this investigation is described in Section 3.1. We also propose in Section 3.2 to benchmark our procedure against existing approaches and to assess our change-point-estimation approach in Section 3.3.

### *3.1. Simulation scheme*

Based on the data coming from our motivating application, we consider two different models for generating the data that we shall refer to as model 1 and model 2 in what follows. For

each model, the complete observed data are $(\mathbf{Y}, Z)$, where $\mathbf{Y}$ is in $\mathbb{R}^n$ and corresponds to the observations of an underlying function, which we shall specify later at the input points $\mathbf{x} = (x_i)_{1\leqslant i\leqslant n} = (10(i-1))_{1\leqslant i\leqslant n}$ with $n = 51$. $Z$ denotes the label of $\mathbf{Y}$ which takes its value in $\mathcal{Z} = \{1, 2, 3, 4\}$. Moreover, for each $z \in \mathcal{Z}$, the associated cluster $\mathcal{C}_z$ is characterized by a number of change points $K_z$, a vector of change points $\mathbf{t}^z$ and a vector of parameters $\boldsymbol{\theta}^z \in \mathbb{R}^{K_z+1}$. Hence, each model is defined by a set of parameters $\{K_z, \mathbf{t}^z, \boldsymbol{\theta}^z : z \in \mathcal{Z}\}$. The values of the parameters that are associated with each model are reported in Tables 1 and 2. For each model, the clusters are distinguishable by both the change points and the parameters.

For each model, the vector $(\mathbf{Y}, Z)$ is simulated according to the following procedure.

(a) The label $Z$ is drawn from a uniform distribution on $\mathcal{Z}$.
(b) We generate $\tilde{\mathbf{t}}^Z = \mathbf{t}^z + \mathcal{U}$, such that $\mathcal{U} = (U, \ldots, U)$, where $U$ is a uniformly distributed random variable on $\{-30, -20, 10, 0, 10, 20, 30\}$.
(c) We generate $\tilde{\boldsymbol{\theta}}^Z = \boldsymbol{\theta}^z + \mathcal{V}$, such that $\mathcal{V} = (V, \ldots, V)$, where $V$ is a uniformly distributed random variable on $[-200, 200]$.
(d) Then, we consider the sequences $(\tilde{t}_0^Z, \ldots, \tilde{t}_{K_Z+1}^Z) = (0, \tilde{\mathbf{t}}^Z, 500)$, $(\tilde{\theta}_0^Z, \ldots, \tilde{\theta}_{K_Z+1}^Z) = (0, \tilde{\boldsymbol{\theta}}^Z)$ and define for $x \in [\tilde{t}_j^Z, \tilde{t}_{j+1}^Z]$, and $j \in \{0, \ldots, K_Z\}$,

$$f_{\tilde{\mathbf{t}}^Z, \tilde{\boldsymbol{\theta}}^Z}(x) = (\tilde{\theta}_{j+1}^Z - \tilde{\theta}_j^Z)\frac{x - \tilde{t}_j^Z}{\tilde{t}_{j+1}^Z - \tilde{t}_j^Z} + \tilde{\theta}_j^Z. \tag{3}$$

(e) Finally, we define $\mathbf{Y}$ such that, for $i \in \{1, \ldots, n\}$,

$$Y_i = f_{\tilde{\mathbf{t}}^Z, \tilde{\boldsymbol{\theta}}^Z}(x_i) + \varepsilon_i, \tag{4}$$

where the $\varepsilon_i$s are independent and identically distributed $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma \in \{1, 5\}$.

Note that the function $f$ that is defined in equation (3) can be seen as another way of writing equation (1).

**Table 1.** Set of parameters for model 1

| $z$ | $K_z$ | $\mathbf{t}^z$ | $\boldsymbol{\theta}^z$ |
|---|---|---|---|
| 1 | 2 | (150, 250) | (1600, 1900, 2000) |
| 2 | 2 | (150, 300) | (1400, 1800, 2200) |
| 3 | 4 | (100, 200, 300, 400) | (300, 1500, 1700, 2000, 2200) |
| 4 | 3 | (50, 150, 300) | (200, 1300, 1800, 2100) |

**Table 2.** Set of parameters for model 2

| $z$ | $K_z$ | $\mathbf{t}^z$ | $\boldsymbol{\theta}^z$ |
|---|---|---|---|
| 1 | 2 | (150, 250) | (1600, 1900, 2000) |
| 2 | 2 | (150, 300) | (1400, 1800, 2200) |
| 3 | 4 | (100, 200, 300, 400) | (300, 1500, 1700, 2000, 2200) |
| 4 | 3 | (150, 250, 300) | (200, 700, 1000, 1600) |

Fig. 2 displays some observations that were generated by using the above simulation scheme for each model and for each $\sigma$. We can see from Fig. 2 that the clustering problem that is associated with model 1 seems to be the most difficult. In model 1, the clusters are indeed completely mixed whereas in model 2 cluster $\mathcal{C}_4$ is well separated from the others. Observe also that the data that are generated have the same behaviour as the data coming from our motivating application: they are non-decreasing and piecewise linear with a small additive noise; see Fig. 1.

### *3.2. Statistical performance*

Following the simulation scheme that was described in Section 3.1, the performance of our procedure is assessed for each model and each $\sigma$ and is compared with two different clustering

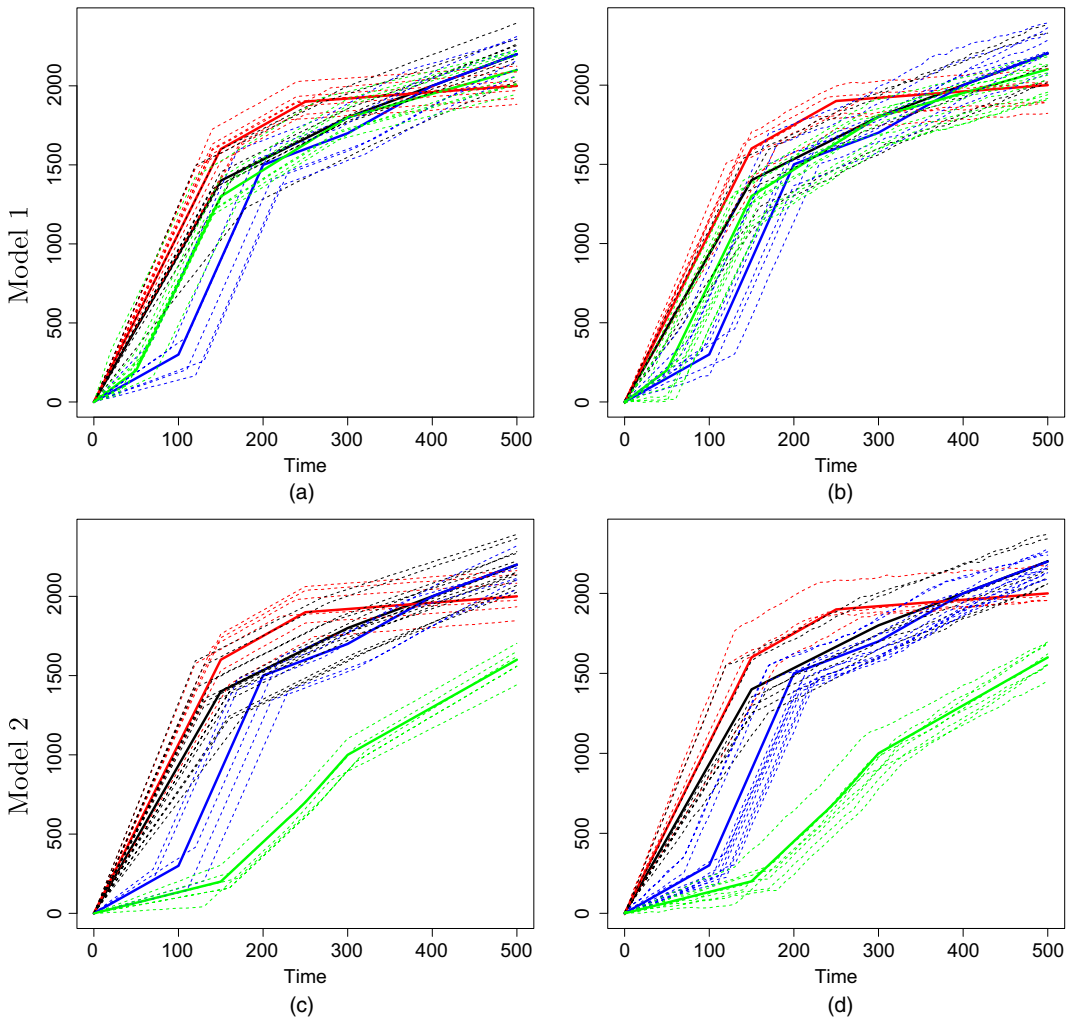


**Fig. 2.** Examples of observations generated from (a), (b) model 1 and (c), (d) model 2 for (a), (c) $\sigma = 1$ and (b), (d) $\sigma = 5$: ——, ——, ——, ——, representative curves of each cluster $f_{\mathbf{t}}^{z}, \theta^{z}$ respectively cluster 1, cluster 2, cluster 3 and cluster 4; - - - -, - - - -, - - - -, - - - -, some examples of the corresponding **Y**

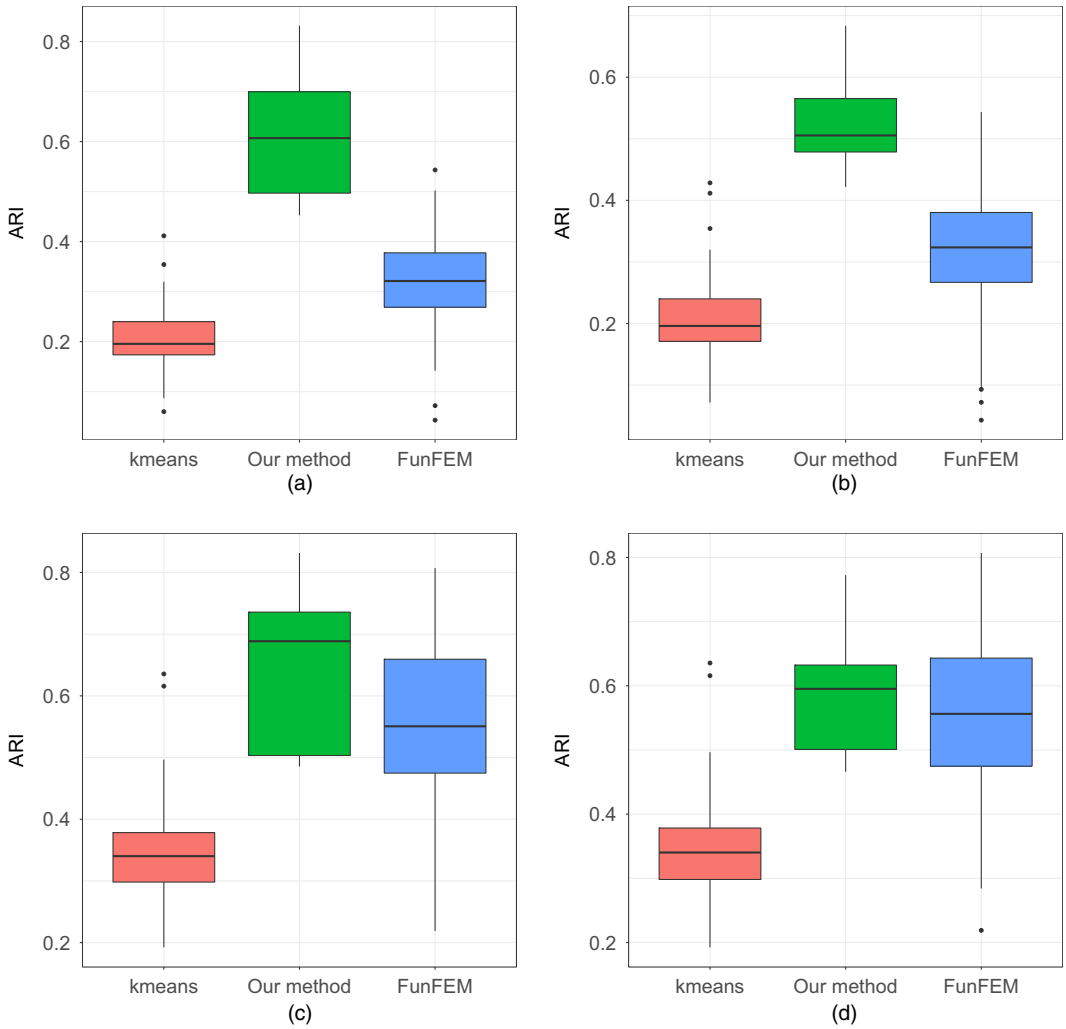


**Fig. 3.** Boxplots of ARI for (a), (b) model 1 and (c), (d) model 2 for (a), (c) $\sigma = 1$ and (b), (d) $\sigma = 5$

methods: The $k$-means algorithm applied to the raw data $\mathbf{Y}$ and the 'FunFEM' procedure (clustering in the discriminative functional subspace) described in Bouveyron *et al.* (2015) and available in the R package `FunFEM`. The FunFEM method is dedicated to the clustering of functional data and is based on a functional mixture model. All the methods are compared thanks to the adjusted Rand index ARI which was defined in Hubert and Arabie (1985) and is often used for clustering validation. It is indeed a measure of agreement between two partitions. Note that the number of clusters $k$ in the $k$-means algorithm is chosen by using the same strategy as we considered in our approach. As far as `FunFEM` is concerned, we used the default parameters.

For each model and for each $\sigma$ in $\{1, 5\}$, we repeat independently 100 times the following steps.

(a) We simulate a sample $\mathcal{D}_N = \{(\mathbf{Y}^1, Z^1) \dots (\mathbf{Y}^N, Z^N)\}$ of size $N = 100$ according to the scheme that was described in Section 3.1.

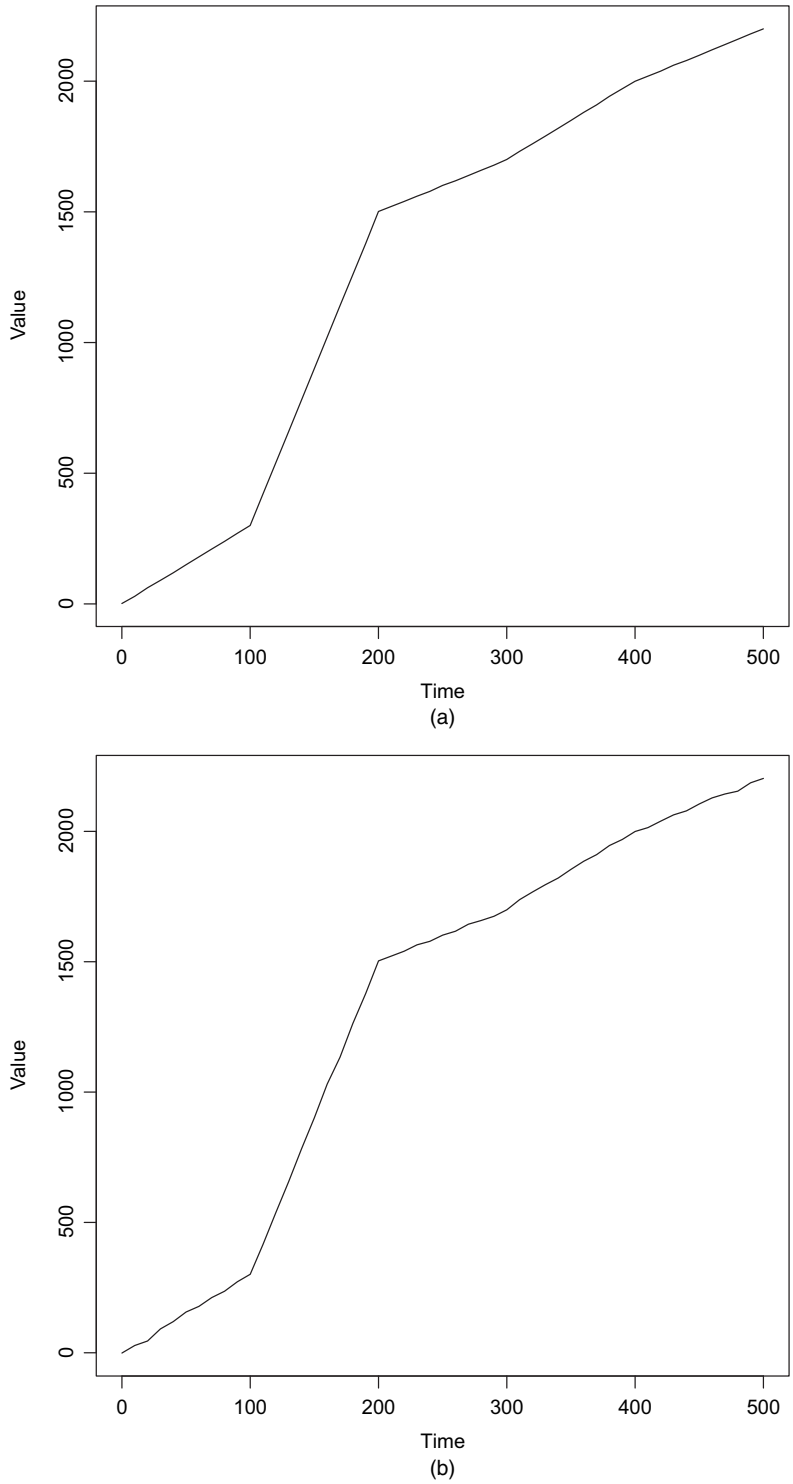


**Fig. 4.** Examples of **Y** belonging to cluster 3 of model 1 for (a) $\sigma = 1$ and (b) $\sigma = 5$

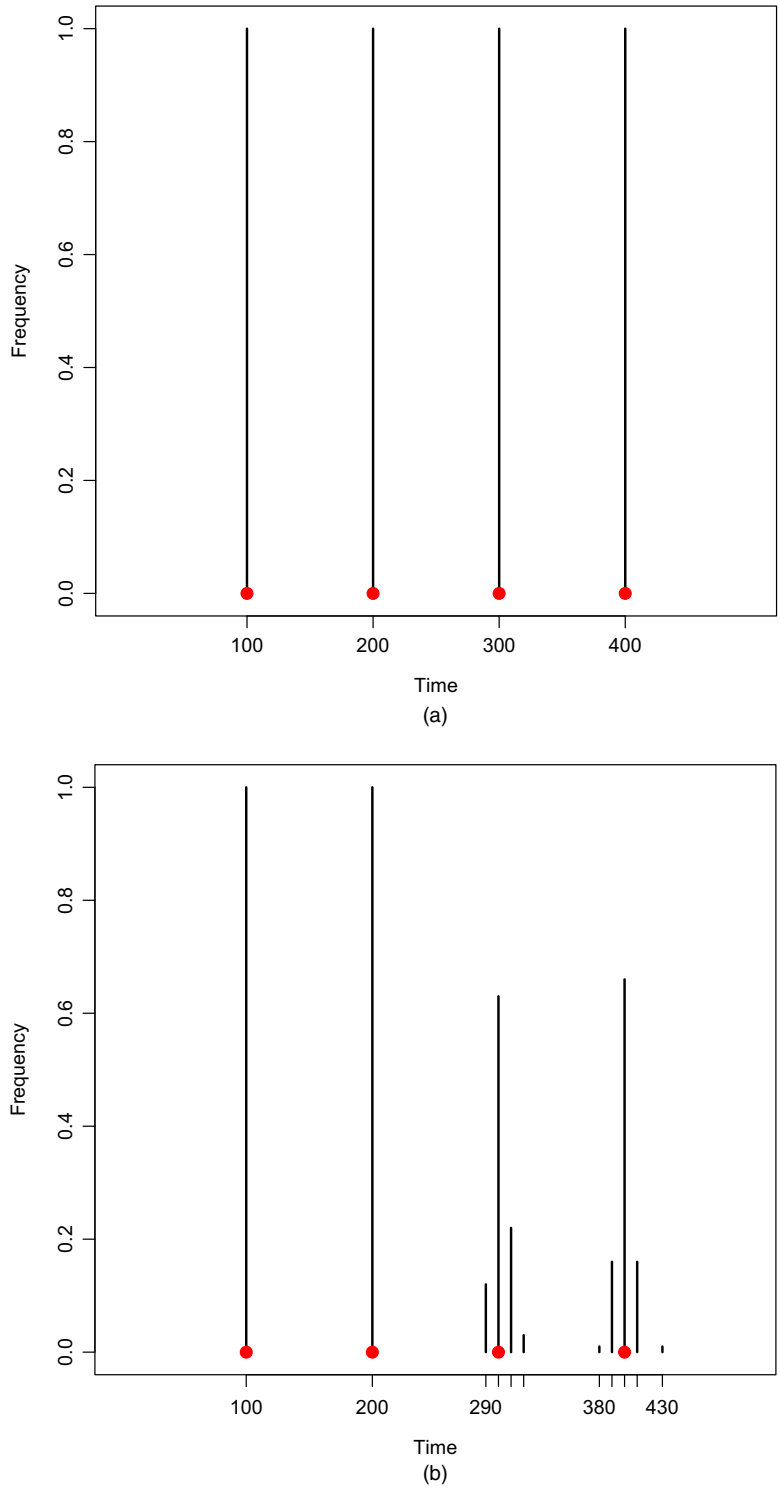


**Fig. 5.** Change-point-estimation frequencies for (a) $\sigma = 1$ and (b) $\sigma = 5$: ●, true change point positions

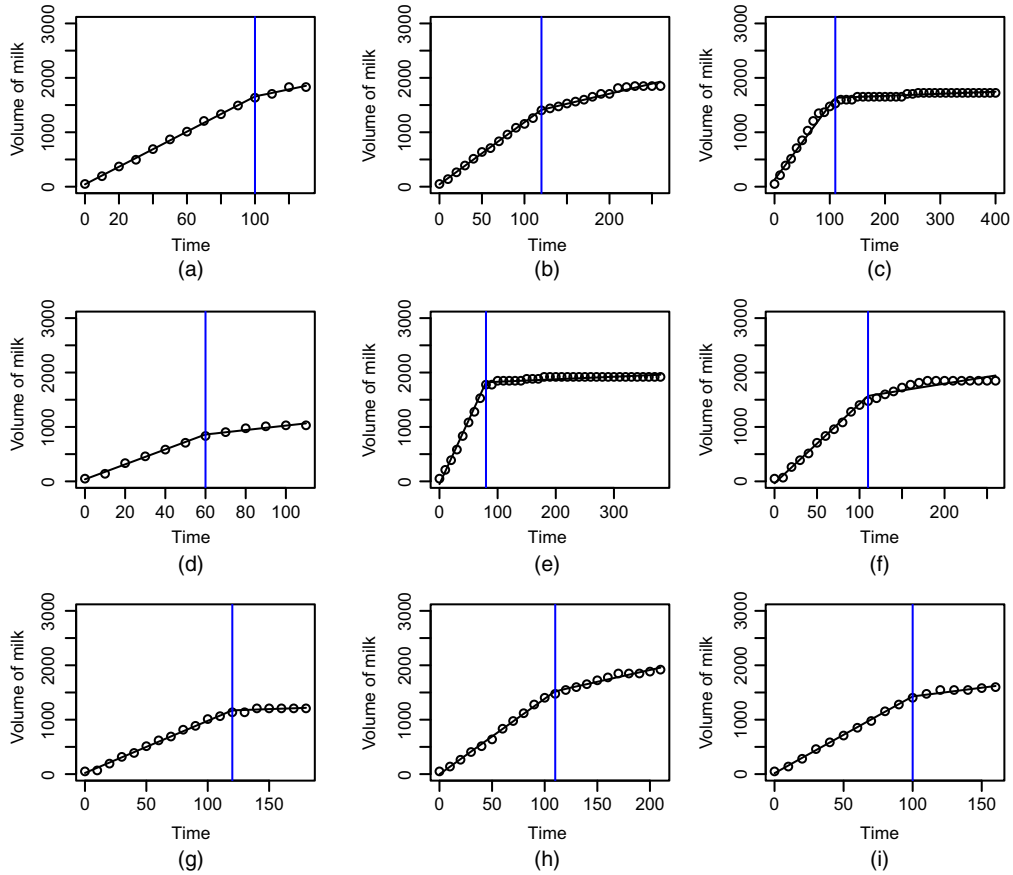


**Fig. 6.** Some examples of milking kinetics belonging to cluster 1 (○, the data; ———, piecewise linear fit obtained thanks to our method; |, position of the change point): (a) goat 250053006212058; (b) goat 250053006212048; (c) goat 250053006212064; (d) goat 250053006212083; (e) goat 250053006213036; (f) goat 250053006213031; (g) goat 250053006213005; (h) goat 250053006213008; (i) goat 250053006213030

(b) We apply each method to $\mathcal{D}_N$.
(c) On the basis of the clustering obtained, we compute ARI.

The results are displayed in Fig. 3 with $K_{\max} = 10$. We can see from Fig. 3 that our method outperforms the others in all cases except for model 2 with $\sigma = 5$ where the performance of our method is on a par with that of FunFEM. Note that applying the $k$-means to a relevant summary measure of $\mathbf{Y}$ significantly improves the clustering performance. Moreover, we observe that, when $\sigma$ increases, the performance of our approach is slightly altered since the change points are more difficult to locate accurately; see Section 3.3.

### *3.3. Assessment of our change-point-estimation procedure*

We provide the following numerical experiments for assessing the change-point-estimation stage of our method. We used the parameters that are associated with cluster 3 of model 1; see Table 1. We repeat 100 times the following steps.

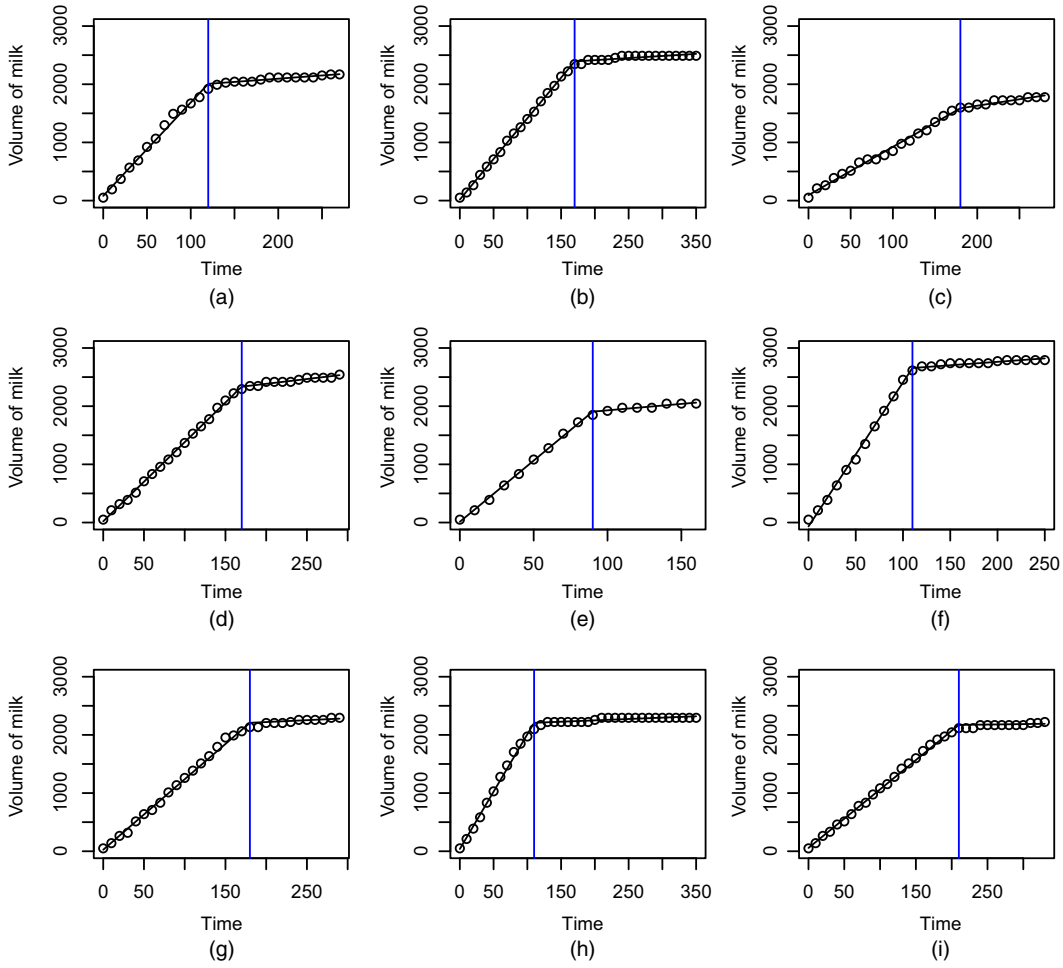


**Fig. 7.** Some examples of milking kinetics belonging to cluster 2 (○, the data; ———, piecewise linear fit obtained thanks to our method; |, position of the change point): (a) goat 250053006213082; (b) goat 250053006212047; (c) goat 250053006213040; (d) goat 250053006212054; (e) goat 250053006213039; (f) goat 250053006212075; (g) goat 250053006212084; (h) goat 250053006212080; (i) goat 250053006212091

(a) We simulate $\mathbf{Y}$ according to equation (4) with $\sigma \in \{1, 5\}$.
(b) We estimate the change points according to the procedure that was described in the first stage of the first step in Section 2.

Some examples of $\mathbf{Y}$ for the two values of $\sigma$ are displayed in Fig. 4. We can see from Fig. 4 that the change points at 300 and 400 are more difficult to detect than the others. It is all the more true when $\sigma = 5$.

Fig. 5 displays the frequency of the number of times where each position has been estimated as a change point. We can see that the change points have all been retrieved and that no spurious change points are provided when $\sigma = 1$. In the case where $\sigma = 5$, although the positions of the true change points are retrieved most of the time, some additional spurious change points are also selected with a very low frequency.

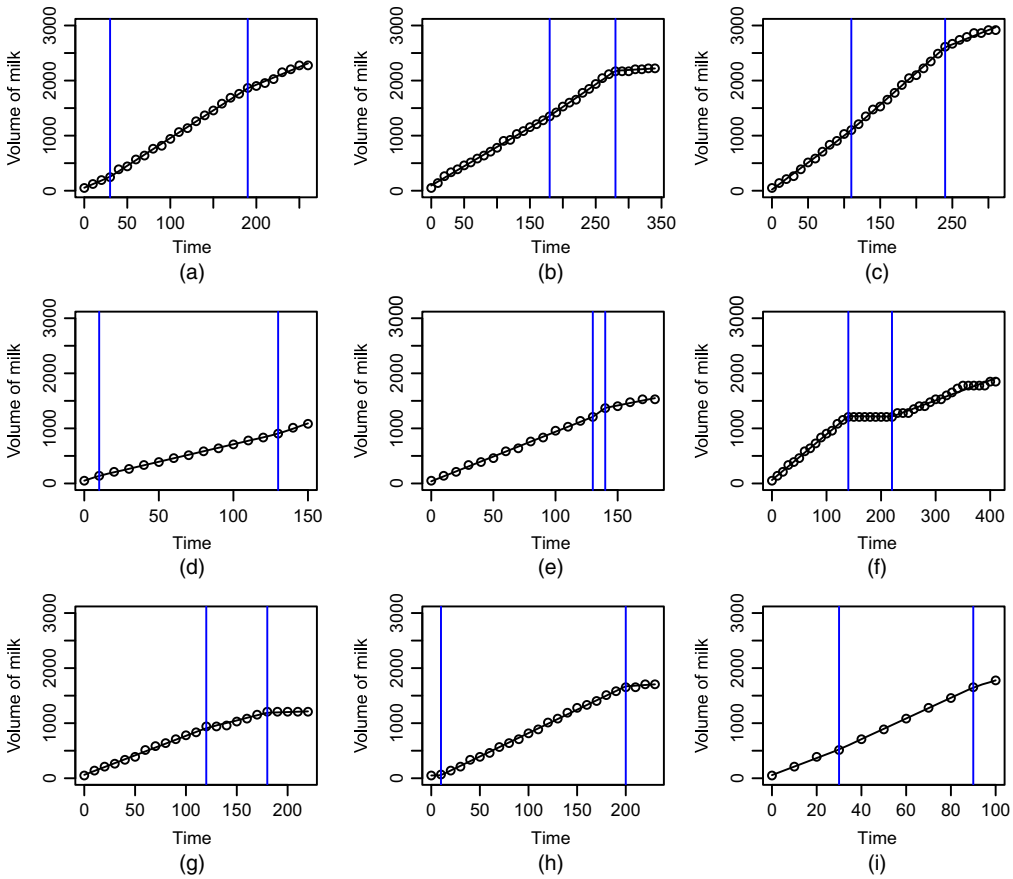


**Fig. 8.** Some examples of milking kinetics belonging to cluster 3 (○, the data; ———, piecewise linear fit obtained thanks to our method; |, position of the change points): (a) goat 250053006212104; (b) goat 250053006212103; (c) goat 250053006212105; (d) goat 250053006213006; (e) goat 250053006213022; (f) goat 250053006213073; (g) goat 250053006213066; (h) goat 250053006213109; (i) goat 250053006213124

## 4. Application

In this section, we apply the methodology that was described in Section 2 to milking kinetics of dairy goats coming from the experimental herd of the Systemic Modelling Applied to Ruminants research unit (Paris, France).

### *4.1. Data description*

The data set contains 100470 milking kinetics of goats of two different breeds: 'Alpine' and 'Saanen'. All these kinetics are morning milking kinetics and several kinetics are available for each goat. The kinetics can also be separated according to parity, which corresponds to the lactation rank, i.e. to the number of times that a goat has given birth and started a new lactation. In the data set considered, there are 276 goats for parity 1 and 191 for parity 2. Goats in parity 2 are also in parity 1 and the kinetics of each goat are observed every day during around 5 months.

### *4.2. Kinetics clustering*

First, note that, on the basis of the shapes of the milking kinetics of this data set, the parameter

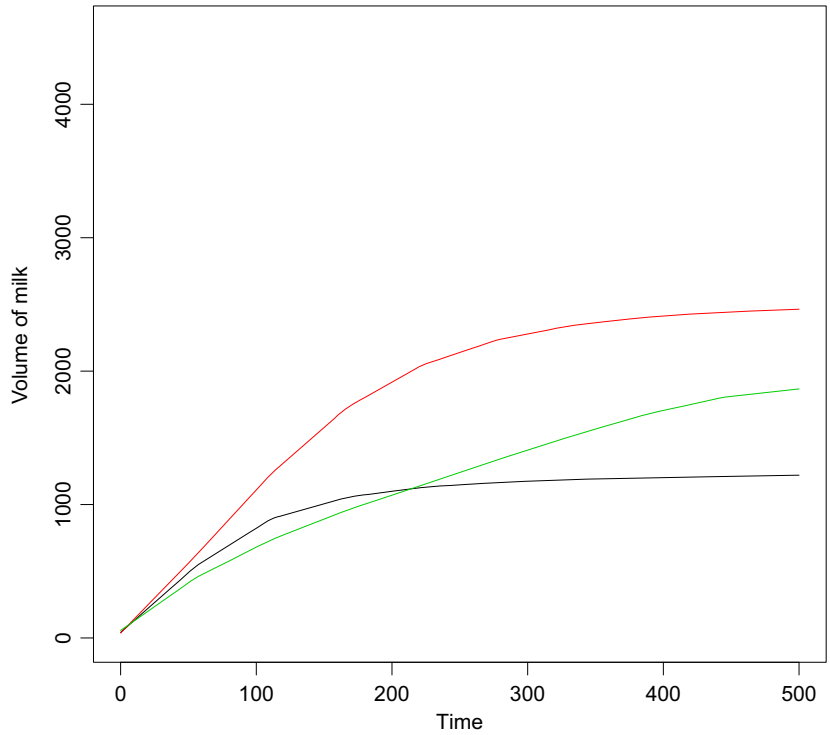


**Fig. 9.** Kinetics average obtained within each of the three clusters: ——, cluster 1; ——, cluster 2; ——, cluster 3

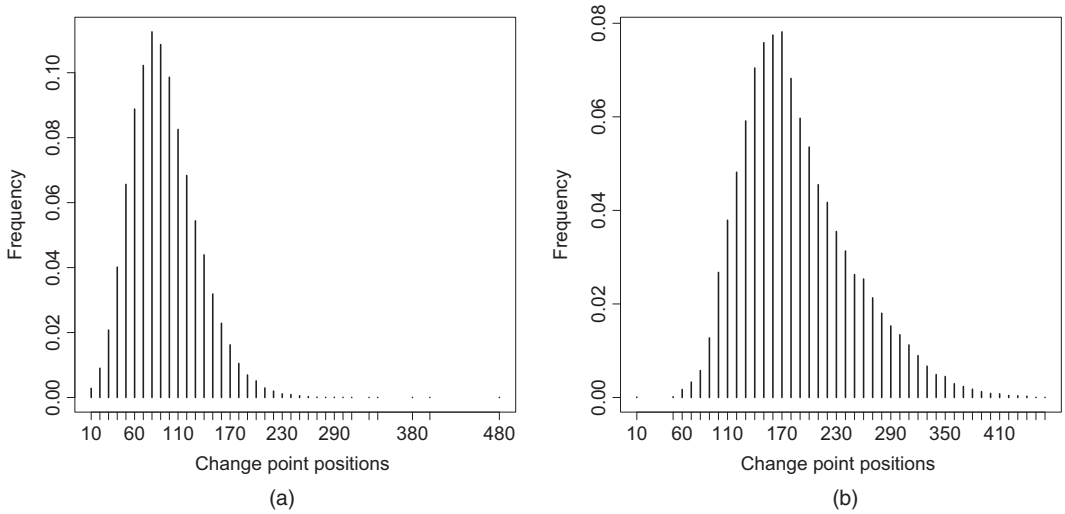


**Fig. 10.** Histograms of the change point positions for (a) cluster 1 and (b) cluster 2

$K_{\max}$ that was defined in the first stage of the first step in Section 2 was set to 2. We obtained three clusters containing 57498, 36757 and 6215 kinetics. Some examples of kinetics belonging to clusters 1, 2 and 3 are displayed in Figs 6, 7 and 8 respectively. The average of the kinetics estimates that were obtained within each cluster is displayed in Fig. 9. We can observe that the

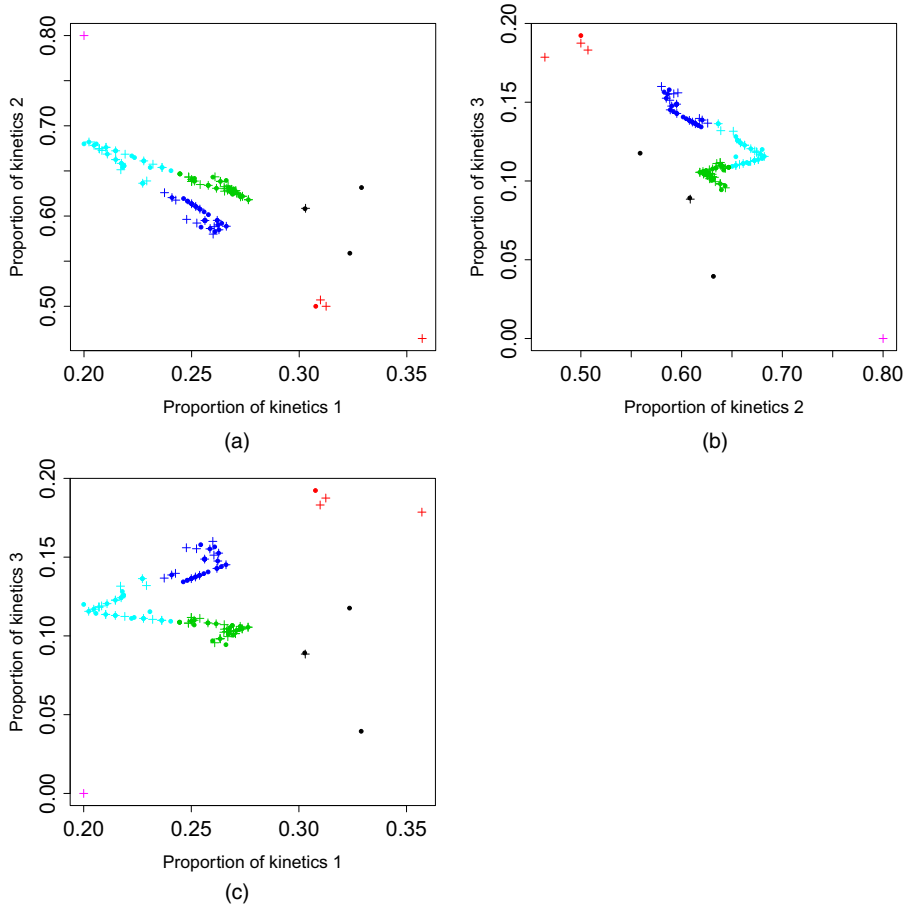


**Fig. 11.** Clustering obtained for goats in parity 1 (six clusters) displayed on the plane having for axes the proportion of kinetics belonging to (a) clusters 1 and 2, (b) 2 and 3, and (c) 1 and 3: ●, ●, ●, Saanen goats; +, +, +, Alpine goats

three clusters can be distinguished in terms of quantity of milk production: cluster 1 has the lowest production, cluster 2 the highest and cluster 3 is between them.

Another difference between the three clusters is the number and the positions of the change points. The number of change points in the kinetics of cluster 1 and cluster 2 is mainly 1 as opposed to cluster 3. Fig. 10 displays the histogram of the change point positions for clusters 1 and 2. We can observe that the change point having the highest frequency is not at the same position for these two clusters. Interestingly, our methodology could distinguish these two clusters thanks to the change point position. This illustrates the potential of our methodology to extract synthetic traits from raw data.

In practice, such clustering may be very useful in the precision farming context to refine selection criteria for breeding programmes, to simplify milking workload or to control udder health. Thanks to the clustering results, we should be able to define a milking profile for each goat. Moreover, we propose in the next section to characterize dairy goats belonging to a given parity.

### *4.3. Parity characterization*

To go further into this analysis, we tried to characterize parities 1 and 2 in terms of the proportion

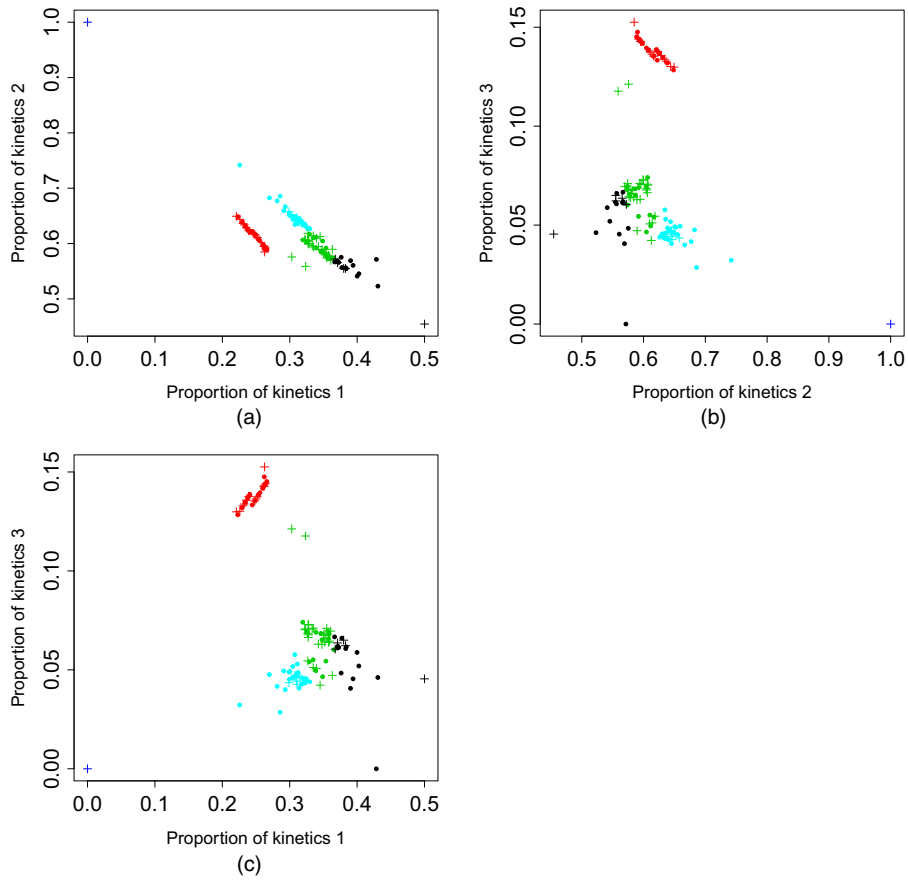


**Fig. 12.** Clustering obtained for goats in parity 2 (five clusters) displayed on the plane having for axes the proportion of kinetics belonging to (a) clusters 1 and 2, (b) 2 and 3 and (c) 1 and 3: ●, ●, ●, Saanen goats; +, +, +, Alpine goats

of kinetics of type 1, 2 or 3 according to the clustering that was previously obtained. We thus created for each goat belonging to a given parity a vector of proportions of its kinetics belonging to cluster 1, 2 or 3. For each parity, the goats are clustered by using the $k$-means algorithm applied to the vectors of proportions. The results are displayed in Figs 11 and 12 for parities 1 and 2 respectively. The number of groups is selected by using the method that was described in Section 2.2. We found six and five groups respectively for parity 1 and parity 2. We can note that there is one goat which produces a large quantity of milk compared with the others for both parities: indeed, in parity 1, 80% of its milking kinetics belongs to cluster 2 and only 20% to cluster 1 whereas, in parity 2, 100% of its milking kinetics belongs to cluster 2.

We also observe from Figs 11 and 12 that, in both parities, the frequency of the milking kinetics belonging to cluster 2 is between 50% and 70%. In parity 2, all the groups have almost the same proportion of milking kinetics belonging to cluster 2 (around 65%). One group (in red) can be distinguished from the others: this group has a higher proportion of milking kinetics belonging to cluster 3 (13%) than do the others (5%). In parity 1, the behaviour is a little different in the sense that the majority of goats have a high proportion of milking kinetics belonging to cluster 2 (around 65%) and a low proportion of milking kinetics belonging to cluster 1 (around

25%). Such results may be interesting in the context of precision breeding since they could help to forecast the production of milk at the different parities.

Further analysis should be perfomed in the future to study how the amounts belonging to each cluster evolve along the lactation course lasting around 150 days in goats. The daily milk yield of a goat for a given parity indeed follows a typical triphasic shape (an increasing, plateau and decreasing phase), each daily milk yield being the sum of the total milk produced during each milking (morning and afternoon milking). Being able to link a particular shape at the milking kinetics scale with a shape at the lactation scale could open perspectives to characterize individual goats better and thus to propose options for individual milking management.

## 5. Conclusion

In this paper, we proposed a novel approach for functional clustering dedicated to the clustering of the milking kinetics of dairy goats. More precisely, we devised a new dimension reduction approach which consists in summarizing each curve (milking kinetics) by a vector containing the coefficients of its projection onto an order 2 $B$-spline basis having estimated knots. These knots correspond to change points estimated by using a novel change-point-estimation method based on the trend filtering approach. Compared with other functional clustering procedures, our main contribution consists in adding these estimated change points in the dimension reduction step. In the course of this study, we have shown that our method has two main features which make it very attractive. Firstly, it is very efficient in terms of statistical performance. Secondly, its very low computational burden makes its use possible on very large data sets.

## References


Abraham, C., Cornillon, P. A., Matzner-Løber, E. and Molinari, N. (2003) Unsupervised curve clustering using b-splines. *Scand. J. Statist.*, **30**, 581–595.

Auger, I. and Lawrence, C. (1989) Algorithms for the optimal identification of segments neighborhoods. *Bull. Math. Biol.*, **51**, 39–54.

Bai, J. and Perron, P. (2003) Computation and analysis of multiple structural change models. *J. Appl. Econ.*, **18**, 1–22.

Bellman, R. (1961) On the approximation of curves by line segments using dynamic programming. *Communs ACM*, **4**, no. 6, 284.

Bouveyron, C., Côme, E. and Jacques, J. (2015) The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Statist.*, **9**, 1726–1760.

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A. (2014) NbClust: an R package for determining the relevant number of clusters in a data set. *J. Statist. Softwr.*, **61**, no. 6, 1–36.

Fearnhead, P., Maidstone, R. and Letchford, A. (2019) Detecting changes in slope with an l0 penalty. *J. Computnl Graph. Statist.*, **28**, 265–275.

Harchaoui, Z. and Lévy-Leduc, C. (2008) Catching change-points with lasso. In *Advances in Neural Information Processing Systems 20* (eds J. C. Platt, D. Koller, Y. Singer and S. T. Roweis), pp. 617–624. Red Hook: Curran Associates.

Harchaoui, Z. and Lévy-Leduc, C. (2010) Multiple change-point estimation with a total variation penalty. *J. Am. Statist. Ass.*, **105**, 1480–1493.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A $K$-means clustering algorithm. *Appl. Statist.*, **28**, 100–108.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn. New York: Springer.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classificn*, **2**, 193–218.

Jacques, J. and Preda, C. (2013) Funclust: a curves clustering method using functional random variables density approximation. *Neurocomputing*, **112**, 164–171.

Jacques, J. and Preda, C. (2014a) Functional data clustering: a survey. *Adv. Data Anal. Classificn*, **8**, 231–255.

Jacques, J. and Preda, C. (2014b) Model-based clustering for multivariate functional data. *Computnl Statist. Data Anal.*, **71**, 92–106.

Killick, R., Fearnhead, P. and Eckley, I. (2012) Optimal detection of changepoints with a linear computational cost. *J. Am. Statist. Ass.*, **107**, 1590–1598.
Krzanowski, W. J. and Lai, Y. T. (1988) A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, **44**, 23–34.
Lavielle, M. (2005) Using penalized contrasts for the change-point problem. *Signl Process.*, **85**, 1501–1510.
Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2016) On optimal multiple changepoint algorithms for large data. *Statist. Comput.*, **27**, 519–533.
Marnet, P. G., Billon, P., Sinapsis, E., Da Ponte, P. and Manfredi, E. (2005) *Machine Milking Ability in Goats: Genetic Variability and Physiological Basis of Milk Flow Rate*, pp. 15–24. Rome: International Committee for Animal Recording.
Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.
Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.-J. (2005) A statistical approach for array CGH data analysis. *BMC Bioinform.*, **6**, no. 27, article 1.
Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis*. New York: Springer Science and Business Media.
Rigaill, G. (2015) A pruned dynamic programming algorithm to recover the best segmentations in 1 to Kmax changes. *J. Soc. Fr. Statist.*, **156**, no. 4, 180–205.
Romero, G., Panzalis, R. and Ruegg, P. (2017) Relationship of goat milk flow emission variables with milking routine, milking parameters, milking machine characteristics and goat physiology. *Animal*, **11**, 2070–2075.
Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L. and Martin, P. (2018) Clustering multivariate functional data in group-specific functional subspaces. *Preprint*. (Available from `https://hal.archives-ouvertes.fr/hal-01652467`.)
Tibshirani, R. J. (2014) Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.*, **42**, 285–323.