XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24–25–26 de outubro de 2013

CLUSTER ANALYSIS OF REGRESSION CURVES

Nora M. Villanueva[1], Marta Sestelo[1] e Javier Roca-Pardiñas[1]

[1]Departamento Estatística e Investigación Operativa. Universidade de Vigo.

RESUMO

A clustering analysis can be perform over regression curves in a nonparametric context. We afford an approach based on estimating and classifying regression curves, if any, into a number of clusters, say K. Classical methods such as k-means can be suitably used for this purpose. The choice of the optimal number of clusters is a challenge problem. In this study, we propose an efficient clustering procedure to classify the regression curves both fixed K *a priori* and unknown K in advance. Our procedure for choosing the number of clusters is a bootstrap based test. Additionally, we demonstrate via a simulation study the well-performance forthe classification of the curves and the choice of the number of clusters. Finally, the methodology was applied to barnacle dataset collected from the Atlantic coast of Galicia.

Palabras e frases chave: Regression curves, k-means, clustering, bootstrap, local polynomial kernel smoothers.

1. INTRODUCTION

One of the main goals of statistical modelling is to understand the dependence of a response variable, Y, with respect to another explanatory variable, X. This type of dependence can be studied through nonparametric regression models, where the relationship between Y and X is modelled without specifying in advance the function that links them. In some situations, study of the regression curves obtained from these type of models can be useful on the comparison of two or more groups, which is an important problem associated with statistical inference.

Let (X_j, Y_j) be J independent random vectors, and assume that they satisfy the following nonparametric regression models, for $j = 1, \ldots, J$

$$Y_j = m_j(X_j) + \varepsilon_j \tag{1}$$

where the error variable ε_j have mean zero and $m_j(X_j) = E(Y_j|X_j)$ is the unknown regression function. Suppose that the covariates X_j have common support R_X. From the formulation of the model in (1) gives rise to several questions of interest, can these curves be classified into groups or clusters and, if so, how many of them?.

Several approaches have beed described in the literature related with cluster analysis techniques (Everitt, 1980). These techniques tend to achieve accuracy in identification of homogeneous groups of individuals in a dataset. Some examples widely known and used in different applications are the k-means algorithm (Macqueen, 1969) or Forgy method (Forgy, 1965). However, these procedures usually require the user to specify the number of clusters K in advance, and it is considered to be one of the biggest drawbacks of these methods.

Our primary objective in this study is to classify regression curves into clusters selecting the optimal number of them by bootstrap techniques (Efron, 1979).

2. METHODOLOGY

The regression model in (1) is estimated using local polynomial kernel smoothers (Fan and Gijbels, 1996; Wand and Jones, 1995). These smoothing techniques estimate the value of the regression curve at a certain point x by averaging locally the values of the response corresponding to the values of the covariate which are close to x.

Given a i.i.d. sample $\{(X_{ij}, Y_{ij})\}_{i=1}^{n_j}$ from (X_j, Y_j), for $j = 1, \ldots, J$ and denote $n = \sum_{j=1}^{J} n_j$, the estimate of one m_j at a point x is given by $\hat{m}_j(x) = \hat{\alpha}_0(x)$, where $\hat{\alpha}_0(x)$ is the first position of the vector $(\hat{\alpha}_0(x), \hat{\alpha}_1(x), \ldots, \hat{\alpha}_R(x))$ which is the minimiser of

$$\sum\nolimits_{i=1}^{n_j} \left\{ Y_{ij} - \sum\nolimits_{r=0}^{R} \alpha_r(x)(X_{ij} - x)^r \right\}^2 \cdot K\left(\frac{X_{ij} - x}{h_j}\right), \tag{2}$$

where K is a kernel function (normally, a symmetric density), h_j is the smoothing parameter or bandwidth and R is the degree of the polynomial.

It is well known that the nonparametric estimates depend heavily on the bandwidths $h_1, \ldots, h_J$ used in the kernel-based algorithm for the estimation of the functions $m_1, \ldots, m_J$. Various methods for an optimal selection have been suggested, such as Generalised Cross-Validation (Golub *et* al., 1979). As a practical solution, the bandwidths used in the nonparametric estimates are automatically selected by cross-validation. However, the use of this technique implies a high computational cost and binning techniques are used to speed up computation.

As already mentioned in the Introduction, the main goal of the present paper is to classify the initial regression curves through a certain number of clusters, say K. If these curves are equal, all of them belong to the same group or cluster. By contrast, if they are different, the curves m_j $(j = 1, \ldots, J)$ will be clustered in more than one cluster.

For this purpose, it is required to obtain a function $\pi : \{1, \ldots, J\} \longrightarrow \{1, \ldots, K\}$, so that each curve m_j is associated to the cluster $\pi(j) \in \{1, \ldots, K\}$. Thus, we propose the use of one of the most popular iterative clustering methods, called the k-means algorithm, which aims to partition the input observations into clusters in which each observation belongs to the cluster with de nearest mean. However, instead of a set of observations, we consider the $\hat{m}_1, \ldots, \hat{m}_J$ curves as input.

Accordingly, given a function π, it is possible to rewrite the model in (1) as follows

$$Y_j = c_{\pi(j)}(X_j) + \varepsilon'_j, \tag{3}$$

where $c_1, \ldots, c_K$ are the functions called centroids and the error variable ε'_j have zero mean.

At this point, having classified the regression functions into a number of clusters —fixed *a priori*— the question that arises in this type of procedures, and has not been totally solved, is to select the optimal number of clusters K. Several strategies for this purpose have been described over the years. A very comprehensive comparative study was carried out by Milligan and Cooper (1985). However, in our knowledge none of the existing methods of determining the number of clusters is completely satisfactory and furthermore, these techniques cannot be employed when data are functions or curves.

We introduce a bootstrap based test that allows us to determine how many clusters are needed. For a given number K, we are interested in testing the null hypothesis of clustering the initial regression curves into K cluster *versus* the alternative in which these curves can be assigned to more than K clusters. For example, to test wether the initial curves can be classified into $K = 1$ clusters is equivalent to say that all the curves are equal. In other words, if this hypothesis is true, this means that all the curves would be included in a unique cluster assuming the equality of them. When the previous hypothesis is rejected, it will be required to test $K + 1$ clusters. If this new hypothesis is again rejected, $K + 2$ clusters should be tested and so on until a certain null hypothesis is accepted.

To test null hypothesis five test statistics are considered:

$$T_1 = \sup_j \sum_{i=1}^{n_j} \left(\hat{m}_j(X_{ij}) - \hat{c}_{\pi(j)}(X_{ij}) \right)^2,$$

$$T_2 = \sup_j \sum_{i=1}^{n_j} \mid \hat{m}_j(X_{ij}) - \hat{c}_{\pi(j)}(X_{ij}) \mid,$$

$$T_3 = \sup_k \sum_{\substack{j=1 \\ j:\pi(j)=k}}^{J} \sum_{i=1}^{n_j} \mid \hat{m}_j(X_{ij}) - \hat{c}_{\pi(j)}(X_{ij}) \mid,$$

$$T_4 = RSS_0 - RSS_1 \quad (\text{Dette}, 1999),$$

$$T_5 = \frac{RSS_0 - RSS_1}{RSS_1} \quad (\text{Fan\&Jiang}, 2005),$$

being $RSS_0 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \hat{c}_{\pi(j)}(X_{ij}))^2$ and $RSS_1 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \hat{m}_j(X_{ij}))^2$.

The procedure to test the null hypothesis is as follows: first, for $j = 1, \ldots, J$, estimate m_j according to the model in (1) using the local polynomial kernel smoothers described above. Second, obtain the function $\pi(j)$ with the k-means algorithm and consequently, estimate the centroids $c_1, \ldots, c_K$. Finally, it is possible to compute the T value —referring to T_1, T_2, T_3 , T_4 and T_5.

It is important to emphasize that, if the null hypothesis is verified, T should be close to zero. The decision rule based on T consists of rejecting the null hypothesis if T is larger than its $(1-\alpha)$-percentile obtained under the null hypothesis. The bootstrap methods are used to approximate the critical values of T, specially the wild bootstrap. The steps are as follows:

Step 1. Compute the T value from the sample as explained above.

Step 2. For $b = 1, \ldots, B$ and for each $j = 1, \ldots, J$ generate bootstrap samples $\left\{X_{ij}, Y_{ij}^{\bullet b}\right\}_{i=1}^{n_j}$ with $Y_{ij}^{\bullet b} = \hat{c}_{\pi(j)}(X_{ji}) + \varepsilon_{ij}^{\bullet b}$, and $\varepsilon_{ij}^{\bullet b}$ being

$$\varepsilon_{ij}^{\bullet b} = \begin{cases} \hat{\varepsilon}_{ij} \cdot \frac{(1-\sqrt{5})}{2} & \text{with probability } p = \frac{5+\sqrt{5}}{10} \\ \hat{\varepsilon}_{ij} \cdot \frac{(1+\sqrt{5})}{2} & \text{with probability } p = \frac{5-\sqrt{5}}{10} \end{cases}$$

where $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{c}_{\pi(j)}(X_{ij})$ are the errors under the null hypothesis, and compute $T^{\bullet b}$ the same way as in **Step 1**.

Finally, the decision rule consists of rejecting the null hypothesis if $T > T^{1-\alpha}$, where $T^{1-\alpha}$ is the empirical $(1-\alpha)$-percentile of values $T^{\bullet 1}, \ldots, T^{\bullet B}$ obtained before.

3. SIMULATION STUDY

Here, we describe the simulation results of our proposed procedure. We consider a scenario where the explanatory covariates X_j, for $j = 1, \ldots, 7$, were drawn from an uniform distribution $[-2, 2]$, and the continuous responses Y_j were generated in accordance to

$$Y_j = m_j(X_j) + \varepsilon_j$$

where

$$m_j(X_j) = \begin{cases} X_j^2 & \text{if} \quad j \in \{1, 2\} \\ 2(X_j^2 - 1) & \text{if} \quad j \in \{3, 4\} \\ 4 - X_j^2 & \text{if} \quad j \in \{5\} \\ 4 - X_j^2 + aX_j & \text{if} \quad j \in \{6, 7\}, \end{cases}$$

ε_j are the errors distributed in accordance to a $N(0, \sigma_j)$ with $\sigma_j(x) = 0.3$, and a is a constant ranging form 0 to 2. 1000 independent samples $\{(X_{ij}, Y_{ij})\}_{i=1}^{n_j}$ were generated from the above model.

To determine the critical values of the test we applied the bootstrap method. Specifically, this entailed 3000 bootstrap samples for calculating type I error and 1000 bootstrap samples for calculating the power under the alternative. Additionally, the performance of the test was checked for different nominal levels (1, 5, 10, 15 and 20%) and for different sample sizes (n_j = 25, 50, 100 and 200).

We explore the validity of the test considering the null hypothesis of classifying the curves into three groups ($K = 3$) under the above model. The value $a = 0$ corresponds to the null hypothesis and as the a value rises, so does the number of clusters to the alternative hypothesis.

Table 1: Type I error (%) for the test based on T, referring to T_1, T_2, T_3, T_4 and T_5.

		Level				
n_j	T	1%	5%	10%	15%	20%
25	T_1	0.1	4.4	9.1	15.3	20.9
	T_2	0.7	5.4	10.7	16.5	22.1
	T_3	0.9	5.6	10.3	17.5	23.4
	T_4	0.7	4.9	12.9	20.1	27.2
	T_5	0.5	3.6	9.6	15.6	20.6
50	T_1	0.8	6.1	11.6	15.9	21.0
	T_2	0.6	6.8	11.4	16.0	21.8
	T_3	1.1	5.3	11.7	16.7	21.8
	T_4	0.9	6.2	11.2	17.5	22.8
	T_5	1.0	5.4	9.9	15.2	20.4
100	T_1	1.0	5.4	10.4	16.8	23.8
	T_2	0.9	6.2	11.2	17.4	24.1
	T_3	0.6	5.8	11.2	16.2	22.3
	T_4	1.2	7.2	12.9	18.4	23.5
	T_5	1.1	6.7	12.1	17.4	22.4
200	T_1	0.5	4.6	10.2	15.8	20.8
	T_2	0.5	5.0	9.6	14.9	20.1
	T_3	0.9	5.0	9.7	14.7	19.6
	T_4	0.7	4.4	10.3	15.6	21.7
	T_5	0.7	4.4	10.0	15.2	21.2

We study the type I error of T_1, T_2, T_3, T_4 and T_5 and the results obtained (expressed in %) are shown in Table 1. All the test statistics perform similarly and reasonably well in most cases, with the level being hold or coming fairly close to the nominal size, especially with large sample sizes. Figure 1 shows the results comparing the power of the proposed test statistics. As can be seen, the behaviour of the power is as expected, i.e. with the increase in the value of a and sample size the proportion of rejections rises. It seems that T_2 performs slightly greater power than the others.

3. APPLICATION TO BARNACLE DATASET

This study was conducted on the Atlantic coast of Galicia (Northwest Spain), which consists of an approximately 1000km long shoreline with extensive rocky stretches exposed to tidal surge and wave action that are settled by the *Pollicipes pollicipes* (Gmelin, 1789) populations targeted for study. Specimens were collected from five sites of the region's Atlantic coastline: Punta do Mouro, Punta Lens, Punta de la Barca, Punta del Boy and Punta del Alba (Figure 2). Two biometric variables of each specimen were measured: RC (Rostro-carinal length, maximum distance across the capitulum between the ends of the rostral and carinal plates) and LT (total length). The idea of this study is to know the relation between both variables, i.e. if the barnacle's growth is similar in all locations or by contrast, if it is possible to detect geographical differentiation in growth.

For each location, nonparametric regression curves were estimated to modelling the dependence between RC and LT (Figure 3). In order to classify them into clusters, we used the proposed procedure obtaining the following results: three estimated curves were attributed to cluster 1 (Punta do Mouro, Punta Lens and Punta del Alba) and two estimated curves were assigned to cluster 2 (Punta de la Barca and Punta del Boy) (Figure 4). As can be seen in Figure 4, the specimens from Punta de la Barca and Punta del Boy have similar morphology. It can be possible because these zones present similar oceanographic characteristic, such as exposed rocky shore.

Acknowledgements Funding for this work was provided by Ministry of Science and Innovation of Spain (MTM2008-03129 and MTM2011-23204) and by Xunta de Galicia (10PXIB300068PR) grants.

BIBLIOGRAPHY

Everitt B. S. (1980). Cluster Analysis. Second Edition, Heinemann, London.

Efron B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1–26.

Fan J., Gijbels I. (1996). Local polynomial modelling and its applications. Chapman and Hall, London.

Fan J., Marron J. (1994). Fast implementation of nonparametric curve estimators. Journal of Computational and Graphical Statistics 3, 35–56.

Forgy E.W. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classification. Biometrics, 21(3), 768–769.

Golub G., Heath M. Wahba G. (1974). Generalized cross-validations as a method for choosing a good ridge parameter. Technometrics, 21(2), 215–223.

Macqueen J.B. (1967). Some methods of classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

Milligan, G., Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159–179.

Sestelo M., Roca-Pardias J. (2011). A new approach to estimation of the length weight relationship of *Pollicipes pollicipes* (Gmelin, 1789) on the Atlantic coast of Galicia (Northwest Spain): some aspects of its biology and management. Journal of Shellfish Research, 30(3), 939–948.

Wand M. P., Jones, M. C. (1995). Kernel smoothing. Chapman and Hall, London.

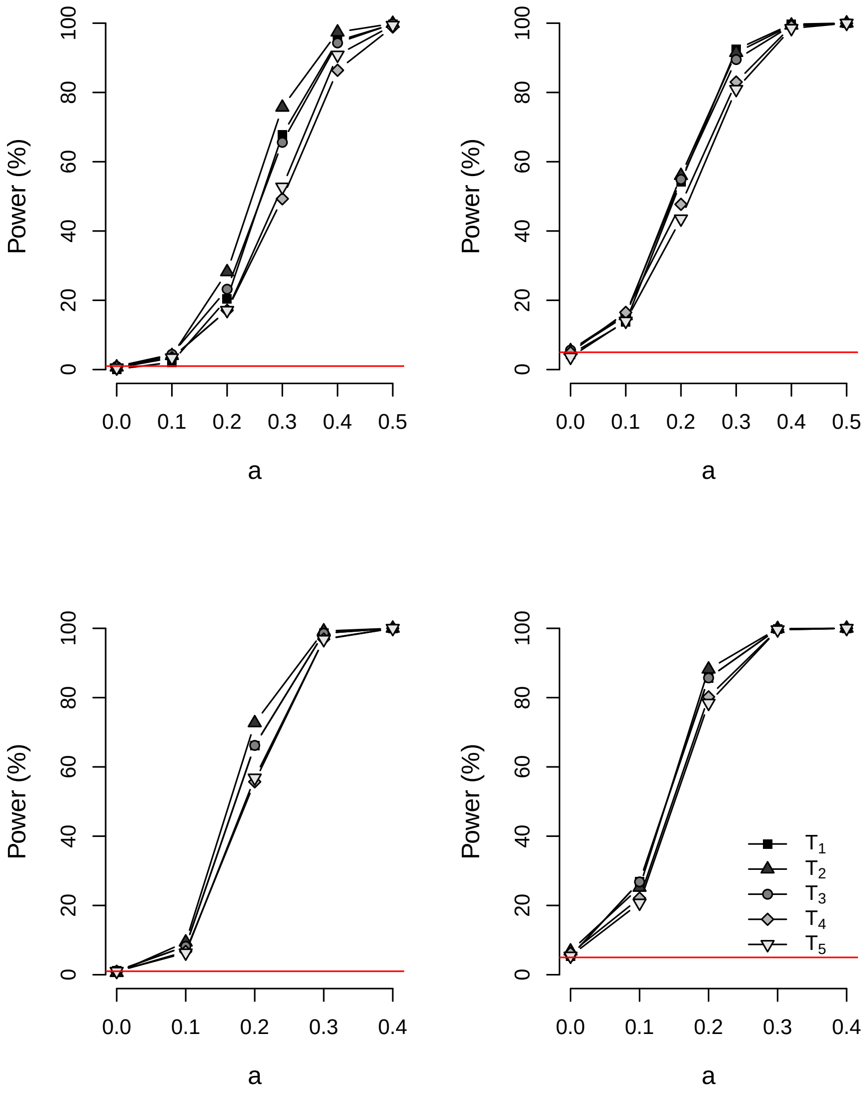

Figure 1: Comparing the performance of all the test statistics for 1 and 5% (first and second column, respectively), and for $n_j = 25$ and 50 (first and second row, respectively).

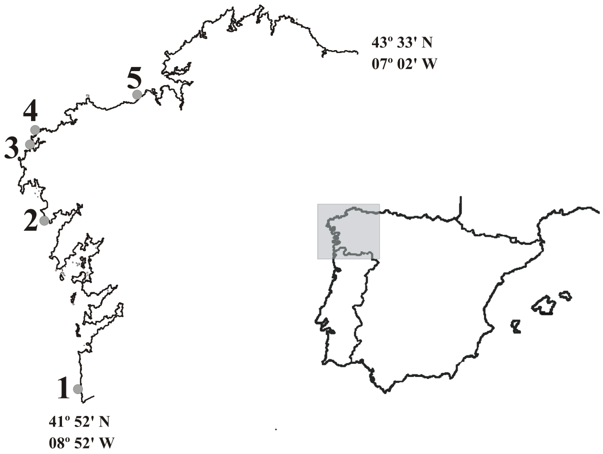

Figure 2: Sampling sites: (1) Punta do Mouro, (2)Punta Lens, (3) Punta de la Barca, (4) Punta del Boy and (5) Punta del Alba.

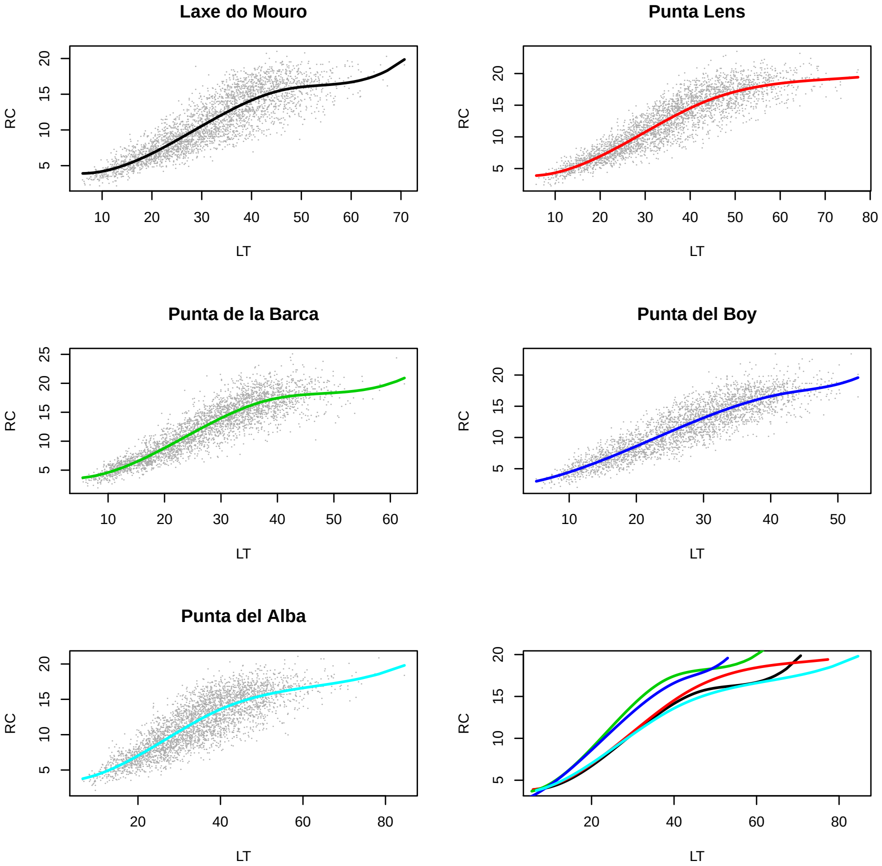

Figure 3: Estimated regression curves colored by location.

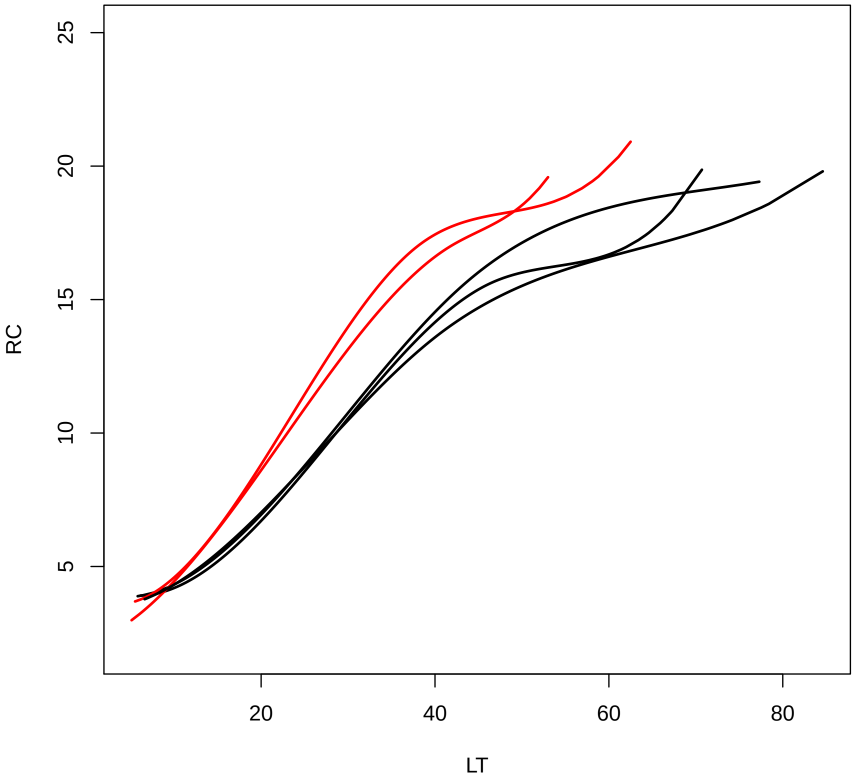

Figure 4: For each location, estimated regression curves colored by clusters. Red lines: estimated curves classified into the cluster 1. Black lines: estimated curves classified into the cluster 2.