Journal of Classification (2022) 39:55–77
https://doi.org/10.1007/s00357-021-09394-5

Partition of Interval-Valued Observations Using Regression

Fei Liu[1] · **L. Billard**[2]

Accepted: 11 June 2021 /Published online: 28 August 2021

Abstract
Both regression modeling and clustering methodologies have been extensively studied as separate techniques. There has been some activity in using regression-based algorithms to partition a data set into clusters for classical data; we propose one such algorithm to cluster interval-valued data. The new algorithm is based on the k-means algorithm of MacQueen (1967) and the dynamical partitioning method of Diday and Simon (1976), with the partitioning criteria being based on establishing regression models for each sub-cluster. This also depends on distance measures between the underlying regression models for each sub-cluster. Several types of simulated data sets are generated for several different data structures. The proposed k-regressions algorithm consistently out-performs the k-means algorithm. Elbow plots are used to identify the total number of clusters K in the partition. The new method is also applied to real data.

Keywords Clusters · k-means algorithm · k-regressions algorithm · Hausdorff distance · City-block distance · Center distance · Simulation methods · Real-data application

1 Introduction

In this article, we adapt the basic dynamical partitioning technique of Diday (1971a, b) and Diday and Simon (1976) for partitioning n observations into K clusters, to develop appropriate algorithms based on regression criteria for interval-valued observations. Interval data are examples of symbolic data first introduced by Diday (1987). Diday's (1971b) dynamical partitioning technique extends the k-means method of MacQueen (1967) for classical data, with several authors (e.g., Chavent et al., 2002) developing k-means methods for selected classes of symbolic data. There are a few papers in the literature which use regression ideas

✉ Fei Liu
louify@gmail.com

L. Billard
lynne@stat.uga.edu

[1] Bank of America, Charlotte, NC 28202, USA

[2] Department of Statistics, University of Georgia, Athens, GA 30602, USA

as a basis for the partitioning criteria in the k-means approach for classical data, starting with the initial development in Charles (1977). We introduce regression based methods to the k-means technique for interval data.

We start in Section 2 with some background literature to the k-means method, and its application to regression methodology. Then, in Section 3, the proposed algorithm for regression-based clustering for interval observations is described. Simulation studies are presented in Section 4; these are used to simulate different sets of data structures. Applications to real data in Section 5 nicely verify the integrity of the proposed algorithm. Advantages and disadvantages are also discussed. Some concluding remarks are found in Section 6.

2 Background

Cluster analysis is a common statistical tool that divides a population (of size n, say) into different sub-populations such that the observations within the same sub-population are as homogeneous as possible while observations from different sub-populations are as non-homogeneous as possible. The fundamental and most well-known clustering method for the partitioning process is the k-means algorithm originally proposed by MacQueen (1967). For a fixed K, the k-means algorithm requires initial clusters to start the process. This initialization could be K seeds or K clusters; a detailed discussion can be found in Anderberg (1973) and Cormack (1971). Then, the algorithm partitions the n observations into K clusters based on the rules under which an observation belongs to a cluster with the nearest mean. A summary of the k-means algorithm and its extensions could be found in Jain et al. (1999), Bock (2007, 2008) and Jain (2010).

Similar to k-means clustering, the cluster-wise linear regression method tries to recover the data structure where the observations are clustered using multiple linear regression models. Cluster-wise linear regression partitions the n observations into K subsets where each observation belongs to its nearest linear regression model. For classical data, the cluster-wise linear regression method is one of the most developed clustering methods in statistics. Analogously with the k-means algorithm, Späth (1979, 1981, 1982) partitioned the data into K subsets and fitted K linear regressions such that the total sum of squares of the errors is locally minimized. DeSarbo and Cron (1988) utilized a maximum likelihood methodology using a mixture of conditional normal distributions to choose the appropriate partition that maximizes the likelihood function, which resulted in a fuzzy cluster-wise linear regression method. The assumptions for ordinary linear regression modeling apply to the cluster-wise linear regression. Wedel and Kistemaker (1989) proposed another maximum likelihood methodology by which a particular observation can belong to only one cluster. Later, Tibshirani et al. (2001) and Shao and Wu (2005) explored methods of determining the number of clusters for a cluster-wise linear regression clustering approach. Zhang (2003) introduced k-harmonic means clustering for the cluster-wise linear regression method, which is less sensitive to the choice of initialization. Rao et al. (2007) and Qian and Wu (2011) extended Späth's (1982) method to one that is more robust by applying M-estimation to the linear regression modeling. Bougeard et al. (2017, 2018) used blocks in a partial least squares setting. A good review can be found in Brusco et al. (2008).

Our concern is with interval-valued observations. Intervals are examples of symbolic data, which can be defined broadly as hypercubes or Cartesian products of distributions in p-dimensional space $\mathbb{R}^p$, in contrast to classical data which are points in $\mathbb{R}^p$. While many

observations are naturally symbolic (e.g., species measurements, temperatures recorded as daily minimum and maximum values), most symbolic data today arise as a result of aggregating the large data sets accumulated by the modern computer, into more manageable forms for analyses. How any particular aggregation proceeds will depend on the scientific questions underlying the investigation at hand. A key feature of such data is that they contain internal variation. For example, suppose an aggregation produced a range of individual values across an interval [10, 20], say. By using a summary statistic such as the midpoint (here 15) for subsequent analyses results in a loss of critical information, since, e.g., this interval cannot be distinguished from a second interval [14, 16] which has the same midpoint as the first. An extensive coverage of the types of symbolic data and their fundamental properties can be found in Bock and Diday (2000), Billard and Diday (2003, 2006), Diday and Noirhomme-Fraiture (2008), Noirhomme-Fraiture and Brito (2011), and Diday (2016), with non-technical introductions in Billard (2011, 2014).

It is important to remember that although observations are aggregated into intervals, each aggregated observation within an interval remains as being from some underlying distribution, e.g., normal distribution. Thus, descriptive statistics calculated from these intervals are still point estimates. It is assumed however that those normally distributed (say) observations are uniformly spread across the interval or sub-intervals for histogram-valued data, analogous with the calculations of histograms (or sample means, etc.) of "group" data in elementary statistics courses. Thus, Bertrand and Goupil (2000) obtained the sample mean as a point value (see Eq. 3). Likewise, for interval-valued observations, a sample variance has a point value, as does each sample regression parameter value, based on n observations (see, e.g., Eq. 9, and Eq. 6, respectively).

Extensions of the k-means algorithm of MacQueen (1967) and the adaptive k-means algorithm of Diday and Simon (1976) to symbolic data are based on the traditional k-means criteria involving distances between observations and the centers as the representation of the obtained clusters; some use a median measure rather than the means. Thus, e.g., for interval-valued data, de Carvalho et al. (2006) consider an L_2 distance; Chavent et al. (2002) use Hausdorff distances; de Souza and de Carvalho (2004), de Souza et al. (2004), and de Carvalho and Lechevallier (2009) apply city-block distances; de Carvalho et al. (2004a, b) use Chebychev distances; and de Carvalho and Lechevallier (2009) apply an adaptive k-means method. Also, for histogram data, Verde and Irpino (2007), Irpino et al. (2006) and Košmelj and Billard (2012) consider Mallows' distance for histogram observations; and Korenjak-Černe et al. (2011) and Batagelj et al. (2015) extend the k-means leaders approach to discrete distributions.

3 Regression-Based *k*-Means for Interval Data

3.1 Regression Model

Suppose we have n observations in a data set with response variable Y and p predictor variables $\mathbf{X} = (X_1, ..., X_p)$. All the response and predictor variables are interval-valued random variables. Let X_{ij}, $i = 1, ..., n$, $j = 1, ..., p$, be the ith realization for the jth predictor variable X_j, denoted by $X_{ij} = [x_{ija}, x_{ijb}]$ with $x_{ija}, x_{ijb} \in \mathbb{R}$ and $x_{ija} \leq x_{ijb}$. Similarly, let Y_i be the ith realization for the response variable Y, denoted by $Y_i = [y_{ia}, y_{ib}]$ with $y_{ia} \leq y_{ib}$, $i = 1, \ldots, n$. Then, the multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}'\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta}^* = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the set of regression coefficients of the $p+1$ variables $\mathbf{X}' = (1, X_1, \ldots, X_p)$, and $\boldsymbol{\epsilon}$ is the error interval vector. In Eq. 1 given the data, $\mathbf{Y}$ is an $n \times 1$ vector and $\mathbf{X}'$ is a $(p+1) \times n$ matrix. Equivalently, Eq. 1 may be written as

$$Y - \bar{Y} = (\mathbf{X} - \bar{\mathbf{X}})'\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\mathbf{X} = (X_1, \ldots, X_p)$ and where $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)$ and $\bar{Y}$ are the sample means defined, respectively, by

$$\bar{X}_j = \frac{1}{2n}\sum_{i=1}^{n}(x_{ija} + x_{ijb}),\ \ j = 1, \ldots, p, \quad \bar{Y} = \frac{1}{2n}\sum_{i=1}^{n}(y_{ia} + y_{ib}). \tag{3}$$

See any of the many elementary texts on regression for properties of the model (e.g., Johnson and Wichern, 2007; Draper & Smith, 1966); see also Xu (2010) for interval observations.

From Eq. 2, the error sum of squares can be written as

$$S = \sum_{i=1}^{n}[Y_i - \bar{Y} - (X_{i1} - \bar{X}_i)\beta_1 - \cdots - (X_{ip} - \bar{X}_p)\beta_p]^2, \tag{4}$$

for given observations $i = 1, \ldots, n$. Then differentiating the right-hand-side of Eq. 4 with respect to $\beta_j,\ \ j = 1, \ldots, p$, we have

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{n}[Y_i - \bar{Y} - (X_{i1} - \bar{X}_i)\beta_1 - \cdots - (X_{ip} - \bar{X}_p)\beta_p](X_{ij} - \bar{X}_j), \tag{5}$$

and setting the derivatives to zero and $\beta_j = \hat{\beta}_j$, we can solve the set of p equations in Eq. 5 to obtain the least squares estimators of β_j. In matrix terms, this becomes

$$\begin{aligned}(\hat{\beta}_1, \ldots, \hat{\beta}_p) &= ((\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}}))^{-1}(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{Y} - \bar{Y}),\\ \hat{\beta}_0 &= \bar{Y} - (\hat{\beta}_1\bar{X}_1 + \cdots + \hat{\beta}_p\bar{X}_p),\end{aligned} \tag{6}$$

where the estimator $\hat{\beta}_0$ pertains from Eq. 1.

To obtain the values of the elements in these matrices, let us re-write

$$\begin{aligned}&(\mathbf{X} - \bar{\mathbf{X}})'\ (\mathbf{X} - \bar{\mathbf{X}})\\ &\quad= \begin{pmatrix} \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2 & \cdots & \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}(X_{ip} - \bar{X}_p)(X_{i1} - \bar{X}_1) & \cdots & \sum_{i=1}^{n}(X_{ip} - \bar{X}_p)^2 \end{pmatrix}_{p\times p}\\ &\quad= \left(\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}_{j'})\right)_{p\times p}\\ &\quad= \big(n \times \mathrm{Cov}(X_j, X_{j'})\big)_{p\times p} \quad j, j' = 1, \ldots, p,\end{aligned} \tag{7}$$

and

$$\begin{aligned}(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{Y} - \bar{Y}) &= \left(\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(Y_i - \bar{Y})\right)_{p\times 1}\\ &= \big(n \times \mathrm{Cov}(X_j, Y)\big)_{p\times 1} \quad j = 1, \ldots, p.\end{aligned} \tag{8}$$

Thus, it is clear that these matrix elements are simply n times the sample covariances between two variables. The question now is to find expressions for these elements for given data; i.e., we need to estimate the elements of the $p \times p$ matrix $(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}}) \equiv \mathbf{M}_{xx}$ (say); and likewise for the elements of the $p \times 1$ matrix $\mathbf{M}_{xy} \equiv (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{Y} - \bar{Y})$.

Consider first the diagonal elements $M_{xx}(j, j) = n\text{Cov}(X_j, X_j) \equiv n\text{Var}(X_j) \equiv \text{SS}_{X_j}$, say. Bertrand and Goupil (2000) showed that for given interval observations, this sum of squares is estimated by

$$\text{SS}_{X_j} = (1/3)\sum_{i=1}^{n}[x_{ija}^2 + x_{ija}x_{ijb} + x_{ijb}^2] - n\bar{X}_j^2, \quad \bar{X}_j = (1/2n)\sum_{i=1}^{n}(x_{ija} + x_{ijb}); \tag{9}$$

this can be re-written as

$$\text{SS}_{X_j} = (1/3)\sum_{i=1}^{n}[(x_{ija} - \bar{X}_j)^2 + (x_{ija} - \bar{X}_j)(x_{ijb} - \bar{X}_j) + (x_{ijb} - \bar{X}_j)^2]. \tag{10}$$

When there is only one observation, i.e., when $n = 1$, Eq. 10 becomes

$$\begin{aligned} \text{SS}_{X_j}(n = 1) &= (1/3)[(x_{ija} - \bar{X}_{ij})^2 + (x_{ija} - \bar{X}_{ij})(x_{ijb} - \bar{X}_{ij}) + (x_{ijb} - \bar{X}_{ij})^2], \\ \bar{X}_{ij} &= (1/2)(x_{ija} + x_{ijb}), \end{aligned} \tag{11}$$

where the single i subscript in Eq. 11 is retained. From this, we see that $\text{SS}_{X_j}(n = 1) \neq 0$ unless $x_{ija} = x_{ijb}$, i.e., unless the observation X_{ij} is a classical point observation. That is, $\text{SS}_{X_j}(n = 1)$ represents the internal variation for this interval X_{ij} observation. Summing over all $i = 1, \ldots, n$ observations, we obtain the Within variation, WithinSS_{X_j}, as

$$\text{WithinSS}_{X_j} = (1/3)\sum_{i=1}^{n}[(x_{ija} - \bar{X}_{ij})^2 + (x_{ija} - \bar{X}_{ij})(x_{ijb} - \bar{X}_{ij}) + (x_{ijb} - \bar{X}_{ij})^2]. \tag{12}$$

Re-arranging, we can show that Eq. 12 can be written as

$$\text{WithinSS}_{X_j} = \sum_{i=1}^{n}[(x_{ijb} - x_{ija})^2/12]. \tag{13}$$

If we look at the variation between observations, viz., BetweenSS_{X_j}, we have

$$\text{BetweenSS}_{X_j} = \sum_{i=1}^{n}[(x_{ija} + x_{ijb})/2 - \bar{X}_j]^2. \tag{14}$$

From Eqs. 13 and 14, it is easy to show that

$$\text{SS}_{X_j} \equiv \text{TotalSS}_{X_j} = \text{WithinSS}_{X_j} + \text{BetweenSS}_{X_j} = M_{xx}(j, j) \tag{15}$$

as given in Eq. 10. It is noted that the expressions on the right-hand-side of Eqs. 13 and 14 are conditional on an assumption that the aggregated values within an interval are uniformly spread across the interval. (As an aside, although this is a so-far universally accepted assumption, different assumptions about the internal interval spread would change the formulae of these two expressions).

The off-diagonal elements $M_{xx}(j, j')$ are obtained analogously, by using the sum of products (SP). Thus, we have the Within SP between the random variables X_j and $X_{j'}$, $\text{WithinSP}_{X_j, X_{j'}}$, as

$$\text{WithinSP}_{X_j, X_{j'}} = \sum_{i=1}^{n}[(x_{ijb} - x_{ija})(x_{ij'b} - x_{ij'a})/12] \tag{16}$$

and the Between SP between X_j and $X_{j'}$, $\text{BetweenSP}_{X_j,X_{j'}}$, as

$$\text{BetweenSP}_{X_j,X_{j'}} = \sum_{i=1}^{n}[(x_{ija} + x_{ijb})/2 - \bar{X}_j][(x_{ij'a} + x_{ij'b})/2 - \bar{X}_{j'}]. \quad (17)$$

Hence, adding Eqs. 16 and 17 and rearranging gives, for $j, j' = 1, \ldots, p$,

$$\begin{aligned} M_{xx}(j, j') &= \text{TotalSP} = \text{WithinSP} + \text{BetweenSP}, \\ M_{xx}(j, j') &= \frac{1}{6}\sum_{i=1}^{n}[2(x_{ija} - \bar{X}_j)(x_{ij'a} - \bar{X}_{j'}) + (x_{ija} - \bar{X}_j)(x_{ij'b} - \bar{X}_{j'}) \\ &\quad + (x_{ij'a} - \bar{X}_{j'})(x_{ijb} - \bar{X}_j) + 2(x_{ijb} - \bar{X}_j)(x_{ij'b} - \bar{X}_{j'})]. \end{aligned} \quad (18)$$

When $j = j'$, Eq. 18 simplifies to Eq. 10. For the special case that the observations are classical point values (with, e.g., $a \equiv [a, a]$), these derivations reduce to those for classical data, as they should. The elements of $\mathbf{M}_{xy}$ are obtained similarly. Hence, the least squares estimators for $\boldsymbol{\beta}^*$ are found by substituting the elements of $\mathbf{M}_{xx}$ and $\mathbf{M}_{xy}$ into Eq. 6.

In the sequel, regression model fits will use the estimators obtained from Eqs. 1–18, and will be referred to as the symbolic variation methodology.

Earlier methods developed for fitting regression models to interval-valued data include the center method of Billard and Diday (2000), whereby a model is fitted to the midpoints of the intervals. Although their method used the range of the predictor variables to find the prediction intervals, the calculation of the model parameters is based on the interval midpoints only, and so omits important information contained in the internal variations of the intervals. In an attempt to redress that problem, in a series of papers, de Carvalho and his colleagues (e.g., Lima Neto et al., 2004; de Carvalho et al., 2004a, b; Lima Neto et al., 2005; Lima Neto & de Carvalho, 2008) transformed the original interval-valued variables into point-valued center and half-range variables; then, a classical regression analysis was conducted on each of the center values and half-range values separately. Also, use of the center-range variables is equivalent to using the interval end-points. We note that all these methods use some form of classical surrogates on one/two single point values of the interval, rather than using the entire interval as in the symbolic variation method. It is easy to show that the variations (SS/SP) on the centers equate to the between variations of the symbolic variation methodology but that the variations on the ranges do not equate to the Within variations of the symbolic variation method. Thus, in their various ways, these methods are not fully utilizing all the variations in the data correctly. There are moreover other statistical issues here; e.g., the assumption that the center and range are independent entities is unsustainable. However, while each of these previous methods advanced the subject at the time, each has its limitations.

3.2 Partitioning Criteria

Our concern is with partitioning our data set into a fixed number K clusters where the observations within each cluster are identified by a specific regression model. Assume that the response variable Y has K different linear relationships with the predictor variables $\boldsymbol{X}$. Let $(\boldsymbol{X}_k, Y_k)$, $k = 1, ..., K$, be the set of observations that follows the kth regression model (i.e., belongs to the kth cluster C_k). Then,

$$Y_k = \boldsymbol{X}_k'\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \quad k = 1, ..., K, \quad (19)$$

where $\boldsymbol{\beta}_k$ is the set of linear coefficients of the predictor variables for the kth regression model, and $\boldsymbol{\epsilon}_k$ is the error vector.

Let $n_k, k = 1, ..., K$, be the number of observations in the kth cluster with $\sum_{k=1}^{K} n_k = n$. We assume the following:

A1 The number of observations n_k satisfies $p < n_k \leq n, k = 1, ..., K$, where p is the number of predictor variables, and n is the total number of observations in the whole data set. It can be shown that $n_k = n$ only if $K = 1$.

A2 The individual errors in a particular cluster k are drawn independently from a normal distribution with mean 0 and variance σ_k^2, $N(0, \sigma_k^2)$, and after aggregation (along the lines of Eq. 33 e.g.) become intervals. The error intervals $\boldsymbol{\epsilon}_k$ are independent from $\boldsymbol{\epsilon}_{k'}$, $k \neq k'$, for $k, k' = 1, ..., K$.

The first assumption (A1) avoids the situation with $n_k < p$ such that there is no linear regression solution for the kth cluster; the second assumption (A2) reduces the computational complexity of the problem. Our goal is to find an optimal partition $P^* = (C_1^*, ..., C_K^*)$ that minimizes the overall residuals of the regression models given the number of clusters K.

Given a partition $P = (C_1, ..., C_K)$, we can fit a linear regression model for each cluster as in Eq. 19. Denote the coefficient estimator of $\boldsymbol{\beta}_k$ by $\hat{\boldsymbol{\beta}}_k$ for $k = 1, ..., K$. Then, the regression residuals for the kth cluster are defined as, for $i \in C_k$,

$$r_{ki} = d(Y_i, \hat{Y}_i), \tag{20}$$

where $d(Y_i, \hat{Y}_i)$ is the distance between the observation Y_i and its predicted value $\hat{Y}_i$. Since $\hat{Y}_i = \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}_k$, Eq. 20 can be rewritten as $r_{ki} = d(Y_i, \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}_k)$. The predictive interval $\hat{Y}_i$ using the symbolic variation method is, for $i \in C_k$,

$$\hat{Y}_i = [\hat{Y}_{ia}, \hat{Y}_{ib}] = [\min_{\boldsymbol{X} \in \mathbb{X}} \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}_k, \max_{\boldsymbol{X} \in \mathbb{X}} \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}_k], \tag{21}$$

where $\mathbb{X} = \{\boldsymbol{X} = (x_j) : x_{ja} \leq x_j \leq x_{jb}, j = 1, ..., p\}$; see Xu (2010). Our goal is to find an optimal partition that minimizes the sum of squared residuals (SSR) given K, viz.,

$$\text{SSR} = \arg\min_{P;\ \hat{\boldsymbol{\beta}}_k} \sum_{k=1}^{K} \sum_{i \in C_k} r_{ki}^2 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} r_{ki}^2. \tag{22}$$

Since r_{ki} in Eq. 22 is defined as the distance between two intervals, Y_i and $\hat{Y}_i$, different definitions of the distance between these two intervals will affect the clustering results of our k-regressions algorithm. For illustrative concreteness, we consider three different distance definitions between two interval variables, specifically, center distance, Hausdorff distance and city-block distance; other distances could be used.

The center distance between two p-dimensional interval observations $\mathbf{Z}_1 = (Z_{11}, \ldots, Z_{1p})$ and $\mathbf{Z}_2 = (Z_{21}, \ldots, Z_{2p})$ with $Z_{ij} = [z_{ija}, z_{ijb}], i = 1, 2, j = 1, \ldots, p$, is defined as

$$d_C(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{j=1}^{p} |z_{1j}^c - z_{2j}^c|, \tag{23}$$

where $z_{ij}^c = (z_{ija} + z_{ijb})/2$, is the midpoint of the interval $Z_{ij}, i = 1, 2, j = 1, \ldots, p$. The Hausdorff (1937) distance is defined as

$$d_H(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{j=1}^{p} \max\{|z_{1ja} - z_{2ja}|, |z_{1jb} - z_{2jb}|\}. \tag{24}$$

The city-block distance is defined as

$$d_{CB}(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{j=1}^{p} [|z_{1ja} - z_{2ja}| + |z_{1jb} - z_{2jb}|]. \tag{25}$$

We observe that there are effectively two optimizations occurring in this process. The first is designed to minimize the sum of squared residuals when fitting the regression model within each cluster (i.e., minimize the regression sum of squared errors $\sum_{i=1}^{n_k} \epsilon_i^2$ for each kth cluster regression, $k = 1, \ldots, K$). This is not the same as minimizing $\sum_{i=1}^{n} \epsilon_i^2$ over all observations ignoring the clusters. The second optimization is to minimize the sum of the squared distances $r_{ki} = d(Y_i, \hat{Y}_i)$ where the predicted value $\hat{Y}_i$ is determined by the regression equation for the kth cluster. We also note that the explained variation in the data is due to a mix of explained variation from the regression fits within clusters and that due to heterogeneity across clusters. For the classical data setting, Brusco et al. (2008) has a nice example illustrating how sometimes the explained variation can be dominated by the within cluster variations, and sometimes by the cluster heterogeneity.

3.3 Partitioning Algorithm

The k-regressions algorithm for each of the three distance definitions of Eqs. 23–25 is the same, except that the distance $d(Y_i, \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}_k)$ used in the allocation step (iii) changes appropriately. Analogously with the algorithm in Späth (1979), we propose a k-regressions cluster-wise algorithm for interval-valued data as follows:

(i) *Initialization*: Choose a partition $P^{(0)} = (C_1^{(0)}, ..., C_K^{(0)})$ randomly from all the possible partitions, or partition the whole data set to K clusters based on some prior knowledge.

(ii) *Representation*: For $k = 1, ..., K$, fit regressions $Y_k = \boldsymbol{X}_k'\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$ to the observations in the kth cluster for partition $P^{(l)} = (C_1^{(l)}, ..., C_K^{(l)})$ where $l = 0, 1, \ldots,$ denotes the lth iteration.

(iii) *Allocation*: For observation Y_i, $i = 1, ..., n$, calculate its distance to its prediction $\hat{Y}_i$ obtained by its k^{th} regression line, $d(Y_i, \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}_k)$, $k = 1, ..., K$, and allocate the observation to its closest line. The updated partition is now $P^{(l+1)} = (C_1^{(l+1)}, ..., C_K^{(l+1)})$.

(iv) *Stop*: Repeat (ii) and (iii) until the improvement of SSR in Eq. 22 is smaller than a predetermined criterion, or the number of iterations reaches a predetermined maximum number.

For the representation step, we apply the symbolic variance method to fit the linear regression model. For the allocation step, the observations are allocated such that, for $k = 1, ..., K$,

$$C_k = \{(\boldsymbol{X}, Y) | d(Y, \boldsymbol{X}'\hat{\boldsymbol{\beta}}_k) \leq d(Y, \boldsymbol{X}'\hat{\boldsymbol{\beta}}_{k'}), \forall k \neq k'\}. \tag{26}$$

Given a data set, the algorithm cannot guarantee a global minimum of SSR. Thus, we repeatedly implement the steps (i)–(iv) a number of times with different initializations and select the solution which has the lowest value of SSR. The selected partition can be further iterated until SSR cannot be reduced anymore. An expanded description is available in Liu (2016).

3.4 Number of Clusters

The k-regressions clustering algorithm is used to implement the cluster-wise regression method given that K is known. However, if we do not have prior knowledge about K, a bad guess of K can mislead the results. Xu (2010) gave a symbolic R-square (R^2) of the symbolic variation method for the linear regression of interval-valued data as

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}, \tag{27}$$

where $\hat{Y}$ is the predicted value of the response variable Y, and $\text{Var}(\cdot)$ is the symbolic variance calculated from Eq. 9. Using the symbolic R^2, we propose the following methods to determine the number of clusters K.

Given a predetermined maximum number of clusters K^{max}, for each $K = 1, \ldots, K^{max}$, calculate the R^2 for each cluster $k = 1, \ldots, K$, denoted by $R_k^{2^{(K)}}$. For the whole data set, the weighted average R^2 for the n observations given K is defined as

$$R^{2^{(K)}} = \sum_{k=1}^{K} w_k^{(K)} R_k^{2^{(K)}}, \tag{28}$$

where $w_k^{(K)} = n_k/n$ is the weight of the R^2 for the kth cluster, and $n_k = |C_k|$ is the number of observations in the kth cluster. From the plot of $(1 - R^{2^{(K)}})$ versus K, the elbow point is the optimal number of clusters K^*.

Determining the optimal number of clusters K by looking for the elbow point can be subjective, especially when the elbow point is not obvious. Analogously with the adjusted R^2 for the linear regression model, we propose an adjusted R^2 to determine the optimal K for the k-regressions algorithm. We know that R^2 for ordinary least squares regression models corresponds to the proportion of variation explained by the model; i.e., R^2 can be defined as

$$R^2 = 1 - \text{SS}_{\text{res}}/\text{SS}_{\text{tot}} = \text{SS}_{\text{reg}}/\text{SS}_{\text{tot}}, \tag{29}$$

where $\text{SS}_{\text{tot}} = \sum_i (Y_i - \bar{Y})^2$ is the total sum of squares, $\text{SS}_{\text{reg}} = \sum_i (\hat{Y}_i - \bar{Y})^2$ is the sum of squares of the regression, $\text{SS}_{\text{res}} = \sum_i (Y_i - \hat{Y}_i)^2$ is the sum of squares of the residuals, and $\bar{Y} = \sum_i Y_i/n$ is the sample mean of Y. The R^2 in Eq. 29 can be rewritten as

$$R^2 = 1 - \text{Var}_{\text{res}}/\text{Var}_{\text{tot}}, \tag{30}$$

where $\text{Var}_{\text{res}} = \text{SS}_{\text{reg}}/n$ and $\text{Var}_{\text{tot}} = \text{SS}_{\text{tot}}/n$. The Var_{res} and Var_{tot} terms are both biased estimators of the residual variation and the population variation, respectively. The adjusted R^2 term adjusts these two variance estimators to be unbiased estimators, so that the adjusted R^2 is defined as

$$\bar{R}^2 = 1 - \frac{\text{SS}_{\text{res}}/\text{df}_\epsilon}{\text{SS}_{\text{tot}}/\text{df}_t}, \tag{31}$$

where $\text{df}_\epsilon = n - p - 1$ is the degree of freedom of the residuals, and $\text{df}_t = n - 1$ is the degree of freedom of the population variation. The adjusted R^2 adjusts the R^2 in Eq. 29 so that it does not always increase.

The k-regressions algorithm fits K regressions on the whole data set, so that the total number of parameters is Kp. For each $K = 1, \ldots, K^{max}$, analogously with the idea of an adjusted R^2 for ordinary least squares regression, we define the adjusted weighted R^2 for

the k-regressions clustering as

$$\begin{aligned} \bar{R}^{2^{(K)}} &= R^{2^{(K)}} - (1 - R^{2^{(K)}}) \tfrac{Kp}{n-Kp-1} \\ &\equiv R^{2^{(K)}} - Q^{(K)}, \end{aligned} \tag{32}$$

where $Q^{(K)} = (1 - R^{2^{(K)}})Kp/(n - Kp - 1)$ is the penalty term.

The adjusted weighted R^2 in Eq. 32 penalizes the $R^{2^{(K)}}$ of Eq. 28 by the factor $Q^{(K)}$ when the number of clusters increases. Since we fit K different linear regression models to the whole data set, the number of parameters is $K(p + 1)$ for the cluster-wise regression methodology.

From Eq. 32, $\bar{R}^{2^{(K)}}$ is always smaller than $R^{2^{(K)}}$. The $\bar{R}^{2^{(K)}}$ increases only if the increase of K improves $R^{2^{(K)}}$ more than the penalized term $Q^{(K)}$. Usually when K increases, $\bar{R}^{2^{(K)}}$ increases and reaches a maximum at a certain value of K, and decreases afterwards. The value of K that maximizes $\bar{R}^{2^{(K)}}$, or equivalently minimizes $1-\bar{R}^{2^{(K)}}$, is the optimal number of clusters K^*. We compare the two methods of determining the optimal K by simulation results in the next section.

4 Simulation Study

We describe how our interval-valued observations are simulated in Section 4.1. Then, in Section 4.2, we compare the k-regressions clustering with the traditional k-means method. How the new algorithm performs is studied in Section 4.3 for several different data set structures.

4.1 Methodology

In practice, most interval data sets arise from aggregating classical data. From this perspective, we propose the following simulation method. The intervals of $\boldsymbol{X}$ are simulated where the interval midpoints $X^{(c)}$ come from a multivariate normal distribution, and the interval ranges $X^{(r)}$ are from exponential distributions. The intervals of X_j, $j = 1, ..., p$, are given by $X_j = [X_j^{(c)} - 0.5X_j^{(r)}, X_j^{(c)} + 0.5X_j^{(r)}]$. The spread of observations within these intervals is assumed to be uniform. For a particular observation i, to obtain the interval Y_i, we randomly draw m values from the uniform distribution $U(x_{ija}, x_{ijb})$ for each $j = 1, ..., p$, denoted by $x_{ij1}, ..., x_{ijm}$. The m is a predetermined number. Then, the interval $Y_i = [y_{ia}, y_{ib}]$ is determined by

$$\begin{aligned} y_{ia} &= \min_{l\in\{1,...,m\}} \{\beta_0 + \beta_1 x_{i1l} + ... + \beta_p x_{ipl} + \epsilon_{il}\}, \\ y_{ib} &= \max_{l\in\{1,...,m\}} \{\beta_0 + \beta_1 x_{i1l} + ... + \beta_p x_{ipl} + \epsilon_{il}\}, \end{aligned} \tag{33}$$

where $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, ..., n$, and $l = 1, ..., m$. This method is practically reasonable. For example, traffic on a particular intersection is recorded multiple times everyday; the minimum and maximum values are recorded as the traffic interval for a day.

A more general case for this method is to assume the number m follows a certain distribution $f(m; \lambda)$, say, such as a geometric distribution or a negative binomial distribution. For each observation i, the m's are the same for all the predictors X_j, $j = 1, ..., p$, but the m's are different for different observations. We have $m_i \overset{\text{iid}}{\sim} f(m; \lambda)$ for $i = 1, ..., n$. The interval Y_i, $i = 1, ..., n$, is now given by Eq. 33 but with m replaced by m_i. By allowing a

random value for m, this simulation method fits more general scenarios. For instance, the daily price for a particular stock is an interval where the lower bound is the minimum price while the upper bound is the maximum price. The prices for the stock are recorded on a transaction base for every trading day, but the number of transactions on each day is not fixed. Instead, it is a random number that follows a certain distribution.

A problem for this simulation method is that it cannot guarantee that the obtained intervals Y_i, $i = 1, ..., n$, internally have the aggregated observations uniformly spread across the interval. However, it can be verified that the observations within an interval Y_i obtained in this way are uniformly distributed for a relatively large m, say, $m \geq 3000$; see Xu (2010). The advantage is that this method is close to how interval data sets are collected in practice.

4.2 Comparison of the *k*-Regressions and *k*-Means Algorithms

The k-means algorithm is designed for spherical data structures. When each of the clusters in a data set is not spherical, the algorithm can fail. For example, if the variables are highly correlated within a cluster and the clusters overlap, it is difficult for the k-means algorithm to recover such clusters. In this section, we give two examples (with m small and large, respectively) where the k-means algorithm fails to recover the true clusters while in contrast the k-regressions algorithm succeeds. The k-means clustering method for interval-valued data is based on the algorithms in Chavent et al. (2002) and de Souza and de Carvalho (2004). In each case, we implement the k-means clustering method based on each of the city-block distance and the Hausdorff distance. For the k-regressions clustering method, we use the center distance for demonstration purposes. Like the k-means algorithm, the k- regressions algorithm does not guarantee a proper convergence given a random start partition. Therefore, multiple initial partitions are needed; the one with minimum SSR (see Eq. 22) is deemed to be the convergence result for the k-regressions algorithm. We call an initial partition that converges to the minimum SSR as being a good initial partition.

Data I

Our first data set (I) is composed of three clusters that follow the equations:

$$(1)\ Y = 142 + 5X + \epsilon_1, \quad (2)\ Y = 53 - 3X + \epsilon_2, \quad (3)\ Y = -43 + 0.6X + \epsilon_3, \qquad (34)$$

respectively, where $\epsilon_1 \sim N(0, 15^2)$, $\epsilon_2 \sim N(0, 12^2)$, and $\epsilon_3 \sim N(0, 7^2)$. We apply the simulation method described in Section 4.1 and set $m = 25$. The observations of these three regression models are simulated separately with 200 observations for each, and then the three data sets are stacked into one data set.

Figure 1a shows the three true clusters with the three linear of Eq. 34, respectively. Figure 1b shows the clustering results based on the k-means algorithm when the city-block distance was used, while Fig. 1c gives the k-means clustering results using the Hausdorff distance for the data. From Fig. 1b and c, we see that both these k-means algorithms cluster at the intersection areas between the three clusters in Eq. 34, which are clearly not the correct clusters.

Figure 1d, e, and f show the clustering process of the k-regressions algorithm with a good initial partition. Figure 1d shows the first (initialization) iteration of the k-regressions algorithm, while Fig. 1e shows the third iteration where the algorithm starts to converge to the true linear regression models in Eq. 34. Figure 1f shows the tenth or final iteration of the k-regressions algorithm where the algorithm converges to the three true linear regres-

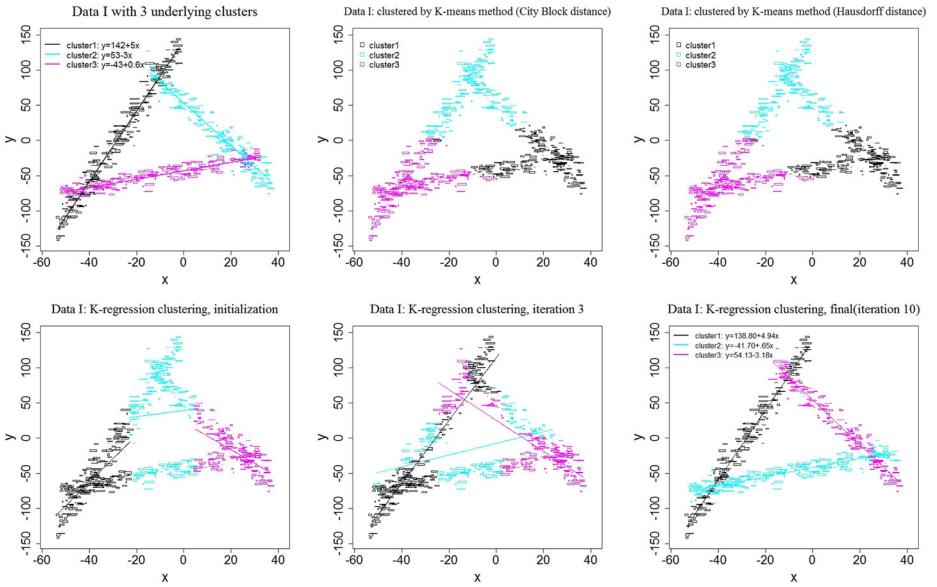

Fig. 1 Comparison between clustering results of k-means algorithm and k-regressions algorithm for Data I

sion models in Eq. 34. The three linear regression models obtained by the k-regressions algorithm are, respectively,

$$(1)\ Y = 138.80 + 4.94X, \quad (2)\ Y = 54.13 - 3.18X, \quad (3)\ Y = -41.7 + 0.65X. \tag{35}$$

These coefficients are close to the true coefficients in Eq. 34. In addition, by comparing the true data set in Fig. 1a and the k-regressions clustering results in Fig. 1f, it is safe to say that the k-regressions algorithm recovers the three true clusters for Data I in Eq. 34. A further investigation shows that all the misclassification observations are from the intersection areas between the three clusters.

Data II

Our second data set (II) contains three clusters. We set $m = 3000$ in Eq. 33. One hundred observations for each regression model are simulated and then all the observations are stacked into one data set. The three linear regression models between the two variables are as follows:

$$(1)\ Y = 150.5 + 4.5X + \epsilon_1, \quad (2)\ Y = 53 - 3X + \epsilon_2, \quad (3)\ Y = -53 + 0.5X + \epsilon_3, \tag{36}$$

respectively, where $\epsilon_1 \sim N(0, 15^2)$, $\epsilon_2 \sim N(0, 12^2)$, and $\epsilon_3 \sim N(0, 7^2)$.

The simulated Data II with a total of 300 observations and 3 clusters is visualized in Fig. 2a where the three true regression lines are also plotted. Figure 2b and c give the clustering results by the k-means algorithm with the city-block distance and the Hausdorff distance. We can see clearly that the k-means algorithm with both the city-block distance and the Hausdorff distance fails to recover the correct clusters.

Figure 2d, e, and f show the progress of the k-regressions algorithm through nine iterations onto the Data II with the good initial partition. Figure 2d is the plot of the three clusters for the first iteration (initialization). Figure 2e shows the third iteration of the algorithm where the three clusters are already close to the three true clusters. The ninth (final) iteration is presented in Fig. 2f and shows clearly the convergence to the correct three clusters

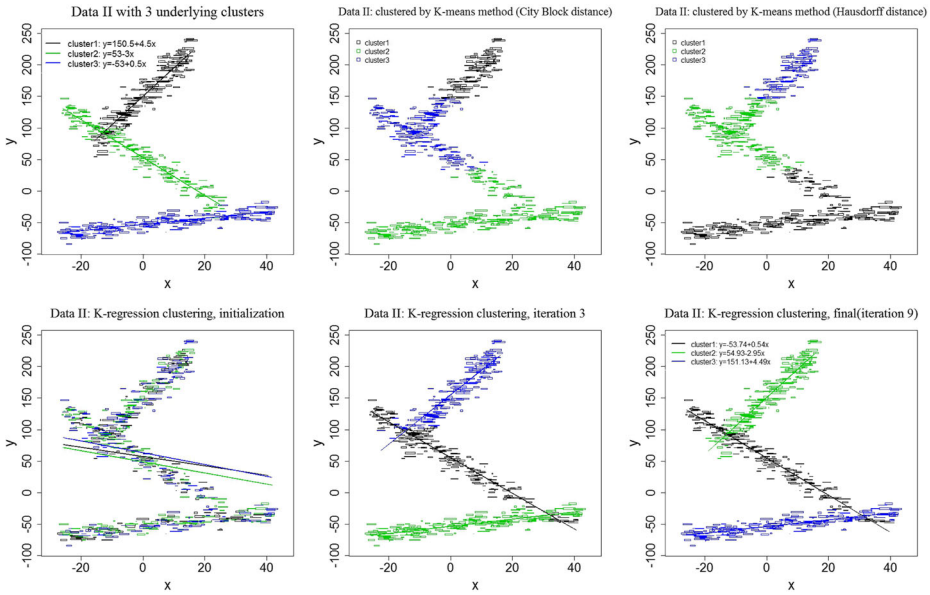

Fig. 2 Comparison between clustering results of k-means algorithm and k-regressions algorithm for Data II

by the algorithm. The estimated linear regression model by the symbolic variation method for the three converged clusters are, respectively,

$$(1)\ Y = 151.13 + 4.49X, \quad (2)\ Y = 54.93 - 2.95X, \quad (3)\ Y = -53.74 + 0.54X. \qquad (37)$$

The estimated coefficients in Eq. 37 and the true coefficients in Eq. 36 are quite close. In addition, by comparing the plot of the original three linear regression models in Fig. 2a and the plot of the converged three clusters in Fig. 2f, it is observed that the k-regressions algorithm successfully recovered the true structure of Data II. Both Data I and II are non-spherical; the k-means algorithm failed to recover the true structure, whereas our method succeeded.

In these simulated examples, we assume that the true number of clusters is known. To decide the optimal number of clusters, we calculate the weighted R-squared, $R^{2^{(K)}}$, for $K = 1, ..., 8$, from Eq. 28. The elbow plot is the plot of $1 - R^{2^{(K)}}$ versus the number of clusters K. Figure 3a and b show the elbow plots for Data I from Eq. 34 and Data II from Eq. 36, respectively. For both data sets, the elbow plots correctly show that the optimal number of clusters is $K = 3$.

4.3 Performance of the *k*-Regressions Algorithm

4.3.1 Data Structures

In this section, we simulate data sets with different structures to investigate the performance of the k-regressions algorithm. We first consider three data sets with $p = 1$. Data A and Data B contain three clusters, while Data C contains four clusters.

Table 1 provides the parameter setup for the three data sets. In Table 1, n is the sample size for each of the clusters; β_0 and β_1 are the coefficients of the linear relation for each

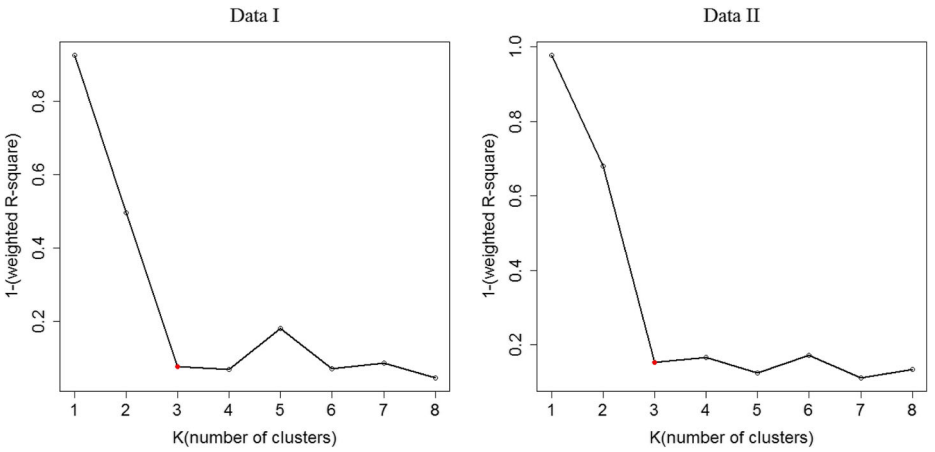

Fig. 3 Determining the number of clusters K by an elbow plot

cluster. The values μ_x and σ_x are the two parameters of the normal distribution $N(\mu_x, \sigma_x)$ from which the interval midpoints of the predictor variable X are drawn. The value λ_x is the parameter of the exponential distribution $\exp(\lambda_x)$ from which the interval ranges of X are drawn. The error terms ϵ_i are drawn from a normal distribution $N(0, \sigma_\epsilon)$ where the values of the parameter σ_ϵ^2 are shown in the row "σ_ϵ" in Table 1.

Thus, the true linear regression equations for the three clusters of Data A have the structure:

$$(1)\ Y = 1.0 + 1.3X,\ \ (2)\ Y = 45 + 1.8X,\ \ (3)\ Y = 45 - 2.5X; \tag{38}$$

those for Data B are as follows:

$$(1)\ Y = 142 + 5X,\ \ (2)\ Y = 33 - 3X,\ \ (3)\ Y = -73 + 0.6X; \tag{39}$$

and those for Data C are as follows:

$$(1)\ Y = 2.0{+}0.8X,\ \ (2)\ Y = 1.0{+}2.3X,\ \ (3)\ Y = 3.0{-}1.8X,\ \ (4)\ Y = 1.0{+}4.3X. \tag{40}$$

We can observe the data structures for Data A, B, and C, in Fig. 4a, b and c, respectively. The regression lines in each plot are the recovered linear lines obtained by the k-regressions

Table 1 Parameter setup for the Data A, B, and C

	Data A			Data B			Data C			
Cluster	1	2	3	1	2	3	1	2	3	4
n	100	100	100	100	100	100	60	60	60	60
β_0	1.0	45.0	45.0	142.0	33.0	−73.0	2.0	1.0	3.0	1.0
β_1	1.3	1.8	−2.5	5.0	−3.0	0.6	0.8	2.3	−1.8	4.3
μ_x	4.0	0.0	8.0	−28.0	12.0	−10.0	4.0	3.0	4.0	3.0
σ_x	12.0	9.6	9.0	10.0	17.0	20.0	4.0	3.0	4.0	3.0
λ_x	1.5	1.3	1.2	1.0	0.9	1.0	10.0	12.0	10.0	12.0
μ_ϵ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
σ_ϵ	5.0	4.0	3.0	6.0	9.0	8.0	1.0	2.0	1.0	4.0

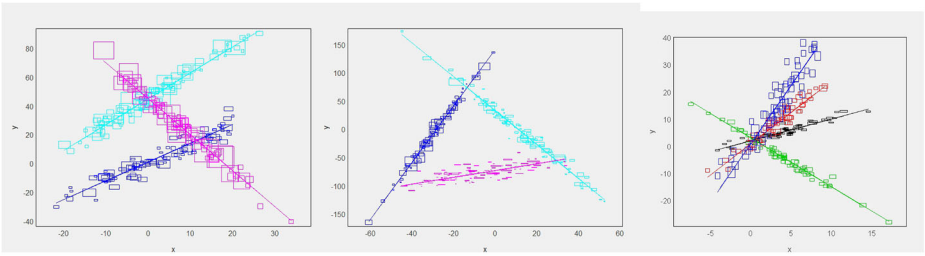

Fig. 4 Data structure for the Data A (**a**), B (**b**), and C (**c**)

algorithm. We can see that different clusters overlap with each other for all three data sets. Especially, for Data C, a large proportion of the four clusters is overlapping. In addition, for a particular data set, each cluster is clustered around a linear regression line.

4.3.2 Application of the *k*-Regressions Algorithm

Consider first Data A. For this particular data set, we generate a random sample that follows its structure as described in Table 1. Then, given the correct number of clusters $K = 3$, we use the k-regressions algorithm to recover the data structure. We try a number of random initial partitions. Based on these different initial partitions, the clustering result with smallest sum of squared residuals of Eq. 22 is set to be the correct convergence for these samples. For each simulation, we tried 50 different random initial partitions to recover its structure. This whole process is one replication. Then, we implement 100 such replications using different seeds to investigate the overall performance of the k-regressions algorithm.

For each distance, and each cluster, the mean and standard deviation (std) of the estimated parameter values from these 100 replications are displayed in Table 2. For example, for the center distance, the mean estimated coefficients of β_0 and β_1 for the cluster 1 are 0.93 and 1.32, respectively; the differences between the estimated values and the true values are small relative to the true values. The standard deviation of the estimated β_0 and β_1 for cluster 1 are 0.17 and 0.01, respectively. Small standard deviations of the coefficients indicate we have stable clustering results. Similar arguments pertain for the other two clusters and for the other distances, though it is noted that the center distances give considerably better fits when comparing the SSR residual values.

For each replication, we tried 50 different initial partitions when applying the k-regressions algorithm to a particular simulation. The number of good initial partitions out of 50, n^*, gives an idea about how difficult it is for the algorithm to converge to the correct cluster by a random initial partition. Out of the 100 replications, we can calculate the mean and standard deviation of n^*, which is shown in the row "n^* out of 50" in Table 2 for the three distances. The SSR is also calculated for each replication. The mean and standard deviation of the SSR out of the 100 replications are presented in the row "SSR".

Table 3 presents the corresponding clustering results for Data B with 100 replications. Table 3 can be interpreted in a similar way as Table 2 for Data A. Given $K = 3$, for all the three distances, the differences between the true coefficients and the mean estimated regression coefficients are all small relative to the coefficient scales. Again, small standard deviations for all the estimated coefficients imply stable clustering results.

Table 2 k-regressions clustering results for Data A

		True	Center		City-block		Hausdorff	
	Parameter	Value	Mean	std	Mean	std	Mean	std
Cluster 1	β_0	1.00	0.93	0.17	4.01	6.16	1.75	3.41
	β_1	1.30	1.32	0.01	1.25	0.55	1.33	0.36
Cluster 2	β_0	45.00	45.00	0.13	44.96	3.42	44.59	2.91
	β_1	1.80	1.82	0.01	1.81	0.33	1.76	0.53
Cluster 3	β_0	45.00	44.65	0.18	44.49	11.03	44.06	3.76
	β_1	−2.50	−2.46	0.01	−2.18	0.77	−2.42	0.26
n^* out of 50	-	-	30.75	8.74	24.73	11.16	30.03	9.55
SSR	-	-	325.69	20.23	2614.00	94.97	2675.46	56.05

The clustering results for Data C are presented in the Table 4. The interpretation of Table 4 for Data C follows in a similar manner as for Table 2 for Data A and Table 3 for Data B. Note that for Data C, a large proportion of the four clusters overlaps, which makes it more difficult to converge to the correct clusters for the k-regressions algorithm. For each replication, we tried 200 different initial partitions. The differences between the true coefficients and the mean estimated coefficients are small relative to the scales of the coefficients. The standard deviations of the coefficients are small for all the estimated coefficients and all the three distances. However, the intercept estimates for clusters 2, 3, and 4 are not as accurate as for cluster 1. This is not surprising given that cluster 1 is more separated from the other three clusters.

Now, let us use the same three data structures, Data A, B, and C, to investigate the performance of determining the optimal number of clusters by the elbow method, and the adjusted R^2. For a particular data set, we generate a random sample based on its parameter setup and implement the k-regressions algorithm for $K = 1, \ldots, 10$. For each K, we try a number of different initial partitions and select the results with smallest SSR as the correct clustering results. The $R^{2^{(K)}}$ from Eq. 28 and $\bar{R}^{2^{(K)}}$ from Eq. 32 are calculated for each of $K = 1, \ldots, 10$. The elbow plot is plotted as K versus $1 - R^{2^{(K)}}$. We also plot the K versus $\bar{R}^{2^{(K)}}$ where the maximum $\bar{R}^{2^{(K)}}$ determines the optimal number of clusters. This

Table 3 k-regressions clustering results for Data B

		True	Center		City-block		Hausdorff	
	Parameter	Value	Mean	std	Mean	std	Mean	std
Cluster 1	β_0	142.00	141.30	2.09	140.77	1.89	141.75	2.01
	β_1	5.00	4.97	0.07	4.95	0.07	4.99	0.07
Cluster 2	β_0	33.00	33.06	1.25	33.39	1.30	33.42	1.28
	β_1	−3.00	−2.99	0.06	−3.00	0.06	−2.99	0.06
Cluster 3	β_0	−73.00	−72.93	0.98	−72.83	1.11	−72.64	1.14
	β_1	0.60	0.60	0.05	0.60	0.05	0.60	0.06
n^* out of 50	-	-	26.55	14.51	33.42	15.70	34.51	14.74
SSR	-	-	1757.41	85.17	4127.63	177.30	2689.33	90.65

Table 4 k-regressions clustering results for Data C

		True	Center		City-block		Hausdorff	
	Parameter	Value	Mean	std	Mean	std	Mean	std
Cluster 1	β_0	2.00	2.06	0.38	3.55	1.76	3.86	2.98
	β_1	0.80	0.81	0.05	0.68	0.17	0.73	0.18
Cluster 2	β_0	1.00	1.32	1.31	3.06	2.74	5.20	4.04
	β_1	2.30	2.36	0.22	2.28	0.50	1.97	0.73
Cluster 3	β_0	3.00	2.90	0.32	3.12	0.34	3.02	0.39
	β_1	−1.80	−1.78	0.04	−1.81	0.05	−1.80	0.05
Cluster 4	β_0	1.00	2.13	1.87	4.29	2.67	4.24	2.82
	β_1	4.30	4.27	0.35	4.04	0.46	4.05	0.48
n^* out of 200	-	-	43.95	45.77	14.73	28.70	24.77	33.35
SSR	-	-	296.25	20.52	777.90	44.84	521.13	27.75

whole process is for one replication, and we implement a total of 20 replications to test the performance of the elbow method and the adjusted R^2.

Figure 5a, b, and c are the elbow plots for $(1 - R^{2(K)})$ against K for Data A, B, and C, respectively, where the grey lines are the elbow plots for the 20 replications, and the blue line is the average $R^{2^{(K)}}$ over the 20 replications. We can see that the elbow plots identify the correct optimal number of clusters for all three data sets, $K = 3$ for Data A and B, and $K = 4$ for C. It is relatively difficult to determine the optimal number of clusters for Data C due to the overlapping of clusters; however, the elbow plots correctly determined the number of clusters for all 20 replications nevertheless. Figure 5d, e, and f show the plots of

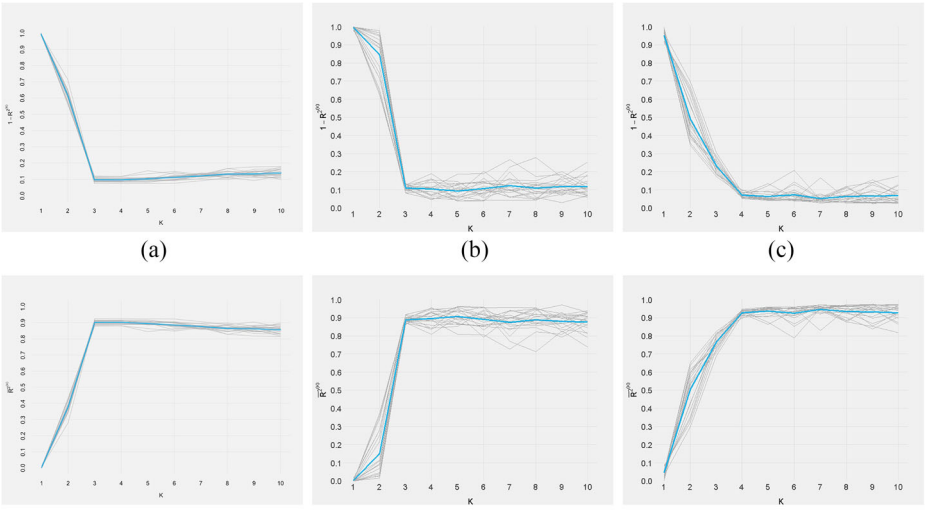

Fig. 5 Elbow plots by weighted R^2 and adjusted R^2 for Data A (**a**) and (**d**), Data B (**b**) and (**e**), and Data C (**c**) and (**f**)

$\bar{R}^{2(K)}$ for Data A, B, and C, respectively. For each of the three data structures, the optimal number of clusters determined by the largest $\bar{R}^{2(K)}$ is mostly larger than the true number of clusters for the 20 replications.

Generally, the elbow method is a stable and reliable method to determine the optimal number of clusters. There could be cases where $R^{2(K)}$ decreases gradually and consistently so that an elbow point is hard to find. Usually such scenarios indicate that there does not exist an optimal number of clusters to separate the data well and subjective judgment needs to be involved for a decision. Fixing a reasonable cutoff for $R^{2(K)}$ is a realistic option in practice. The $\bar{R}^{2(K)}$, adjusted R^2, usually overestimates the optimal number of clusters and so is not a good method to determine the optimal number of clusters.

4.3.3 $p > 1$

Consider now three simulated data sets D, E, F, with $p = 3, 5, 5$, respectively. The parameter setups are given in Table 5; tables and figures are shown in the Supplementary Materials S1. The sample size for each cluster in Data D is 200, while the sample size for each cluster in Data E and in Data F is 300 observations. As before, there are 100 replications for each case. The estimated parameter values for each of the three distances, are given in Tables 6, 7 and 8, for Data D, E, and F, respectively. The respective elbow plots are shown in Fig. 10 with a and d showing the weighted $1 - R^{2(K)}$ plot and the adjusted $\bar{R}^{2(K)}$ plot for Data D, Fig. 10b and e for Data E, and Fig. 10c and f for Data F, respectively. As for the $p = 1$ cases, these estimated results compare well with the true values; and the fits are good as determined by the SSR values and the elbow plots, again corroborating the merits and usefulness of the proposed k-regressions algorithm. Tables 6–8 also show the time (in secs) to run the 100 replications. From these, we see there is very little difference between these times as p increases; rather, if anything, any time differences are due to different distance measures used, though the standard deviations are large relative to the respective sample means.

5 Real Data Application

The methodology introduced herein is applied to the faces data set of Leroy et al. (1996). The predictor variable is X = length between the outer corners of the eyes, i.e., eye-span, and the response variable is Y = length between the inner corners of the eyes, i.e., the length of the bridge of the nose. There are $n = 27$ interval-valued observations (see Table 9 in Supplementary Materials S2). The resulting $K = 3$ clusters, obtained from the k-regressions algorithm, are shown in Fig. 6. The maximum number of clusters was set at $K = 5$. We see from the elbow plot in Fig. 7 that the optimal number is the $K = 3$ of Fig. 6. The respective regression equations are as follows:

$$\text{Cluster 1 (black):} \quad Y = -87.52 + 0.91X + \epsilon, \tag{41}$$

$$\text{Cluster 2 (red):} \quad Y = -143.19 + 1.20X + \epsilon, \tag{42}$$

$$\text{Cluster 3 (green):} \quad Y = -80.17 + 0.89X + \epsilon. \tag{43}$$

For all three clusters, the length between the two outer corners of the eyes (X) is positively correlated with the length between the two inner corners of the eyes (Y). However, the relationship between the two lengths is a bit different for different clusters.

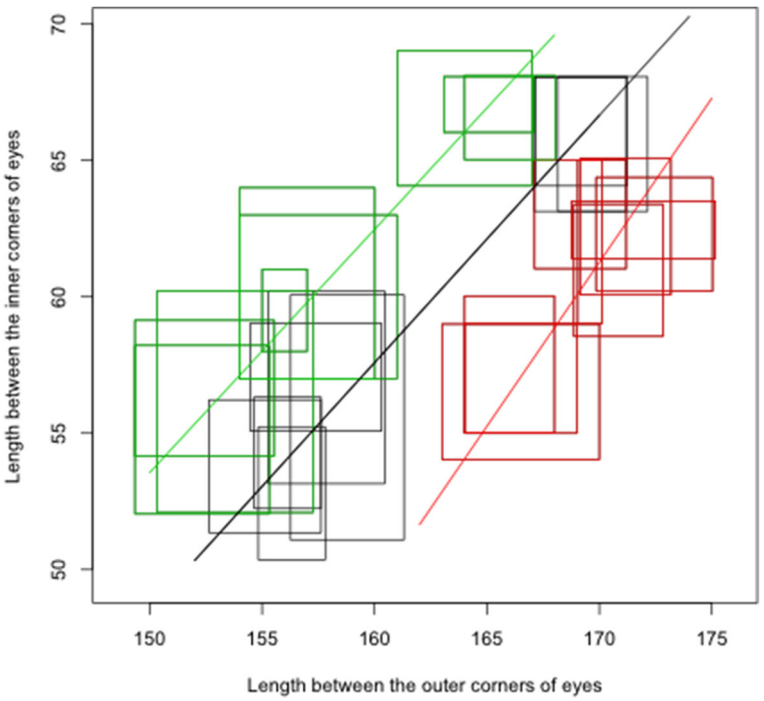

Fig. 6 Three k-regressions clusters for the faces data

The clusters consist of $C_1 = \{7, 8, 9, 19, 20, 21, 25, 26, 27\}$ faces, $C_2 = \{4, 5, 6, 10, 11, 12, 13, 14, 15\}$ faces, and $C_3 = \{1, 2, 3, 16, 17, 18, 22, 23, 24\}$ faces, respectively. As it so happens, these faces are in fact three observations from each of nine individuals. The algorithm (naturally unaware of this fact) correctly always puts the measurements for the same individual into the same cluster, thus enhancing the credibility of the algorithm.

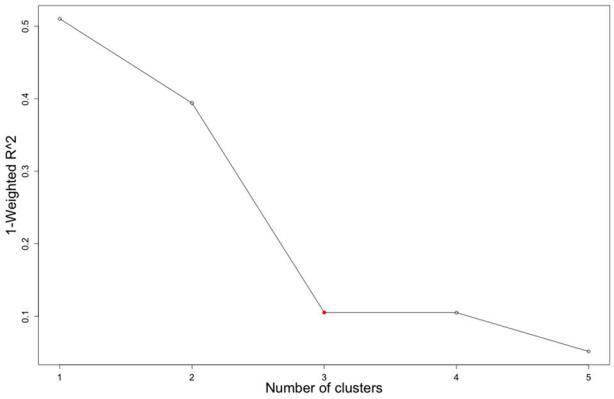

Fig. 7 Elbow plots faces data set clustering

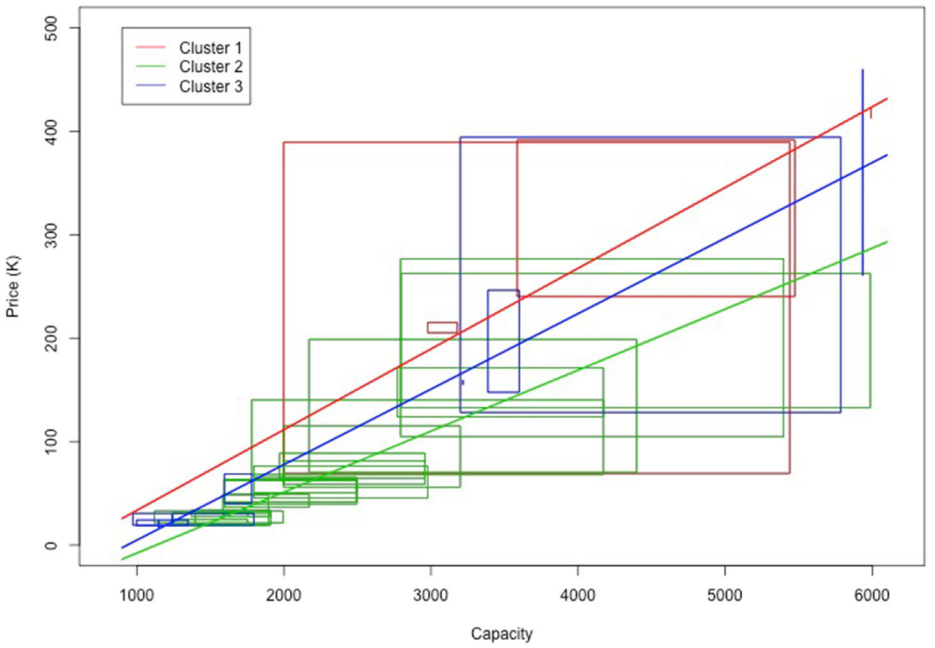

Fig. 8 Three k-regressions clusters for the cars data

A second data set is the cars data set discussed in de Carvalho et al. 2010. There are $n = 33$ cars. The k-regressions algorithm is applied where the response variable is $Y =$ price and the regression variable is $X =$ engine capacity (data are in Table 10 in S2). The k-regressions algorithm gave the $K = 3$ clusters as shown in Fig. 8. The respective regression

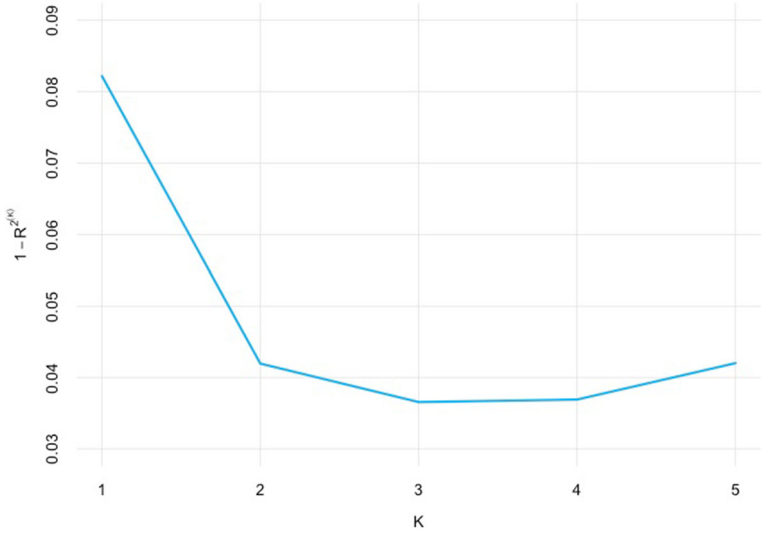

Fig. 9 Elbow plots cars data set clustering

equations (where the data are standardized) are as follows:

$$\text{Cluster 1 (red):} \quad Y = -44.53 + 78.05X + \epsilon, \tag{44}$$

$$\text{Cluster 2 (green):} \quad Y = -66.78 + 58.98X + \epsilon, \tag{45}$$

$$\text{Cluster 3 (blue):} \quad Y = -68.24 + 73.00X + \epsilon. \tag{46}$$

The elbow plot is shown in Fig. 9 suggesting that the optimal $K = 3$. The clusters consist of $C_1 = \{11, 15, 16, 22\}$ cars, $C_2 = \{1, 2, 3, 6, 7, 8, 9, 10, 12, 13, 14, 18, 20, 21, 26, 29, 30, 31, 32, 33\}$ cars, and $C_3 = \{3, 4, 16, 18, 22, 23, 24, 26, 27\}$ cars.

These cars data were also analysed by de Carvalho et al. 2010 using the same variables. Importantly, they also obtained three clusters. However, it is difficult to make any further comparison, as they apply separate classical regressions to each of the interval centers and half-ranges point values and so (apart from the deficiencies discussed in Section 3.1) it is not possible to produce cluster-regression fits along the lines of Eqs. 44–46 above.

6 Conclusion

The use of a regression-based approach to partition a data set into its relevant clusters, as originally introduced by Charles (1977), exists for classical data. For interval data, we have introduced a k-regressions algorithm to partition a set of interval-valued observations into its underlying clusters. The algorithm was tested on a wide variety of simulated data sets and two real application sets of interval data. It worked well. The new k-regressions algorithm demonstrably excelled when compared with the well-known k-means algorithm methodology for partitioning.

Supplementary Information The online version contains supplementary material available at (10.1007/s00357-021-09394-5).

References

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Batagelj, V., Kejžar, N., & Korenjak-Černe, S. (2015). Clustering of modal valued symbolic data. Machine Learnin. arXiv:1507.06683.

Bertrand, P., & Goupil, F. (2000). Descriptive statistics for symbolic data. In H.-H. Bock, & E. Diday (Eds.) *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data* (pp. 103–124). Berlin: Springer.

Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining*, *4*, 149–156.

Billard, L. (2014). The past's present is now. What will the present's future bring? In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.) *Past, present, and future of statistical science* (pp. 323–334). New York: Chapman and Hall.

Billard, L., & Diday, E. (2000). Regression analysis for interval-valued data. In H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, & M. Schader (Eds.) *Data analysis, classification, and related methods* (pp. 369–374). Berlin: Springer.

Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal American Statistical Association*, *98*, 470–487.

Billard, L., & Diday, E. (2006). *Symbolic data analysis: conceptual statistics and data mining*. Chichester: Wiley.

Bock, H.-H. (2007). Clustering methods: A history of k-means algorithms. In P. Brito, P. Bertrand, G. Cucumel, & F. de Carvalho (Eds.) *Selected contributions in data analysis and classification* (pp. 161–172). Berlin: Springer.

Bock, H.-H. (2008). Origins and extensions of the k-means algorithm in cluster analysis. *Journal Électronique d'Histoire des Probabilités et Statistics*, *4*, 1–18.

Bock, H.-H., & Diday, E. (2000). *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*. Berlin: Springer.

Bougeard, S., Abdi, H., Saporta, G., & Niang, N. (2018). Clusterwise analysis for multiblock component methods. *Advances in Data and Analysis of Classification*, *12*, 285–313.

Bougeard, S., Cariou, V., Saporta, G., & Niang, N. (2017). Prediction for regularized clusterwise multiblock regression. *Applied Stochastic Models for Business and Industry*, *34*, 852–867.

Brusco, M. J., Cradit, J. D., Steinley, D., & Fox, G.L. (2008). Cautionary remarks on the use of clusterwise regression. *Multivariate Behavioral Research*, *43*, 29–49.

Charles, C. (1977). *Regression typologique et reconnaissance des formes thèse de 3ème cycle*. Université de, Paris, Dauphine.

Chavent, M., Lechevallier, Y., Jajuga, K., Sokolowski, A., & Bock, H.-H. (2002). Dynamical clustering of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In *Classification, clustering, and data analysis* (pp. 53–60). Berlin: Springer.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society A*, *134*, 321–367.

de Carvalho, F. A. T., Lima Neto, E. A., & Tenorio, C.P. (2004a). A new method to fit a linear regression model for interval-valued data. In *Lecture notes in computer science, KI2004 advances in artificial intelligence* (pp. 295–306). Springer.

de Carvalho, F. A. T., de Souza, R. M. C. R., & Silva, F.C.D. (2004b). A clustering method for symbolic interval-type data using adaptive Chebyshev distances. In A. L. C. Bazzan, & S. Labidi (Eds.) *LNAI 3171* (pp. 266-275). Berlin: Springer.

de Carvalho, F. A. T., Brito, M. P., & Bock, H.-H. (2006). Dynamic clustering for interval data based on l_2 distance. *Computational Statistics*, *21*, 231–250.

de Carvalho, F. A. T., & Lechevallier, Y. (2009). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, *42*, 1223–1236.

de Carvalho, F. A. T., Saporta, G., & Queiroz, D.N. (2010). A clusterwise center and range regression model for interval-valued data. In Y. Lechevallier, & G. Saporta (Eds.) *Proceedings in computational statistics COMPSTAT 2010* (pp. 461–468). Berlin: Springer.

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, *5*, 249–282.

de Souza, R. M. C. R., & de Carvalho, F. A. T. (2004). Clustering of interval data based on city-block distances. *Pattern Recognition Letters*, *25*, 353–365.

de Souza, R. M. C. R., de Carvalho, F. A. T., Tenóio, C. P., & Lechevallier, Y. (2004). Dynamic cluster methods for interval data based on Mahalanobis distances. In D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul (Eds.) *Classification, clustering, and data analysis* (pp. 251–360). Berlin: Springer.

Diday, E. (1971a). Une nouvelle méthode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. *Revue de Statistique Appliquée*, *2*, 19–33.

Diday, E. (1971b). La méthode des nuées dynamiques. *Revue de Statistique Appliquée*, *19*, 19–34.

Diday, E. (1987). Introduction à l'approche symbolique en analyse des données. *Premier Jouneles Symbolique-Numerique*, CEREMADE, Universite Paris - Dauphine, 21–56.

Diday, E. (2016). Thinking by classes in data science: The symbolic data analysis paradigm. *WIRES Computational Statistics*, *8*, 172–205.

Diday, E., & Noirhomme-Fraiture, M. (2008). *Symbolic data analysis and the SODAS software*. Chichester: Wiley.

Diday, E., & Simon, J. C. (1976). Clustering analysis. In K. S. Fu (Ed.) *Digital pattern recognition* (pp. 47–94). Berlin: Springer.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. New York: Wiley.

Hausdorff, F. (1937). *Set theory (translated into English by J. R. Aumann 1957)*. New York: Chelsey.

Irpino, A., Verde, R., & Lechevallier, Y. (2006). Dynamic clustering of histograms using Wasserstein metric. In A. Rizzi, & M. Vichi (Eds.) *COMPSTAT 2006* (pp. 869–876). Berlin: Physica-Verlag.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*, 651–666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*, 263–323.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*, 6th edn. New Jersey: Prentice-Hall.

Korenjak-Černe, S., Batagelj, V., & Pavešić, B. J. (2011). Clustering large data sets described with discrete distributions and its application on TIMSS data set. *Statistical Analysis and Data Mining*, *4*, 199–215.

Košmelj, K., & Billard, L. (2012). Mallows'l^2 distance in some multivariate methods and its application to histogram-type data. *Metodološki Zvezki*, *9*, 107–118.

Leroy, B., Chouakria, A., Herlin, I., & Diday, E. (1996). Approche géométrique et classification pour la reconnaissance de visage. *Reconnaissance des Forms et Intelligence Artificelle*, INRIA and IRISA and CNRS, France, 548–557.

Lima Neto, E. A., & de Carvalho, F. A. T. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis*, *52*, 1500–1515.

Lima Neto, E. A., de Carvalho, F. A. T., & Freire, E.S. (2005). Applying constrained linear aggression models to predict interval-valued data. In U. Furbach (Ed.) *Lecture notes in computer science, KI: advances in artificial intelligence* (pp. 92–106). Brelin: Springer.

Lima Neto, E. A., de Carvalho, F. A. T., & Tenorio, C.P. (2004). Univariate and multivariate linear regression methods to predict interval-valued features. In *Lecture notes in computer science, AI 2004, advances in artificial intelligence* (pp. 526–537). Berlin: Springer.

Liu, F. (2016). *Cluster analysis for symbolic interval data using linear regression method*. Doctoral Dissertation, University of Georgia.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. LeCam, & J. Neyman (Eds.) *Proceedings of the 5th berkeley symposium on mathematical statistics and probability*, (Vol. 1 pp. 281–299). Berkeley: University of California Press.

Noirhomme-Fraiture, M., & Brito, M. P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, *4*, 157–170.

Qian, G., & Wu, Y. (2011). Estimation and selection in regression clustering. *European Journal of Pure and Applied Mathematics*, *4*, 455–466.

Rao, C. R., Wu, Y., & Shao, Q. (2007). An M-estimation-based procedure for determining the number of regression models in regression clustering. *Journal of Applied Mathematics and Decision Sciences*, Article ID 37475.

Shao, Q., & Wu, Y. (2005). A consistent procedure for determining the number of clusters in regression clustering. *Journal of Statistical Planning and Inference*, *135*, 461–476.

Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing*, *22*, 367–373.

Späth, H. (1981). Correction to algorithm 39: clusterwise linear regression. *Computing*, *26*, 275.

Späth, H. (1982). A fast algorithm for clusterwise linear regression. *Computing*, *29*, 175–181.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B*, *63*, 411–423.

Verde, R., & Irpino, A. (2007). Dynamic clustering of histogram data: Using the right metric. In P. Brito, P. Bertrand, G. Cucumel, & F. de Carvalho (Eds.) *Selected contributions in data analysis and classification* (pp. 123–134). Berlin: Springer.

Wedel, M., & Kistemaker, C. (1989). Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing*, *6*, 45–59.

Xu, W. (2010). *Symbolic data analysis: interval-valued data regression*. Doctoral Dissertation, University of Georgia.

Zhang, B. (2003). Regression clustering. In X. Wu, A. Tuzhilin, & J. Shavlik (Eds.) *Proceedings third IEEE international conference on data mining* (pp. 451–458). California: IEEE Computer Society Publishers.