Off the beaten track: a new linear model for interval data

Sónia Dias[a,*], Paula Brito[b]

[a] *Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal*
[b] *Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Portugal*

Abstract

We propose a new linear regression model for interval-valued variables. The model uses quantile functions to represent the intervals, thereby considering the distributions within them. In this paper we study the special case where the Uniform distribution is assumed in each observed interval, and we analyse the extension to the Symmetric Triangular distribution. The parameters of the model are obtained solving a constrained quadratic optimization problem that uses the Mallows distance between quantile functions. As in the classical case, a goodness-of-fit measure is deduced. Two applications on up-to-date fields are presented: one predicting duration of unemployment and the other allowing forecasting burned area by forest fires.

Keywords: interval data, linear regression, Symbolic Data Analysis, quantile functions

1. Introduction

The extensive and complex data that emerged in the last decades made it necessary to extend and generalize the classical concept of data sets. Data tables where the cells contain a single quantitative or categorical value were no longer sufficient. More complex data tables were needed, with cells that include more accurate and complete information, e.g. expressing the variability or imprecision

*Corresponding author
Email addresses: sdias@estg.ipvc.pt (Sónia Dias), mpbrito@fep.up.pt (Paula Brito)

of the records of each observed unit. Data with variability occur when each unit represents a specific group/class, or when values express characteristics that float along a period of time. Imprecise data occur when an interval associated with each unit under analysis represents the uncertain value of the record. This paper comes within the framework of *Symbolic Data Analysis* (SDA) (Billard and Diday, 2006; Noirhomme-Fraiture and Brito, 2011; Brito, 2014), which is concerned with data presenting variability. This variability may emerge due to the aggregation of single observations. In particular, we consider symbolic variables whose realizations are intervals, i.e., *interval-valued variables* (Bock and Diday, 2000; Billard and Diday, 2006).

Data at aggregated level are also used in other frameworks, as for example Granular Computing (see Pedrycz, 2014). Information granules are groups of individual observations capturing the semantics of the abstract entities of interest. When the data are numerical, granules may be defined by cartesian products of intervals. Information granules may be expressed and treated considering fuzzy, rough, interval and probabilistic models. A comprehensive treatise on the subject may be found in Pedrycz (2013).

The linear regression models for interval-valued variables previously proposed are very different from the one presented in this work. Most linear regression models developed for interval-valued variables in the context of SDA are descriptive (see Billard and Diday, 2000, 2002, 2006; Lima Neto and De Carvalho, 2008, 2010; Giordani, 2014); nevertheless, some papers recently published propose probabilistic models and inference studies (see Ahn et al., 2012; Lima Neto et al., 2011; Brito and Duarte Silva, 2012). The noteworthy descriptive models allow predicting a response variable from p explicative variables. However those models do not treat the intervals as such, rather they require the adjustment of classical linear regression models for their lower and upper bounds or for their centers and half ranges. In other words, these models are based on differences between real values and do not quantify the closeness between intervals. Therefore, the elements estimated by the models may fail to build an interval; to solve this problem the *Constrained Center and Range Method (CCRM)* proposed by

Lima Neto and De Carvalho (2010) imposes non-negative constraints in the linear regression between the half ranges of the intervals. More recently, Giordani (2014) proposed a Lasso approach, that as Lima Neto and De Carvalho (2008, 2010) considers two regression models, one for the centers and another for the half ranges, but whose parameters are related.

Linear regression models have also been proposed for intervals representing imprecise data. Aznar and Guijarro (2007) proposed a programming optimization technique that allows predicting real values and estimating regression parameters from imprecise information, represented by intervals. A review on some linear regression analysis for imprecise interval-valued data may be found in Blanco-Fernández et al. (2013).

The linear regression model proposed here is a descriptive method that comes within the SDA framework. Borrowing the idea of using quantile functions to represent variables expressing variability from Dias and Brito (2015), we propose a model that allows considering intervals as such, i.e., we do not fit separate linear models to the corresponding bounds or midpoints and ranges. The innovations of the proposed model are the following: 1) the model considers the distribution within the intervals; in this paper the Uniform distribution is in general assumed, and the extension to the Symmetric Triangular distribution is investigated, other distributions may also be considered; 2) the intervals are represented by quantile functions; 3) the model allows predicting a response variable from p explicative variables and the predicted range of values always constitutes an interval; 4) the linear relations between the centers and half ranges, induced by the model, are jointly obtained, but although related, these relations are different; 5) a goodness-of-fit measure is derived from the model.

In Section 2, we introduce the representation of intervals by quantile functions and present the linear regression model for interval-valued variables. Section 3 reports a simulation study and discusses its results. In Section 4, real applications are presented. Finally, Section 5 concludes the paper, pointing out directions for future research.

2. Regression Model with interval-valued variables

In this section a linear regression model for interval-valued variables and the respective goodness-of-fit measure are proposed assuming, in general, the Uniform distribution within the intervals. The model is however quite flexible: it reduces to the classical linear regression when it is applied to degenerate intervals (classical numerical variables) and it may be adapted to other distributions within the intervals.

2.1. Representation of the intervals by quantile functions

According to Bock and Diday (2000) and Billard and Diday (2006) interval-valued variables are formally defined as follows.

Definition 2.1. *Y is an interval-valued variable when to each unit $j \in \{1, \ldots, n\}$ of the set under study corresponds an interval $Y(j)$ of real numbers. $Y(j)$ may be represented by the interval $I_{Y(j)} = \left[\underline{I}_{Y(j)}, \overline{I}_{Y(j)}\right]$; alternatively, the interval $Y(j)$ may be represented by its center $c_{Y(j)} = \frac{\overline{I}_{Y(j)} + \underline{I}_{Y(j)}}{2}$ and half-range $r_{Y(j)} = \frac{\overline{I}_{Y(j)} - \underline{I}_{Y(j)}}{2}$, then $I_{Y(j)} = \left[c_{Y(j)} - r_{Y(j)}; c_{Y(j)} + r_{Y(j)}\right]$.*

Definition 2.2. *The symbolic mean of an interval-valued variable Y is defined in Billard and Diday (2006) as:*

$$\overline{Y} = \frac{1}{n} \sum_{j=1}^{n} c_{Y(j)}.$$

Irpino and Verde (2006) proposed using quantile functions to represent empirical distributions. In this work we particularize this representation to intervals. Notice that a similar representation has already been considered by Bertoluzza et al. (1995) and named "parametrization of the interval". Assuming an Uniform distribution in each interval $Y(j)$, we may represent $Y(j)$ by the respective empirical quantile function as follows:

$$\Psi^{-1}_{Y(j)}(t) = \underline{I}_{Y(j)} + \left(\overline{I}_{Y(j)} - \underline{I}_{Y(j)}\right) t, \qquad 0 \leq t \leq 1.$$

or using the center $c_{Y(j)}$ and half-range $r_{Y(j)}$ as

$$\Psi^{-1}_{Y(j)}(t) = c_{Y(j)} + r_{Y(j)}(2t-1), \qquad 0 \leq t \leq 1.$$

As empirical quantile functions are the inverse cumulative distribution functions, which in the particular case of the intervals are continuous linear functions with domain $[0,1]$, we shall use the usual arithmetic operations with functions. However, when we use arithmetic operations with quantile functions, problems may arise. Since for all intervals $\underline{I}_Y \leq \overline{I}_Y$, the quantile function that represents an interval is always a non-decreasing function. The addition of quantile functions is a non-decreasing function, but when we multiply a quantile function by a negative real number we obtain a function that is not non-decreasing. Consider an interval $I_Y = [c_Y - r_Y, c_Y + r_Y]$ and let $-I_Y = [-c_Y - r_Y, -c_Y + r_Y]$ be the respective symmetric interval; if $\Psi^{-1}_Y(t) = c_Y + r_Y(2t-1), \quad 0 \leq t \leq 1$, is the quantile function that represents I_Y, the quantile function that represents $-I_Y$ is $-\Psi^{-1}_Y(1-t) = -c_Y + r_Y(2t-1), \quad 0 \leq t \leq 1$ (which is non-decreasing) - and not $-\Psi^{-1}_Y(t)$. Some properties met by the usual symmetric elements are not met when these elements are ranges of values. The addition of interval I_Y with $-I_Y$ is not the null interval, so that the difference between ranges of values does not provide information on how dissimilar the intervals are. The difference between two equal intervals is an interval with symbolic mean (as defined by Billard and Diday (2006)) zero i.e., an interval with center zero and symmetric bounds. For more details about the behavior of quantile functions, see Dias (2014).

Since we need to measure the similarity between predicted and observed intervals, it is necessary to select an adequate distance. In this paper the Mallows distance will be the considered to evaluate the similarity between intervals. In recent literature (Arroyo and Maté, 2009; Irpino and Verde, 2015), the Mallows distance is considered an adequate measure to evaluate the similarity between distributions. Arroyo and Maté (2009) study several di,stances and conclude that, in addition to the interesting properties for error measurement (positive definiteness, symmetry, and triangle inequality) the Mallows distance has intuitive interpretations related to the Earth Mover's Distance (EMD), and it is the

one that better adjusts to the concept of distance as assessed by the human eye. This distance has also been considered in cluster analysis for histogram data, proposed by Irpino and Verde (2006). In a classification context, Hofer (2014) uses the EMD distance between distributions, also making a link with the Mallows distance. In spite of the fact that the Mallows distance has mainly been applied to distributions, intervals being considered a special case, the application of this distance to intervals is not completely new. The Mallows distance is a particular case of the Bertoluzza distance, used in the literature to measure the distance between two intervals (Bertoluzza et al., 1995); a generalization of this distance is also used in linear regression models for interval-valued random sets (Blanco-Fernández et al., 2011; González-Rodríguez et al., 2007; Gil et al., 2002).

Definition 2.3. *Given two quantile functions* $\Psi^{-1}_{X(j)}(t)$ *and* $\Psi^{-1}_{Y(j)}(t)$ *representing the values of interval-valued variables* X *and* Y *for an observation* j*, the square of the Mallows distance is defined as follows (Mallows, 1972):*

$$D^2_M(\Psi^{-1}_{X(j)}, \Psi^{-1}_{Y(j)}) = \int_0^1 (\Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t))^2 dt.$$

Irpino and Verde (2006) have simplified the expression for the squared Mallows distance between two distributions represented by the quantile functions associated with the corresponding histograms, assuming the Uniform distribution within the histograms' intervals. Particularizing for interval-valued variables, we may then rewrite the squared Mallows distance as the sum of the squared differences between the centers and one third of the square of the differences between the half-ranges. We notice that the weight of the difference between the centers is larger than the weight of the difference between the half-ranges.

Proposition 2.1. *Given two quantile functions* $\Psi^{-1}_{X(j)}(t)$ *and* $\Psi^{-1}_{Y(j)}(t)$ *representing the values of interval-valued variables* X *and* Y *for an observation* j*, and assuming uniformity within each interval, the squared Mallows distance may be expressed as:*

$$D_M^2(\Psi_{X(j)}^{-1}, \Psi_{Y(j)}^{-1}) = (c_{X(j)} - c_{Y(j)})^2 + \frac{1}{3}(r_{X(j)} - r_{Y(j)})^2$$

where $c_{X(j)}, c_{Y(j)}$ *are the centers and* $r_{X(j)}, r_{Y(j)}$ *are the half-ranges of the intervals* $X(j)$ *and* $Y(j)$*, respectively, with* $j \in \{1, 2, \ldots, n\}$ *.*

2.2. The Interval Distributional regression model

The proposed model, defined for p explicative variables, is the first linear regression model within the SDA framework that predicts intervals from other intervals without decomposing them in their bounds or centers and half ranges. Following the principle of a quantile function representation for data with variability (Dias and Brito, 2015), in this model the observations of the interval-valued variables are represented by quantile functions, assuming a specific distribution within the intervals. However, as mentioned above, when we multiply a quantile function by a negative number we obtain a function that is not a non-decreasing function and consequently is not a quantile function. As a consequence, the functional linear relation between interval data may not be just an adaptation of the classical model, since, if the linear relation model between interval-valued variables were the classical linear model then, if the values of the parameters were negative, the function predicted for $Y(j)$ might not be a quantile function ($\Psi_{\widehat{Y}(j)}^{-1}(t)$ might be a decreasing function). For the predicted element to be a quantile function it would be necessary to impose non-negativity constraints on the parameters of the model. However, these restrictions would always force a direct linear relation between the variables. Therefore, and although non-negative constraints on the parameters are necessary (the parameters of the model cannot be negative) it is obviously required to define a model that allows for a direct or inverse linear relation between the response variable Y and the independent variables X_i. For this reason, in our model both the quantile functions corresponding to the observations of the explicative interval-valued variables, $\Psi_{X_i(j)}^{-1}(t)$ and the quantile functions that represent the respective symmetric intervals, $-\Psi_{X_i(j)}^{-1}(1-t)$, are included. Consequently, the linear relation between the intervals is not necessarily direct, even

though positivity constraints are imposed on the parameters. For each unit it is hence possible to predict response quantile functions/intervals from other quantile functions/intervals.

The *Interval Distributional (ID) regression model* is defined as follows.

Definition 2.4. *Consider the interval-valued variables* $X_1; X_2; \ldots; X_p$. *The quantile functions that represent the range of values that these variables take for each unit* j *are denoted by* $\Psi^{-1}_{X_1(j)}(t)$, $\Psi^{-1}_{X_2(j)}(t), \ldots,$ $\Psi^{-1}_{X_p(j)}(t)$ *and the quantile functions that represent the respective symmetric intervals are denoted* $-\Psi^{-1}_{X_1(j)}(1-t), -\Psi^{-1}_{X_2(j)}(1-t), \ldots,$ $-\Psi^{-1}_{X_p(j)}(1-t)$, *with* $t \in [0,1]$. *For the response variable* Y, *each quantile function* $\Psi^{-1}_{Y(j)}$ *may be expressed as* $\Psi^{-1}_{Y(j)}(t) = \Psi^{-1}_{\widehat{Y}(j)}(t) + e_j(t)$ *where* $\Psi^{-1}_{\widehat{Y}(j)}(t)$ *is the predicted quantile function for unit* j, *obtained from*

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = v + \sum_{i=1}^{p} a_i \Psi^{-1}_{X_i(j)}(t) - \sum_{i=1}^{p} b_i \Psi^{-1}_{X_i(j)}(1-t) \tag{1}$$

with $t \in [0,1]$; $a_i, b_i \geq 0$, $i \in \{1, 2, \ldots, p\}$ *and* $v \in \mathbb{R}$.

When we assume uniformity within the observed intervals, and since $\Psi^{-1}_{X_i(j)}(t) = c_{X_i(j)} + (2t-1) r_{X_i(j)}$ and $-\Psi^{-1}_{X_i(j)}(1-t) = -c_{X_i(j)} + (2t-1) r_{X_i(j)}$, the predicted quantile function $\Psi^{-1}_{\widehat{Y}(j)}$ in expression (1) may be rewritten as follows:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \sum_{i=1}^{p} (a_i - b_i)\, c_{X_i(j)} + v + \sum_{i=1}^{p} (a_i + b_i)\, r_{X_i(j)}\, (2t-1) \tag{2}$$

with $t \in [0,1]$; $a_i, b_i \geq 0$, $i \in \{1, 2, \ldots, p\}$, and $v \in \mathbb{R}$.

The lower and upper bounds of each $I_{\widehat{Y}(j)}$ are obtained from expression (2) for $t = 0$ and $t = 1$, respectively, as follows:

$$\Psi^{-1}_{\widehat{Y}(j)}(0) = \sum_{i=1}^{p} a_i \left(c_{X_i(j)} - r_{X_i(j)}\right) - \sum_{i=1}^{p} b_i \left(c_{X_i(j)} + r_{X_i(j)}\right) + v$$

$$\Psi^{-1}_{\widehat{Y}(j)}(1) = \sum_{i=1}^{p} a_i \left(c_{X_i(j)} + r_{X_i(j)}\right) - \sum_{i=1}^{p} b_i \left(c_{X_i(j)} - r_{X_i(j)}\right) + v,$$

For each unit j, the predicted interval $I_{\widehat{Y}(j)}$ may then be obtained from

$$I_{\widehat{Y}(j)} = \left[\sum_{i=1}^{p} \left(a_i \underline{I}_{X_i(j)} - b_i \overline{I}_{X_i(j)} \right) + v, \sum_{i=1}^{p} \left(a_i \overline{I}_{X_i(j)} - b_i \underline{I}_{X_i(j)} \right) + v \right]. \tag{3}$$

The error, for each unit j, is a linear function, but not necessarily a quantile function, given by $e_j(t) = \Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t), \qquad t \in [0,1]$.

To be possible to define a linear regression model in the requested conditions, it was necessary to include in the model two parameters for each explicative variable - one associated with the quantile function that represents each interval $X_i(j)$ and the other associated with the quantile function that represents the respective symmetric interval. This solution increases the number of parameters to estimate, it is therefore important to bear in mind that the number of observations should be higher than the total number of parameters. Notice that a similar situation occurs for other models, as the MinMax, CRM, CCRM, where the bounds, or the centers and half ranges, of the intervals are separately estimated, and therefore two parameters must be estimated for each explicative variable.

For the *ID Model*, the center $c_{\widehat{Y}}(j)$ and the half range $r_{\widehat{Y}}(j)$ (or the bounds) predicted for the interval-valued variable Y may be described, respectively, by a classical linear relation for the centers $c_{X_i}(j)$ and by a classical linear relation for the half ranges $r_{X_i}(j)$ (or the bounds), of the explicative interval-valued variables. These linear relations, obtained from (3), are the following:

$$c_{\widehat{Y}}(j) = \sum_{i=1}^{p} (a_i - b_i)\, c_{X_i(j)} + v; \qquad r_{\widehat{Y}}(j) = \sum_{i=1}^{p} (a_i + b_i)\, r_{X_i(j)} \tag{4}$$

with $a_i, b_i \geq 0$, $i \in \{1, 2, \ldots, p\}$ and $v \in \mathbb{R}$.

Considering the expression that allows predicting the centers and considering that $v = \overline{Y} - \sum_{i=1}^{p} (a_i - b_i)\, \overline{X}_i$, (Dias and Brito, 2015), it is easily proven that the sum of the errors between the observed and predicted centers of the intervals is zero.

From these expressions we may observe that the parameters that define the linear regressions between the centers and between the half ranges are not the same but are related. In spite of the fact that this model is defined between intervals and that the relation may be direct or inverse, it always induces a direct linear relation between the half ranges. The direct or inverse relation

between the interval-valued variables is always in accordance with the linear relation between the centers. An interval-valued variable X_k is in direct linear relation with Y when $a_k > b_k$ and the linear relation is inverse if $a_k < b_k$. When we predict one interval-valued variable Y from only one interval-valued variable X, the following relations result from (4):

Proposition 2.2. *Consider the intervals predicted from an interval-valued variable X by the ID Model. From (4) we may conclude that:*

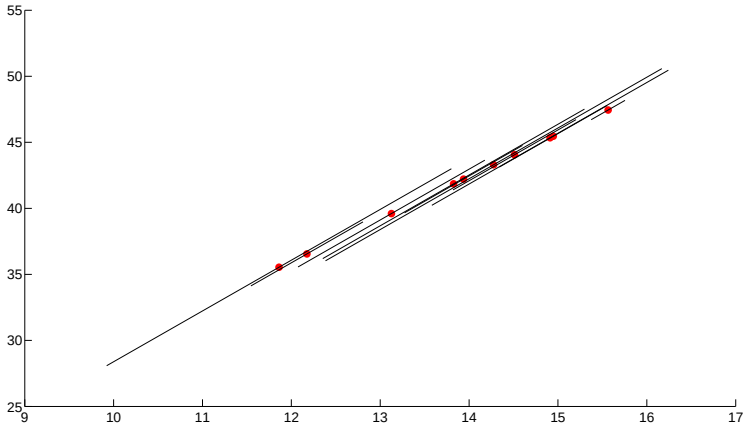
1. *The centers of the predicted intervals $\widehat{Y}(j)$ are obtained from a classical linear relation on the centers of the observed intervals of the variable X.*
2. *The ratio of the half ranges of the predicted intervals $\widehat{Y}(j)$ to the half ranges of $X(j)$, for $j \in \{1, \ldots, n\}$, is constant.*

Two interval-valued variables X and Y may be represented in a scatter plot. In this representation, each interval may be represented by a line joining the points $(\underline{I}_{X(j)}, \underline{I}_{Y(j)})$ and $(\overline{I}_{X(j)}, \overline{I}_{Y(j)})$, these lines are the diagonals of the rectangles that more usually are used to represent scatter plots of intervals (Billard and Diday, 2006). Figure 1 illustrates the results in Proposition 2.2 for a perfect linear relation: it induces a perfect linear regression between the centers of the intervals and the ratio (slope) of the ranges of the intervals is constant for all observations. In Figure 1(a) the linear relation is direct because $a > b$, in Figure 1(b) as $a < b$ the relation is inverse.

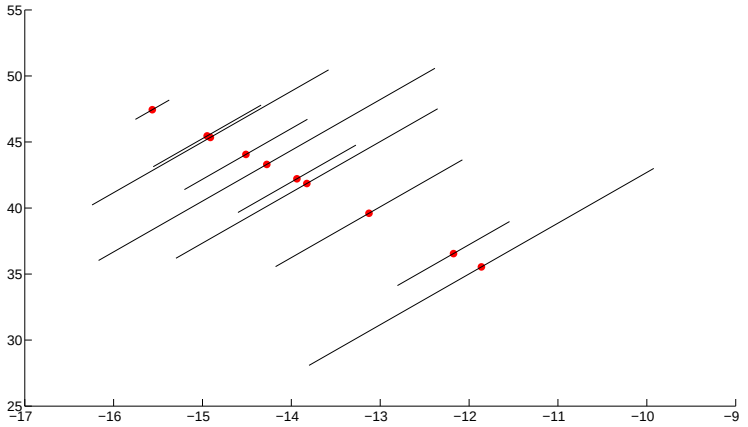
2.3. Parameters of the ID Model

The non-negative parameters of the *ID Model* in Definition 2.4, are determined solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns. The distance used to quantify the dissimilarity between the predicted and the observed quantile function is the Mallows Distance (Mallows, 1972) as in Proposition 2.1.

Consider the centers $c_{Y(j)}$ and half ranges $r_{Y(j)}$ of the observed intervals $Y(j)$ and the centers $c_{\widehat{Y}(j)}$ and half ranges $r_{\widehat{Y}(j)}$ of the predicted intervals $\widehat{Y}(j)$.



(a) $\Psi^{-1}_{\widehat{Y}(j)}(t) = -2.6 + 3.5\Psi^{-1}_{X(j)}(t) - 0.3\Psi^{-1}_{X(j)}(1-t)$.



(b) $\Psi^{-1}_{\widehat{Y}(j)}(t) = -2.6 + 0.3\Psi^{-1}_{X(j)}(t) - 3.5\Psi^{-1}_{X(j)}(1-t)$.

Figure 1: Scatter plots of the intervals and the respective centers.

The quadratic optimization problem to be solved to obtain the parameters of

the model is then:

$$\begin{array}{ll} \min & \sum_{j=1}^{n}\left[\left(c_{Y(j)}-\sum_{i=1}^{p}\left(a_i-b_i\right)c_{X_i(j)}-v\right)^2+\frac{1}{3}\left(r_{Y(j)}-\sum_{i=1}^{p}\left(a_i+b_i\right)r_{X_i(j)}\right)^2\right] \\ \text{subject to} & -a_i,-b_i\leq 0, \qquad i\in\{1,2,\ldots,p\}, \quad \text{with} \quad v\in\mathbb{R}. \end{array} \tag{5}$$

The optimization problem (5) may be written in matricial form as a classical constrained quadratic optimization problem or alternatively, as a constrained least squares problem.

Consider the $n-$dimensional vectors of the observed centers and half ranges of the response variable Y : $\mathbf{y^c}=\left(c_{Y(1)},\ldots,c_{Y(n)}\right)^T$, $\mathbf{y^r}=\left(r_{Y(1)},\ldots,r_{Y(n)}\right)^T$; and the vector of the parameters of the model with dimension $2p+1$, $\mathbf{b}=(a_1,b_1,\ldots,a_p,b_p,v)^T$. From the vectors $\mathbf{x^c(j)}=\left(c_{X_1(j)},-c_{X_1(j)},\ldots,c_{X_p(j)},-c_{X_p(j)},1\right)$, $\mathbf{x^r(j)}=\left(r_{X_1(j)},r_{X_1(j)},\ldots,r_{X_p(j)},r_{X_p(j)},0\right)$; we may build the following $n\times(2p+1)$ matrices:

$$\mathbf{X^c}=[\mathbf{x^c(1)}\quad\mathbf{x^c(2)}\quad\ldots\quad\mathbf{x^c(n)}]^T \quad\text{and}\quad \mathbf{X^r}=[\mathbf{x^r(1)}\quad\mathbf{x^r(2)}\quad\ldots\quad\mathbf{x^r(n)}]^T.$$

The minimization problem (5) may then be rewritten in matricial form as :

$$\begin{array}{ll} \min & \left\|\mathbf{y^c}-\mathbf{X^c b}\right\|^2+\frac{1}{3}\left\|\mathbf{y^r}-\mathbf{X^r b}\right\|^2 \\ \text{subject to} & -a_i,-b_i\leq 0, \qquad i\in\{1,2,\ldots,p\}, \quad \text{with} \quad v\in\mathbb{R}. \end{array} \tag{6}$$

As the parameters for the centers and half ranges may not be obtained independently, we may rewrite the optimization problem (5) as the following least squares problem:

$$\begin{array}{ll} \min & \left\|\begin{bmatrix}\mathbf{y^c}\\ \frac{1}{\sqrt{3}}\mathbf{y^r}\end{bmatrix}-\begin{bmatrix}\mathbf{X^c}\\ \frac{1}{\sqrt{3}}\mathbf{X^r}\end{bmatrix}\mathbf{b}\right\|^2=\left\|\mathbf{Y}-\mathbf{Xb}\right\|^2 \\ \text{subject to} & -a_i,-b_i\leq 0, i\in\{1,2,\ldots,p\}, \quad \text{with} \quad v\in\mathbb{R}. \end{array} \tag{7}$$

Several methods may be found in the literature to solve the constrained least squares problem (7) and therefore the constrained quadratic optimization problem (5). As the quadratic function to optimize is convex and the feasible region

as well, it may be ensured that the vectors that verify the *Kuhn Tucker conditions* (see Winston, 1994) are the vectors where the function reaches the smallest value i.e., are the optimal solutions. In cases where the objective function is strictly convex we may ensure that the optimal solution is unique.

In the case of a single linear regression we obtain explicit expressions for the parameters a, b and v of the model. These are obtained for different conditions imposed on the relations between the centers and half ranges of the intervals of the explicative and response variables, which emerge from the non-negativity of the parameters a and b. Four cases must be considered, as in Proposition 2.3 below.

Proposition 2.3. *Consider the minimization problem (5) with only one explicative variable X. When the function to minimize is strictly convex and the centers of all intervals of the explicative variable are not all the same, the optimal solution $\boldsymbol{b}^* = (a^*, b^*, v^*)$, i.e., the values of the parameters of the ID Model where the objective function reaches the minimum value, are as follows:*

I. $a^* = 0; \quad b^* = 0; \quad v^* = \overline{Y} \quad$ *if*

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} = \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) = 0;$$

II. $a^* = 0;\ b^* = \dfrac{\sum\limits_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} - \sum\limits_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)}{\sum\limits_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum\limits_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}; v^* = \overline{Y} + b^* \overline{X}$

if

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \geq \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)$$

and

$$\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 \leq -\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2;$$

III. $a^* = \dfrac{\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} + \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$; $b^* = 0$; $v^* = \overline{Y} - a^* \overline{X}$

if

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \geq -\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)$$

and

$$\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 \geq \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2;$$

IV. $a^* = \dfrac{\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2}{2 \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$;

$b^* = \dfrac{-\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2}{2 \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$;

$v^* = \overline{Y} - (a^* - b^*)\,\overline{X}$ *if*

$$-\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 \leq \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2$$

and

$$\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 \leq \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2.$$

Proof. The proof is given in *Supplementary Material*. □

As the *ID Model* uses both the quantile function $\Psi^{-1}_{X_i(j)}(t)$ representing the interval $X_i(j)$ and the quantile function $-\Psi^{-1}_{X_i(j)}(1-t)$ representing the respective symmetric interval, we analyze the behavior of the model when these functions are collinear. Proposition 2.4 below allows deducing the collinearity conditions.

Proposition 2.4. *The quantile functions* $\Psi^{-1}_{X_i(j)}(t) = c_{X_i(j)} + r_{X_i(j)}(2t-1)$ *and* $-\Psi^{-1}_{X(j)}(1-t) = -c_{X_i(j)} + r_{X_i(j)}(2t-1)$ *with* $t \in [0,1]$ *that represent the intervals* $I_{X_i(j)}$ *and* $-I_{X_i(j)}$*, respectively, for* $j \in \{1, \ldots, n\}$*, are collinear if at least one of the following conditions occurs:*

1. *the interval* $I_{X_i(j)}$ *has* $c_{X_i(j)} = 0$*, i.e., the interval is symmetric;*
2. $r_{X_i(j)} = 0$*, i.e., the interval is reduced to a real number.*

When all quantile functions $\Psi^{-1}_{X_i(j)}(t)$ and $-\Psi^{-1}_{X_i(j)}(1-t)$ are collinear, expression (1) reduces to the classical linear regression model: between the centers when all intervals of the explicative interval-valued variables are degenerate or between the half ranges when all intervals of the explicative interval-valued variables are symmetric (see expressions (4)).

When, for all observations of the explicative variables, collinearity between $\Psi^{-1}_{X_i(j)}(t)$ and $-\Psi^{-1}_{X_i(j)}(1-t)$ occurs, the optimization problem has an optimal solution which is not unique since the quadratic function to optimize is not strictly convex (the columns of $\mathbf{X}$ in the optimization problem (7) are linearly dependent). However, all values of the parameters where the smallest value is attained allow obtaining the same model, that in these cases is a classical model between the centers or the half ranges.

The optimal solution of the quadratic optimization problem (5) verifies the *Kuhn Tucker conditions* (Winston, 1994; Dias and Brito, 2015).

2.4. Model evaluation measures

Let $\left(a_1^*, b_1^*, \cdots, a_p^*, b_p^*, v^*\right)$ be an optimal solution of the optimization problem in (5). According to Dias and Brito (2015) we may prove that:

- the symbolic mean of the predicted values is $\overline{\widehat{Y}} = \sum_{i=1}^{p} (a_i^* - b_i^*)\, \overline{X_i} + v^*$;
- using the Mallows distance, and as in classical regression, the total variation may be decomposed into sum of squares due to error and sum of squares due to regression, according to:

$$\sum_{j=1}^{n} D_M^2 \left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right) = \sum_{j=1}^{n} D_M^2 \left(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t)\right) + \sum_{j=1}^{n} D_M^2 \left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right)$$

This decomposition allows defining the goodness-of-fit measure for the *ID Model* for interval-valued variables.

Definition 2.5. *Consider the observed and predicted ranges of values of the interval-valued variable Y represented, respectively, by their quantile functions $\Psi^{-1}_{Y(j)}(t)$ and $\Psi^{-1}_{\widehat{Y}(j)}(t)$ with $t \in [0,1]$. Consider also the symbolic mean of the interval-valued variable Y, $\overline{Y}$. The goodness-of-fit measure Ω is given by*

$$\Omega = \frac{\sum_{j=1}^{n} D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right)}{\sum_{j=1}^{n} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)} = \frac{\sum_{j=1}^{n}\left(\left(c_{\widehat{Y}(j)} - \overline{Y}\right)^2 + \frac{1}{3} r^2_{\widehat{Y}(j)}\right)}{\sum_{j=1}^{n}\left(\left(c_{Y(j)} - \overline{Y}\right)^2 + \frac{1}{3} r^2_{Y(j)}\right)}.$$

As in classical linear regression, where the coefficient of determination R^2 ranges from 0 to 1, the goodness-of-fit measure Ω also ranges between 0 and 1 (Dias, 2014). The goodness-of-fit measure Ω is used to evaluate the linearity of the *ID Model*.

To measure the dissimilarity between the observed and predicted intervals, it is usual to compute the lower and the upper bound Root Mean Square Errors (Lima Neto and De Carvalho, 2008, 2010),

$$RMSE_L = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\underline{I}_{\widehat{Y}(j)} - \underline{I}_{Y(j)}\right)^2}; \quad RMSE_U = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\overline{I}_{\widehat{Y}(j)} - \overline{I}_{Y(j)}\right)^2}$$

and a measure defined with the Mallows distance proposed by Irpino and Verde (2015):

$$RMSE_M = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \Psi^{-1}_{Y(j)}(t)\right)^2 dt}.$$

2.5. Flexibility of the ID Model

2.5.1. The ID Model with degenerate intervals

The *ID Model* under uniformity is a theoretical generalization of the descriptive classical linear regression model. This generalization is in accordance

with the purpose of SDA, where symbolic variables are defined to generalize the classical concept of variable. Hence, the statistical concepts and methods defined for symbolic variables should also generalize the classical ones.

The *ID linear regression Model* may be written for classical variables whose values may be considered as degenerate intervals (the upper and lower bounds are equal, i.e. $\overline{I}_{X_i(j)} = \underline{I}_{X_i(j)} = u_{X_i(j)}$). In this particular case, expression (1) that allows predicting the values of response variable Y, is simplified to $\widehat{y}(j) = v + \sum_{i=1}^{p} (a_i - b_i)\, u_{X_i(j)}$ with $a_i, b_i \geq 0$, $i \in \{1, 2, \ldots, p\}$ and $v \in \mathbb{R}$. As we have referred above, in this situation, the function to optimize is not strictly convex, and therefore more than one optimal solution exists. However, for all optimal solutions for a_i and b_i we obtain the same diference $a_i - b_i$ and since no constraint is imposed on $a_i - b_i$, we have in this case a classical linear regression model. Moreover, the goodness-of-fit measure for interval-valued variables is also a generalization of the coefficient of determination R^2 for classical variables.

2.5.2. The ID Model assuming the Symmetric Triangular distribution

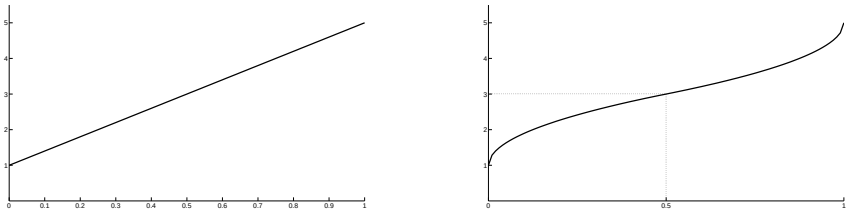
The flexibility of the *ID Model* goes beyond the reduction to the classical model and the generalization to histogram-valued variables (Dias and Brito, 2015). When applied to interval-valued variables, the method may be defined considering different distributions within the intervals. In this paper a detailed study for the Uniform distribution is presented. To consider other distributions within the intervals, it is necessary to include new definitions and deduce new results. It is important to notice that almost all theoretical results presented in this section take into account the uniformity condition. To put in evidence the potential of the proposed model, we present in this subsection some results for the case where a Symmetric Triangular distribution is assumed within the intervals of the observed interval-valued variables.

Consider a interval-valued variable Y for which the Symmetric Triangular distribution is assumed within all intervals $Y(j)$, with $j \in \{1, 2, \ldots, n\}$. Each interval $Y(j)$ with center $c_{Y(j)}$ and half-range $r_{Y(j)}$, may be represented by the

empirical quantile function defined as follows:

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} c_{Y(j)} - r_{Y(j)} + r_{Y(j)}\sqrt{2t}, & 0 \leq t \leq \frac{1}{2} \\ c_{Y(j)} + r_{Y(j)} - r_{Y(j)}\sqrt{2(1-t)}, & \frac{1}{2} < t \leq 1 \end{cases} \tag{8}$$

In this case the empirical quantile functions that represent the intervals are piecewise irrational functions with domain $[0, 1]$ and with $t = \frac{1}{2}$ as the boundary point between the two expressions.



(a) Uniform distribution. (b) Symmetric Triangular distribution.

Figure 2: Quantile functions assuming two different distributions.

In this case, again, $-\Psi^{-1}_{Y(j)}(t)$ is not the quantile function that represents the symmetric of the interval $Y(j)$. As in the case where the Uniform distribution is assumed, it is necessary to make a function transformation besides the multiplication by -1. The quantile function that represents the symmetric of the interval $Y(j)$ is given by

$$-\Psi^{-1}_{Y(j)}(1-t) = \begin{cases} -c_{Y(j)} - r_{Y(j)} + r_{Y(j)}\sqrt{2t}, & 0 \leq t \leq \frac{1}{2} \\ -c_{Y(j)} + r_{Y(j)} - r_{Y(j)}\sqrt{2(1-t)}, & \frac{1}{2} \leq t1 \end{cases}$$

In SDA, uniformity within the intervals is generally assumed. Therefore, Definition 2.2 of symbolic mean for interval-valued variable of Billard and Diday (2006) was deduced considering the density function of this distribution. It may be easily proven that, when considering the Triangular density function for the symmetric case, we obtain the same expression for the symbolic mean.

From Definition 2.5 and considering the quantile functions defined as in expression (8), the squared Mallows distance may now be rewritten as the sum

of the squared differences between the centers and one-sixth of the square of the differences between the half-ranges. We notice that the weight of the difference between the half-ranges with the Sf Symmetric Triangular distribution is lower than with the Uniform distribution.

Proposition 2.5. *Given two quantile functions $\Psi^{-1}_{X(j)}(t)$ and $\Psi^{-1}_{Y(j)}(t)$ representing the values of interval-valued variables X and Y for an observation j, and assuming the Symmetric Triangular distribution within each interval, the squared Mallows distance may be expressed as:*

$$D^2_M(\Psi^{-1}_{X(j)}, \Psi^{-1}_{Y(j)}) = (c_{X(j)} - c_{Y(j)})^2 + \frac{1}{6}(r_{X(j)} - r_{Y(j)})^2$$

where $c_{X(j)}, c_{Y(j)}$ are the centers and $r_{X(j)}, r_{Y(j)}$ are the half-ranges of the intervals $X(j)$ and $Y(j)$, respectively, with $j \in \{1, 2, \ldots, n\}$.

The *ID linear regression model* proposed in Definition 2.4 may be applied to interval-valued variables under these new conditions and the predicted quantile function in this case is the following:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \begin{cases} v + \sum_{i=1}^{p}[(a_i - b_i)c_{X_i(j)} - (a_i + b_i)r_{X_i(j)}(1 - \sqrt{2t})], & 0 \leq t \leq \frac{1}{2} \\ v + \sum_{i=1}^{p}[(a_i - b_i)c_{X_i(j)} - (a_i + b_i)r_{X_i(j)}(1 + \sqrt{2(1-t)})], & \frac{1}{2} < t \leq 1 \end{cases} \tag{9}$$

The closeness between the expressions of the Mallows distance when the Uniform or Symmetric Triangular distributions are assumed (see Definitions 2.1 and 2.5, respectively), leads to a similar behavior of the *ID Model* in both situations in spite of the different type of expressions of the quantile functions.

The parameters of the model are now obtained by solving a quadratic optimization problem very similar to the one presented in expression (5):

$$\begin{aligned} \min \quad & \sum_{j=1}^{n}\left[\left(c_{Y(j)} - \sum_{i=1}^{p}(a_i - b_i)\, c_{X_i(j)} - v\right)^2 + \tfrac{1}{6}\left(r_{Y(j)} - \sum_{i=1}^{p}(a_i + b_i)\, r_{X_i(j)}\right)^2\right] \\ \text{subject to} \quad & -a_i, -b_i \leq 0, \qquad i \in \{1, 2, \ldots, p\}, \quad \text{with} \quad v \in \mathbb{R}. \end{aligned}$$

The conditions, in Section 2.4, that allow deducing the goodness-of-fit measure associated to the *ID Model* are also verified and consequently, Definition 2.5 for the measure named Ω, may also be rewritten for interval-valued variables under the condition of the Symmetric Triangular distribution.

In the applied examples studied in Section 4 - Applications to real data sets, we will present the *ID Model* results assuming both distributions.

3. Experiments with simulated data

In this section we evaluate the performance of the proposed method, as done in the literature, see (Irpino and Verde, 2015; Blanco-Fernández et al., 2011). With this objective, two studies are presented: 1) a simulation study designed to evaluate the behavior of the ID Model parameters under different conditions and considering two levels of linearity, as evaluated by the measure Ω and 2) experiments with interval data simulated with different configurations, aimed at analysing the performance of the ID model.

3.1. Evaluating the behavior of the ID Model parameters

3.1.1. Building the data sets

To build the simulated symbolic data tables it is necessary to generate the observations of the explicative interval-valued variables X_i, $i = \{1, \ldots, p\}$ and of Y, the response variable.

- To obtain the n observations of variable X_i, we start by uniformly simulating 5000 real values for each unit. Then from these $x_{ji}(w)$ values we build the corresponding intervals considering the minimum and maximum values. We considered different patterns of variability:

  i) *Low variability* - when the intervals associated with variable X_i have similar centers and small half ranges.
  $x_{ji}(w) \sim \mathcal{U}(\delta_1(j), \delta_2(j))$ are randomly generated considering for each $j \in \{1, \ldots, m\}$ and $i \in \{1, 2, 3\}$:

- $i = 1 : \delta_1(j) \sim \mathcal{U}(17, 19)$ and $\delta_2(j) \sim \mathcal{U}(21, 23)$ (centers around 20 and half ranges around 2);
- $i = 2 : \delta_1(j) \sim \mathcal{U}(13.5, 15)$ and $\delta_2(j) \sim \mathcal{U}(15, 16.5)$ (centers around 15 and half ranges around 0.5);
- $i = 3 : \delta_1(j) \sim \mathcal{U}(8, 10)$ and $\delta_2(j) \sim \mathcal{U}(10, 12)$ (centers around 10 and half ranges around 1).

ii) *High variability* - when the intervals associated with variable X_i have similar centers and high half ranges.
$x_{ji}(w) \sim \mathcal{U}(\delta_3(j), \delta_4(j))$ are randomly generated considering for each $j \in \{1, \ldots, n\}$ and $i \in \{1, 2, 3\}$:

- $i = 1 : \delta_3(j) \sim \mathcal{U}(4, 6)$ and $\delta_4(j) \sim \mathcal{U}(34, 36)$ (centers around 20 and half ranges around 15);
- $i = 2 : \delta_3(j) \sim \mathcal{U}(4, 6)$ and $\delta_4(j) \sim \mathcal{U}(24, 26)$ (centers around 15 and half ranges around 10);
- $i = 3 : \delta_3(j) \sim \mathcal{U}(1, 3)$ and $\delta_4(j) \sim \mathcal{U}(17, 19)$ (centers around 10 and half ranges around 8).

iii) *Mixed* variability - when the intervals associated with the variable X_i have diverse half ranges and centers. In this case, the intervals of X_i have centers between 8 and 22 and half ranges between 1 and 15.

- The intervals that are the observations of the interval-valued variable Y are obtained in two steps:

1. First, we consider a perfect linear regression, without error, given by $\Psi^{-1}_{Y^*(j)}(t) = v + \sum_{i=1}^{p} a_i \Psi^{-1}_{X_i(j)}(t) - \sum_{i=1}^{p} b_i \Psi^{-1}_{X_i(j)}(1-t)$ for particular values of the parameters a_i, b_i and v. The interval-valued variables X_i and Y^* are in a perfect linear relation.
2. To disturb the perfect linear relations we introduce an error function, $\Psi^{-1}_{Y(j)}(t) = \Psi^{-1}_{Y^*(j)}(t) + e_j(t)$. This error function is a linear function defined by $e_j(t) = \widetilde{c}(j) + (2t-1)\,\widetilde{r}(j), \quad t \in [0, 1]$. The values of $\widetilde{c}(j)$

and $\widetilde{r}(j)$ are randomly selected from intervals with low or high variation depending on whether we want the linear regression between the variables to be better (error level I) or worse (error level II).

For each $j \in \{1, \ldots, n\}$, the values of $\widetilde{c}(j)$ and $\widetilde{r}(j)$ are randomly generated as follows:

i) Level I: $\widetilde{c}(j) \sim 0.1 \times \mathcal{U}(-Mr, Mr)$ and $\widetilde{r}(j) \sim 0.1 \times \mathcal{U}(-mr, mr)$;

ii) Level II: $\widetilde{c}(j) \sim \mathcal{U}(-Mr, Mr)$ and $\widetilde{r}(j) \sim \mathcal{U}(-mr, mr)$

where $Mr = \max\limits_{j \in \{1,\ldots,n\}} \left\{ r_{Y^*(j)} \right\}$ and $mr = \min\limits_{j \in \{1,\ldots,n\}} \left\{ r_{Y^*(j)} \right\}$.

Note 1. *The values of $\widetilde{r}(j)$ have a limitation: each half range $r_{Y(j)}$ in the quantile function $\Psi^{-1}_{Y(j)}(t)$, that results from the perturbation of $\Psi^{-1}_{Y^*(j)}(t)$ by the error function $e_j(t)$, is obtained by $r_{Y(j)} = r_{Y^*(j)} + \widetilde{r}(j)$, for each unit j. As it is not imposed that the error function is a quantile function, the values of $\widetilde{r}(j)$ may be negative but cannot be lower than $-r_{Y^*(j)}$ else for this unit j the half range $r_{Y(j)}$ would be negative.*

Note 2. *To select the error levels, a preliminary simulation study has been made, to analyze the behavior of the error function and see if it is possible to establish a relation with Ω. Therefore, it is important to understand: 1) how much it is necessary to disturb the model to obtain a weak/strong linear relation between intervals, i.e. how small/large should be the values composing the error function for which the coefficient of determination Ω evaluates the linear relation as weak or strong and 2) whether the pattern of variability influences the values of Ω. From this preliminary study, we concluded that the disturbance of the centers must be sufficiently large to be "detected" by Ω. It can be observed that, independently from $\widetilde{r}(j)$, the values of Ω are only lower than 0.5 when the values of M_r are close or larger than the maximum of the half ranges of the intervals $\left[\underline{I}_{Y^*(j)}, \overline{I}_{Y^*(j)}\right]$, with $j \in \{1, \ldots, n\}$. For this reason, to obtain similar values of Ω for different patterns of variability, the error functions depend on the values of the half ranges involved in the linear relations.*

Four selections of parameters were considered. For $p = 1 : a = 2$; $b = 1$; $v = -1$ (a and b are close); $a = 2$; $b = 8$; $v = 3$ (a is smaller than b); $a = 6$; $b = 0$; $v = 2$ (a is larger than b) and for $p = 3 : a_1 = 2$; $b_1 = 1$; $a_2 = 0.5$; $b_2 = 3$; $a_3 = 1.5$; $b_3 = 1$; $v = -1$. For each selection, we analyze the behavior of the *ID Model* considering symbolic data tables with different sample sizes ($n = 10;\ 30;\ 100;\ 250$); the three different patterns of variability in the explicative variables and the two levels of linearity between the interval-valued variables, as defined above. For each case, 1000 data tables were generated. The mean values of the obtained results are organized in tables provided in the Supplementary Material.

3.1.2. Results and discussion

The results obtained for the *ID Model* with one or three explicative variables are in general similar (see tables in Supplementary Material) and therefore we will analyze in detail only the results obtained when $p = 1$.

In general, the results show that the behavior of the parameters' estimation is independent from the number of explicative variables and from the values of the parameters selected for the model. For each selection of parameters, three patterns of variability in the explicative variables X_i were considered, each of them with two levels of linearity and a different behavior was observed.

When we consider an error function of level II, the behavior of the estimated parameters is more unstable and their mean values are more distant from the original values. Consequently, the mean square errors (MSE) for the parameters are not close to zero. This is not surprising because other models may exist that adjust the interval data better. The behavior of the parameters is more unstable when the number of observations is low and when the variability in the explicative variables is *high*.

When the linear relation between the variables is strong (error functions of level I), the estimated parameters are close to the initial parameter values and the closeness is more obvious as the number of units in the sample increases. The values of the MSE, associated with each parameter, decrease and approach zero

as the number of observations increases. The values of the MSE and the values of the standard deviation associated with the mean value of the parameters are more distant from zero when the half ranges of the intervals of Y are larger - which occurs when the variability of X is similar and high and when the values of the parameters are far apart. The justification may be that, under these conditions, the disturbance is "larger", in absolute terms.

The boxplots in Figures 3 to 5 illustrate the behavior of the parameters for the situations where $p = 1$, the original values of the parameters are $a = 2; b = 8; v = 3$ and when the level I error is considered.

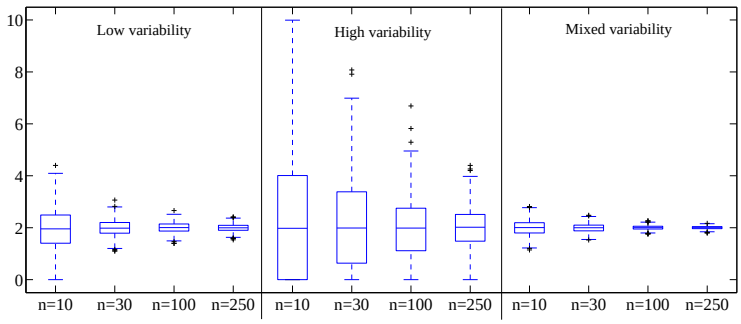


Figure 3: Boxplots of the estimates of the parameter a, under different conditions, when *ID Model* with $a = 2; b = 8; v = 3$ is applied and when the level I error is considered.

We may then conclude that when an error function of level I is considered, the estimates of the independent parameter may be quite different from the original. Consequently, the standard deviation associated with the mean value of the independent parameters and the respective values of the MSE are much higher than the values of the standard deviation and the MSE associated with the other parameters. The larger instability of the independent parameters (see Figure 5) may be explained because they are obtained from the other parameters: $v^* = \overline{\widehat{Y}} - \sum_{i=1}^{p} (a_i^* - b_i^*)\, \overline{X_i}$.

This simulation also allowed confirming the empirical consistency of the parameters' estimation. Estimated values close to the original parameters, with

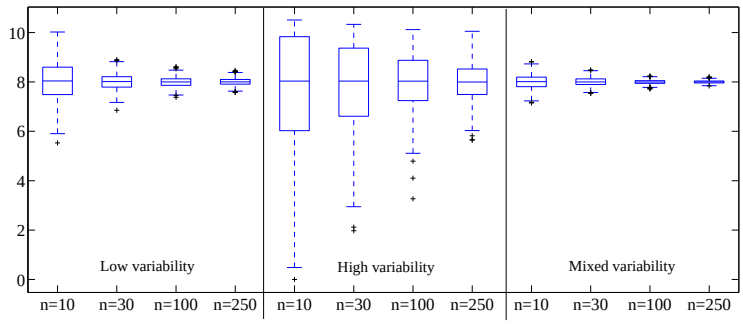


Figure 4: Boxplots of the estimates of the parameter b, under different conditions, when *ID Model* with $a = 2; b = 8; v = 3$ is applied and when the level I error is considered.

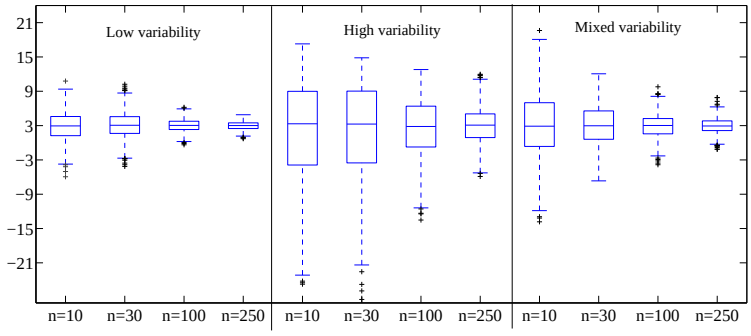


Figure 5: Boxplots of the estimates of the parameter v, under different conditions, when *ID Model* with $a = 2; b = 8; v = 3$ is applied and when the level I error is considered.

associated MSE that approach zero as the number of observations increases, was to be expected when the linear relation is strong, i.e. Ω is close to one.

To evaluate the behavior of the parameters' estimates under different conditions and considering two levels of linearity, it is important to verify that Ω has the expected behavior, according to the preliminary study (see *Note 2*). The models slightly disturbed (error level I) present values of Ω close to one. On the other hand, when the error function causes a high disturbance in the linear relation (error level II), the values of Ω are more distant from one and closer

to zero. It is important to bear in mind that the level of the error takes into account the variability within the intervals of Y^*. Also, the mean values of Ω are consistent with the respective mean values of $RMSE_M$; $RMSE_L$ and $RMSE_U$. In general, as expected, in each situation and for the respective pattern of variability in the explicative variables, the highest values of Ω correspond to the lowest values of $RMSE_M$. Small values in $RMSE_M$; $RMSE_L$ and $RMSE_U$ mean that the observed and predicted intervals are close and in this case we expected values of Ω close to one.

3.2. Analysing ID performance

The main goal of this section is to evaluate the performance of the *ID* linear regression model, and compare it to that of the *CRM* (Lima Neto and De Carvalho, 2008) and the *CCRM* (Lima Neto and De Carvalho, 2010).

It should be noticed, however, that given the nature of the method, and the constraints imposed, it cannot be expected that the *ID Model* under uniformity provides better prediction results than *CRM* or *CCRM*. *ID* models directly interval-valued variable Y from interval-valued variables X_i, providing one single model. Under uniformity, this model then induces linear relations between the centers and betwen the half ranges, which are related and not independent. On the other hand, both *CRM* and *CCRM* model centers and half ranges separately, using classical linear regression, and not imposing any connection between the two models. Therefore, the search space is larger for *CRM* and *CCRM*, which then implies that more accurate predictions should be expected. Notice that *CCRM*, imposing a constraint on the half ranges model, already often reduces the quality of predictions as compares to *CRM* (Lima Neto and De Carvalho, 2010).

This study is based on synthetic interval-valued data sets built according to the following strategy:

To obtain the n observations of variable X_i, $i = \{1, \ldots, p\}$ we started by simulating uniformly 5000 real values for each observation, and then built the corresponding interval from the minimum and maximum obtained values, as in

the previous simulation. From the observations of the intervals X_i, the centers $c_{X_i(j)}$ and half ranges $r_{X_i(j)}$ of each interval are computed. The intervals that are the observations of the response variable Y are obtained as follows:

$$c_{Y(j)} = a_0 + \sum_{i=1}^{p} a_i c_{X_i(j)} + e_c(j) \text{ and } r_{Y(j)} = b_0 + \sum_{i=1}^{p} b_i r_{X_i(j)} + e_r(j).$$

Considering the process described above, we defined that the intervals associated with the variables X_i would have centers between 18 and 40 and half ranges between 0 and 4. The situations considered comprise cases of single regression (cases A to H) and of multiple regression with three explicative variables (cases I to L). For the single regression, in cases A, B, C and D the relation between the half ranges is positive whereas in cases E, F, G and H it is negative. Different levels of linearity between the centers and between the half ranges are considered, according to the values of the errors $e_c(j)$ and $e_r(j)$. The relations between the centers and between the half ranges, as well as the levels of linearity considered in this simulation, are presented in Table 1.

Table 1: Relations and levels of linearity between the centers and between the half ranges.

Cases	Parameters	Center error	Half Range error
A	$a_0 = 3;\ a_1 = 2;\ b_0 = 1;\ b_1 = 0.5$	$e_c(j) \sim \mathcal{U}(-5, 5)$	$e_r(j) \sim \mathcal{U}(0.5, 1.5)$
B		$e_c(j) \sim \mathcal{U}(-5, 5)$	$e_r(j) \sim \mathcal{U}(1, 5)$
C		$e_c(j) \sim \mathcal{U}(-20, 20)$	$e_r(j) \sim \mathcal{U}(0.5, 1.5)$
D		$e_c(j) \sim \mathcal{U}(-20, 20)$	$e_r(j) \sim \mathcal{U}(1, 5)$
E	$a_0 = 3;\ a_1 = 2;\ b_0 = 1;\ b_1 = -3$	$e_c(j) \sim \mathcal{U}(-5, 5)$	$e_r(j) \sim \mathcal{U}(0.5, 1.5)$
F		$e_c(j) \sim \mathcal{U}(-5, 5)$	$e_r(j) \sim \mathcal{U}(1, 5)$
G		$e_c(j) \sim \mathcal{U}(-20, 20)$	$e_r(j) \sim \mathcal{U}(0.5, 1.5)$
H		$e_c(j) \sim \mathcal{U}(-20, 20)$	$e_r(j) \sim \mathcal{U}(1, 5)$
I	$a_0 = 3;\ a_1 = 2;\ a_2 = 1;$ $a_3 = -0.5;\ b_0 = 1;\ b_1 = 0.25;$ $b_2 = -0.5;\ b_3 = 1$	$e_c(j) \sim \mathcal{U}(-5, 5)$	$e_r(j) \sim \mathcal{U}(0.5, 1.5)$
J		$e_c(j) \sim \mathcal{U}(-5, 5)$	$e_r(j) \sim \mathcal{U}(1, 5)$
K		$e_c(j) \sim \mathcal{U}(-20, 20)$	$e_r(j) \sim \mathcal{U}(0.5, 1.5)$
L		$e_c(j) \sim \mathcal{U}(-20, 20)$	$e_r(j) \sim \mathcal{U}(1, 5)$

For each case we simulated $1000 + 250 \times 100 + 30 \times 100$ observations that were

grouped into two learning data sets, one with $n = 30$ and another with $n = 250$, and a test set comprising 1000 observations. For each of the two learning data sets, 100 replicates were created. The models *ID*, *CRM* and *CCRM* were applied to the learning sets and the parameters of the respective model were predicted. With these parameters, we applied the 100 replicates created for each of the models to the test set. The dissimilarity between the observed and predicted intervals is measured by the $RMSE_M$. The mean values of these measures are presented in Supplementary Material.

We note that centers on the one hand, and half ranges on the other hand, are modeled according to a linear regression that is not obtained from the *ID Model*, and could not be recovered by the method (since the linear relations between centers and between half ranges do not present the restriction induced by the *ID Model*).

The results for all three methods in single and multiple regression are quite similar, although, as expected, the *ID Model* performs slightly worse than *CRM* and *CCRM*. The differences are however not too important, considering that the data was generated by separate linear models for the centers and half ranges, and therefore in conditions following *CRM* and *CCRM*, but not *ID Model*.

4. Applications to real data sets

In this section we show that when analysing data with variability, considering interval-valued variables is a good option as an alternative to summarizing data by central statistics (e.g., mean, median), to allow for a representation by classical real variables. In each case we analyze the relation between variables representing the variability within each observation, and we evaluate the performance of the *ID Model*.

4.1. Employment data: Relation between time of unemployment and years of activity

The 2008 Portuguese Labour Force Survey provides individual information about the people that live in Portugal. Here we analyze whether the time of

unemployment (in months) is related to the previous time of activity (in years). However, we are not interested in performing this study for each individual, but rather to determine what happens in certain categories, such as 'young women from the North of Portugal'. Since each of these categories consists of several individuals, the observed "values" are no longer single points but intervals. In this case, the symbolic data table is built considering that the units are classes of individuals obtained by crossing $gender \times region \times age \times education$. There are two genders (female (F), male (M)), four regions (North (N), Center (C), Lisbon and Tagus Valley (L), South (S)), three age groups (15 to 24 (A1), 25 to 44 (A2), 45 to 64 (A3)) and three levels of education (basic (B), secondary (S) and graduate (G)), leading to $2 \times 4 \times 3 \times 3 = 72$ possible classes (categories). The time of unemployment and the time of activity are now interval-valued variables.

Table 2 represents a portion of the symbolic table resulting from the original data, for the variables U (time of unemployment) and E (time of activity before unemployment). Only 58 classes (units) were created, since there are no cases corresponding to the remaining 14.

Table 2: Symbolic data table where time of unemployment (U) and time of activity before unemployment (E) are interval-valued variables (partial view).

Units	U	E	Units	U	E	Units	U	E
$F \times C \times A1 \times B$	$[3; 49]$	$[0; 4]$	$F \times N \times A3 \times S$	$[0; 123]$	$[23; 35]$	$M \times L \times A3 \times B$	$[1; 244]$	$[22; 57]$
$F \times C \times A1 \times S$	$[1; 6]$	$[0; 2]$	$F \times S \times A1 \times B$	$[1; 52]$	$[1; 7]$	$M \times L \times A3 \times S$	$[2; 65]$	$[25; 50]$

We analyze the linear relation between the logarithm of the time of unemployment, LNU ($LNU = ln(U + 2)$), and the time of activity before unemployment, E, for the classes of individuals as described above.

The prediction, with the *ID Model*, of the quantile functions for the interval-valued variable LNU, is in this case:

$$\Psi^{-1}_{\widehat{LNU}(j)}(t) = 2.2277 + 0.0779\Psi^{-1}_{E(j)}(t) - 0.0503\Psi^{-1}_{E(j)}(1-t), \quad t \in [0, 1].$$

Equivalently, the predicted interval for each unit j, is given by

$$\left[2.2277 + 0.0276c_{E(j)} - 0.1282r_{E(j)},\ 2.2277 + 0.0276c_{E(j)} + 0.1282r_{E(j)}\right].$$

The value of the goodness-of-fit measure Ω is 0.7715, that shows that 77% of the total variation in the interval-valued variable LNU is explained by the linear relation. According to the interpretation given in Section 2.2, the interval-valued variables E and LNU have a linear relation that tends to be direct, because the estimate of parameter $a = 0.0779$ is slightly higher than that of $b = 0.0503$. For the set of classes of individuals which the data refers to, when the symbolic mean of the time of activity before unemployment increases one year, the symbolic mean of the LNU (in months) increases 0.0276. From the linear relation between the half-ranges induced by the model (see Table 3), it is possible to conclude that the variability of time unemployment and the variability of time of activity before unemployment present a ratio of 0.13.

We compare the *ID Model* resulting from assuming the Uniform distribution (identified by *ID*) and the Symmetric Triangular distribution (identified by ID_T) with those obtained by other models proposed within the SDA context: *CM* (Billard and Diday, 2000); *MinMax Method (MinMax)* (Billard and Diday, 2002); *Billard and Diday Method (BD)* (Billard and Diday, 2006); *CRM* (Lima Neto and De Carvalho, 2008) and *CCRM* (Lima Neto and De Carvalho, 2010). In Tables 3 and 4 we present the obtained model expressions, the Root Mean Square Error (RMSE) values and the mean of the areas between the observed and predicted quantile functions ($\overline{Area}$).

In this example, the *CRM* and *CCRM* models are the same because in the *CRM* the parameters estimated for the half ranges are all non-negative, i.e. the constraints imposed in the *CCRM* are met. The linear regression induced by the *ID Model* for the centers of the intervals is the one obtained by the models for which a linear regression between the centers is considered. In this situation, the values of the parameters (to the second decimal) are the same for both distributions considered within the observations of the interval-valued variables. However, for the Symmetric Triangular distribution the value of $\Omega = 0.7086$ is slightly lower.

Applying the Leave-One-Out method, we may verify that the $RMSE$ values and the mean of the areas between the observed and predicted quantile functions

Table 3: Expressions of the symbolic linear regression models for interval-valued variables in Table 2.

Models	Expressions that allow predicting the intervals
ID	$\Psi^{-1}_{\widehat{LNU}(j)}(t) = 2.2277 + 0.0779\Psi^{-1}_{E(j)}(t) - 0.0503\Psi^{-1}_{E(j)}(1-t)$ $c_{\widehat{LNU}(j)} = 2.2277 + 0.0276\, c_{E(j)}$ *and* $r_{\widehat{LNU}(j)} = 0.1282\, r_{E(j)}$
ID_T	$\Psi^{-1}_{\widehat{LNU}(j)}(t) = 2.2277 + 0.0779\Psi^{-1}_{E(j)}(t) - 0.0503\Psi^{-1}_{E(j)}(1-t)$ $c_{\widehat{LNU}(j)} = 2.2277 + 0.0276\, c_{E(j)}$ *and* $r_{\widehat{LNU}(j)} = 0.1282\, r_{E(j)}$
CM	$c_{\widehat{LNU}(j)} = 2.2277 + 0.0276\, c_{E(j)}$
BD	$\widehat{LNU}(j) = 1.9009 + 0.0468\, E(j)$
MinMax	$\underline{I}_{\widehat{LNU}(j)} = 1.2236 + 0.0206\underline{I}_{E(j)}$ *and* $\overline{I}_{\widehat{LNU}(j)} = 2.8704 + 0.0436\overline{I}_{E(j)}$
CRM	$c_{\widehat{LNU}(j)} = 2.2277 + 0.0276\, c_{E(j)}$ *and* $r_{\widehat{LNU}(j)} = 0.5321 + 0.0855\, r_{E(j)}$
CCRM	$c_{\widehat{LNU}(j)} = 2.2277 + 0.0276\, c_{E(j)}$ *and* $r_{\widehat{LNU}(j)} = 0.5321 + 0.0855\, r_{E(j)}$

Table 4: Comparison of the Root Mean Square Error values when Leave-One-Out is not/is applied together with the proposed models for the data in Table 3.

Models	Without Leave-One-Out				With Leave-One-Out			
	$RMSE_L$	$RMSE_U$	$RMSE_M$	$\overline{Area}$	$RMSE_L$	$RMSE_U$	$RMSE_M$	$\overline{Area}$
ID	0.5745	0.6710	0.4679	0.3674	0.5866	0.6829	0.4797	0.3773
ID_T	0.5745	0.6710	0.4197	0.3674	0.5866	0.6829	0.4317	0.3773
CM	1.1622	1.3146	0.7759	0.6192	1.1651	1.3188	0.7818	0.6246
BD	1.0368	1.1499	0.7255	0.5730	1.0410	1.1589	0.7337	0.5805
MinMax	0.4725	0.7329	0.4621	0.3700	0.4940	0.7568	0.4781	0.3829
CRM	0.4457	0.6541	0.4397	0.3565	0.4597	0.6740	0.4539	0.3682
CCRM	0.4457	0.6541	0.4397	0.3565	0.4597	0.6740	0.4539	0.3682

are very similar in all models compared, see Table 4. As expected, the $RMSE$ values when the Leave-One-Out method is applied are slightly higher than those obtained without the Leave-One-Out.

The values of the $RMSE$ and the mean values of the areas between the observed and predicted quantile functions allow comparing the predicted and

the observed intervals of the response variable LNU. These measures may be used for an independent comparison of results. The mean values of the areas are in accordance with the other measures. From Table 4 we may conclude that the *ID Model* and *CRM (and CCRM)* provide similar results. It may also be observed that the values of the $RMSE_L$, $RMSE_U$ and the *mean of the Area* are the same when the two distributions studied in this paper are considered. It is when the Symmetric Triangular distribution is assumed that the value of $RMSE_M$ is the lowest, with and without the Leave-One-Out.

4.2. Forest fires data: predicting burned area in the northeast region of Portugal

This study concerns forest fire data from the Montesinho natural park, in the northeast of Portugal. The original data was collected from January 2000 to December 2003 using two sources. The first database was collected by the responsible for the Montesinho fire occurrences, registering date, spatial location, type of vegetation involved, the six components of the FWI system (FFMC, DMC, DC, ISI, BUI, FWI) and the total burned area. The second database was collected by the Bragança Polytechnic Institute, and contains the weather observations recorded (temperature, relative humidity, wind, rain) by a meteorological station located in the center of the Montesinho park. Details are described in Cortez and Morais (2007).

For this study we selected the response variable *area* (the forest burned area (in ha)) and three explicative variables that better explain the hectares of area burned: DMC (index from the FWI system that is a numeric rating of the average humidity content of lightly compacted organic layers of moderate depth); *wind* (wind speed in km/h); rh (relative humidity in percentage). As in the study of Cortez and Morais (2007), the response variable *area* was transformed with a $ln(x+1)$ function and we represent it as $LNarea$. We aggregated the information according to the coordinates of the spatial location within the Montesinho park map, thereby obtaining the symbolic data (macrodata). The units (higher units) of this study are locations defined by spatial coordinates, the observations of the variables DMC, *wind*, rh and $LNarea$ for each location

are organized in intervals. To build these macrodata we only considered the places and records in which forest fires occurred. Under these conditions, from 269 first-level units in the microdata, 33 higher-level units were obtained (after spatial aggregation), the locations with forest fires recorded. In nine of these 33 locations only one fire occurred and consequently the "symbolic values" associated with all variables are degenerate intervals. Table 5 presents the data with the first five records, organized in two different ways. In the even columns we have symbolic variables obtained from the spatial aggregation of the set of records associated with each variable. In the odd columns we registered the logarithm of the total burned area $LNareaT$, and the mean values of the DMC, $wind$ and rh for each place, i.e., the classical variables.

Table 5: Data with information about the total burned area and other four variables: $LNarea$, DMC, $wind$ and rh, organized according the spatial location (partial view).

Units	$LNarea$	$LNareaT$	DMC	$\overline{DMC}$	$wind$	$\overline{wind}$	rh	$\overline{rh}$
1	[0.44; 5.37]	8.26	[51.30; 163.20]	110.98	[1.80; 5.40]	3.48	[31; 53]	39.75
2	[0.29; 4.27]	16.18	[91.30; 276.30]	144.83	[2.20; 6.70]	4.53	[29; 73]	44.43
3	[0.36; 4.43]	20.75	[87.70; 276.30]	144.66	[1.80; 7.60]	3.69	[28; 88]	50.0
4	[0.90; 4.57]	10.11	[126.50; 149.30]	137.68	[2.20; 3.10]	2.55	[27; 42]	32.25
5	[0.42; 5.31]	27.68	[3.60; 231.10]	124.03	[0.90; 7.20]	3.75	[22; 79]	44.36

The symbolic models that allow predicting the intervals of $LNarea$ from the explicative interval-valued variables $DMC, wind$ and rh, for each location j, are presented in Table 6.

This example aims at illustrating the behavior of the *ID Model* in a situation of multiple regression, assuming the Uniform and the Symmetric Triangular distributions and compare it with that of other proposed models. For all these models, if we know the interval for the DMC, wind and relative humidity for one location, we may predict the range of burned forest area. Since in this case the microdata is known, a comparison with two approaches based on classic linear

Table 6: Comparison of the symbolic linear regression models for the data in Table 5.

Models	Expressions that allow predicting the intervals
ID	$\Psi^{-1}_{\widehat{LNarea}(j)}(t) = 1.0307 + 0.0062\Psi^{-1}_{DMC(j)}(t) - 0.0005\Psi^{-1}_{DMC(j)}(1-t) +$ $+0.2742\Psi^{-1}_{wind(j)}(t) + 0.0076\Psi^{-1}_{rh(j)}(t) - 0.0153\Psi^{-1}_{rh(j)}(1-t)$ $c_{\widehat{LNarea}(j)} = 1.0307 + 0.0057c_{DMC(j)} + 0.2742c_{wind(j)} - 0.0077c_{rh(j)}$ $r_{\widehat{LNarea}(j)} = 0.0067r_{DMC(j)} + 0.2742r_{wind(j)} + 0.0229r_{rh(j)}$
ID_T	$\Psi^{-1}_{\widehat{LNarea}(j)}(t) = 1.0172 + 0.0062\Psi^{-1}_{DMC(j)}(t) - 0.0005\Psi^{-1}_{DMC(j)}(1-t) +$ $+0.2780\Psi^{-1}_{wind(j)}(t) + 0.0075\Psi^{-1}_{rh(j)}(t) - 0.0152\Psi^{-1}_{rh(j)}(1-t)$ $c_{\widehat{LNarea}(j)} = 1.0172 + 0.0057c_{DMC(j)} + 0.2780c_{wind(j)} - 0.0077c_{rh(j)}$ $r_{\widehat{LNarea}(j)} = 0.0067r_{DMC(j)} + 0.2780r_{wind(j)} + 0.0227r_{rh(j)}$
CM	$c_{\widehat{LNarea}(j)} = 1.0016 + 0.0057c_{DMC(j)} + 0.2825c_{wind(j)} - 0.0078c_{rh(j)}$
BD	$\widehat{LNarea}(j) = 2.4393 - 0.0023DMC(j) + 0.0079wind(j) - 0.0039rh(j)$
MinMax	$\underline{I}_{\widehat{LNarea}(j)} = -1.0974 + 0.0039\underline{I}_{DMC(j)} + 0.4819\underline{I}_{wind(j)} + 0.0325\underline{I}_{rh(j)}$ $\overline{I}_{\widehat{LNarea}(j)} = 0.9308 + 0.0061\overline{I}_{DMC(j)} + 0.2546\overline{I}_{wind(j)} + 0.0043\overline{I}_{rh(j)}$
CRM	$c_{\widehat{LNarea}(j)} = 1.0016 + 0.0057c_{DMC(j)} + 0.2825c_{wind(j)} - 0.0078c_{rh(j)}$ $r_{\widehat{LNarea}(j)} = 0.1356 + 0.0074r_{temp(j)} + 0.2233r_{wind(j)} + 0.0206r_{rh(j)}$
CCRM	$c_{\widehat{LNarea}(j)} = 1.0016 + 0.0057c_{DMC(j)} + 0.2825c_{wind(j)} - 0.0078c_{rh(j)}$ $r_{\widehat{LNarea}(j)} = 0.1356 + 0.0074r_{DMC(j)} + 0.2233r_{wind(j)} + 0.0206r_{rh(j)}$

regression models will also be considered.

For the *ID Model* the value of the goodness-of-fit measure is $\Omega = 0.5506$, for the case of the ID_T*Model* this value is a bit lower, $\Omega = 0.4280$.

As all of the estimated parameters of the model associated with the half ranges in *CRM* are non-negative, the expression for the half ranges in *CCRM* is the same. Moreover, the relations between the centers and half ranges induced by the *ID Model* are very similar to those obtained by *CRM* and *CCRM*. This behavior was also observed in the previous example, but it should be underlined that it is not always the case.

In Figure 6 we may observe the prediction of the burned area ($LNarea$) for five locations, obtained with the symbolic methods in Table 6. The locations

are represented in the xx-axis and the observed and predicted intervals are represented in the yy-axis.

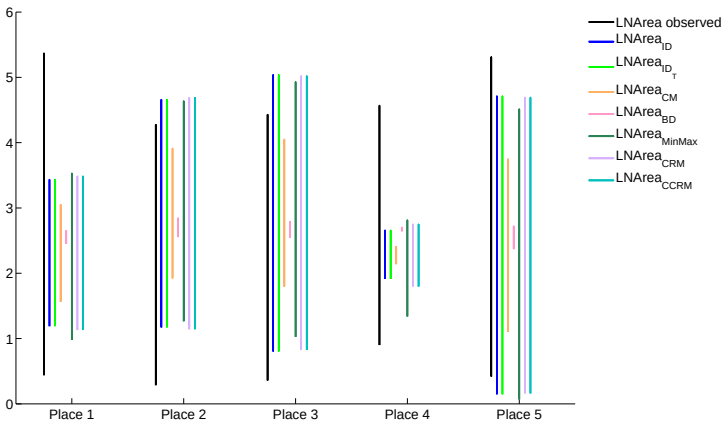


Figure 6: Observed and predicted intervals of the burned area in five locations of the Montesinho natural park.

In Table 7, for the models selected, it is possible to compare the measures $RMSE_L$, $RMSE_U$, $RMSE_M$ and the mean value of the areas between the predicted and observed quantile functions ($\overline{Area}$), with and without Leave-One-Out.

As observed in the example of Section 4.1, when the Leave-One-Out method is not applied, the results of the $RMSE$ and the $\overline{Area}$, for the *CRM, CCRM*, *ID* and ID_T models are again very similar. These results are also similar to the ones obtained for the *MinMax* model. The values of $RMSE$ with and without the Leave-One-Out method are in general close. The slight difference observed for the *ID Model* shows that the model was not overfitting the data. The lowest value of the $RMSE_M$ with and without the Leave-One-Out method is observed when the Symmetric Triangular distribution in the *ID Model* is assumed.

Since in this situation the microdata is known, alternatively to the symbolic study we performed 1) a classical study and 2) another approach that we named

Table 7: Comparison of the Root Mean Square Error values when Leave-One-Out is not/is applied together with the proposed models for the data in Table 5.

Models	Without Leave-One-Out				With Leave-One-Out			
	$RMSE_L$	$RMSE_U$	$RMSE_M$	$\overline{Area}$	$RMSE_L$	$RMSE_U$	$RMSE_M$	$\overline{Area}$
ID	0.7824	1.1246	0.8773	0.7157	0.9056	1.2787	1.0148	0.8236
ID_T	0.7818	1.1251	0.8529	0.7159	0.9071	1.2819	0.9931	0.8258
CM	1.0382	1.2558	0.9483	0.7744	1.1267	1.3767	1.0756	0.8761
BD	1.6874	1.8036	1.2411	1.0229	1.7139	1.8730	1.3155	1.0812
$MinMax$	0.7598	1.1087	0.8666	0.7084	0.9154	1.2595	1.0053	0.8172
CRM	0.7837	1.1168	0.8758	0.7127	0.9258	1.2872	1.0239	0.8296
$CCRM$	0.7837	1.1168	0.8758	0.7127	0.9258	1.2872	1.0239	0.8296

classic-symbolic.

1. The classical study allows predicting the value of the total burned area from the mean values of the three weather-related variables, associated with each location: DMC, $wind$ and rh (see Table 5). However, in this case we lose the variability of the data and the predicted results are less informative. The classical linear regression model for these data is: $\widehat{LNareaT}(j) = -8.3425 - 0.0617\overline{DMC}(j) + 2.1984\overline{wind}(j) + 0.5908\overline{rh}(j)$. The classical coefficient of determination of the model is $r^2 = 0.1051$, i.e., only 10.5% of the variance of the value of the total burned area is explained by the variations of the DMC component of the FWI system and the weather variables wind and relative humidity. Although the burned area of the forest seems to be influenced by the FWI system components and weather factors in the model, in this case it is not the classical linear regression that better explains the relations between these variables.
2. The classic-symbolic study is an alternative to analyze data with variability using classical methods. The first step is the prediction of the burned area. Applying classical linear regression to all values observed for each

first level unit, the microdata, the obtained model is:

$$\widehat{LNarea}(j) = 2.0857 + 0.0110DMC(j) - 0.0061wind(j) + 0.0446rh(j).$$

The value of $r^2 = 0.0104$ associated with this model shows that there is no linear relation between the variables in the microdata.
Afterwards, the predicted values are aggregated by location obtaining, for each specific location, the range of hectares of the burned area. As after the aggregation the elements are of different nature, the behavior between the variables may be different. Comparing the observed and predicted intervals obtained by the classic-symbolic approach, we verify that we do not obtain good predictions. In this case, the Root Mean Square Error values are $RMSE_L = 1.5217$, $RMSE_U = 3.5896$ and $RMSE_M = 2.3139$.

The results obtained show that the relation between the classical variables is not linear, nevertheless the symbolic models allow obtaining reasonable spatial predictions for the burned area of forest fires.

5. Conclusion

In this paper we propose a linear regression model for interval-valued variables which allows analyzing situations where the data have intrinsic variability. This is done by assuming a distribution within the observed intervals. The classical approach that consists in reducing the observations to measures of central tendency is not the most adequate, since much information is lost. The proposed method provides a model relating the interval-valued variables directly and as a whole, as an alternative to considering separate relations for the lower and upper bounds or centers and half ranges. In this paper, the Uniform distribution is considered for modeling the within observations variability. The *ID Model* has however the potential of considering other distributions in the intervals associated with the observations of the interval-valued variables. In this paper we show the *ID Model* developed for the Symmetric Triangular distribution in the intervals. As in most studies of *Symbolic Data Analysis* uniformity is assumed

within the observed intervals, all other necessary concepts would also have to be redefined. Experiments with simulated and real data show that the method has a good performance. For prediction purposes, the user should consider different alternative methods, and retain the one providing more accurate results in the problem at hand.

As future research perspectives, other models and methods in Symbolic Data Analysis based on linear relations between interval-valued variables, such as logistic regression and discriminant analysis, may now be developed using the proposed approach.

Acknowledgements

This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

The authors would like to thank the anonymous reviewers for their interest, careful reading and precious suggestions, which very much helped improve the manuscript.

References

Ahn, J., Peng, M., Park, C., Jeon, Y., 2012. A resampling approach for interval-valued data regression. Statistical Analysis and Data Mining 5 (4), 336–348.

Arroyo, J., Maté, C., 2009. Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting 25 (1), 192–207.

Aznar, J., Guijarro, F., 2007. Estimating regression parameters with imprecise input data in an appraisal context. European Journal of Operational Research 176 (3), 1896–1907.

Bertoluzza, C., Blanco, N. C., Salas, A., 1995. On a new class of distances between fuzzy numbers. Mathware & Soft Computing 2 (2), 71–84.

Billard, L., Diday, E., 2000. Regression analysis for interval-valued data. In: Kiers, H., Rasson, J.-P., Groenen, P., Schader, M. (Eds.), Data Analysis, Classification and Related Methods, Proc. IFCS'00. Springer Berlin Heidelberg, pp. 369–374.

Billard, L., Diday, E., 2002. Symbolic regression analysis. In: Jajuga, K., Sokolowski, A., Bock, H.-H. (Eds.), Classification, Clustering, and Data Analysis, Proc. IFCS '02. Springer Berlin Heidelberg, pp. 281–288.

Billard, L., Diday, E., 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Ltd.

Blanco-Fernández, A., Colubi, A., González-Rodríguez, G., 2013. Towards Advanced Data Analysis by Combining Soft Computing and Statistics. Vol. 285. Springer-Verlag Berlin, Ch. Linear Regression Analysis for Interval-valued Data Based on Set Arithmetic: A Review, pp. 19–31.

Blanco-Fernández, A., Corral, N., González-Rodríguez, G., 2011. Estimation of a flexible simple linear model for interval data based on set arithmetic. CSDA 55 (9), 2568–2578.

Bock, H.-H., Diday, E. (Eds.), 2000. Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag Berlin.

Brito, P., 2014. Symbolic data analysis: another look at the interaction of Data Mining and Statistics. WIREs Data Mining and Knowledge Discovery 4 (4), 281–295.

Brito, P., Duarte Silva, A., 2012. Modelling interval data with Normal and Skew-Normal distributions. Journal of Applied Statistics 39 (1), 3–20.

Cortez, P., Morais, A., 2007. A data mining approach to predict forest fires using meteorological data. In: Neves, J., Santos, M. F., Machado, J. (Eds.), New Trends in Artificial Intelligence, Proc. EPIA 2007. APPIA, pp. 512–523.

Dias, S., 2014. Linear regression with empirical distributions. Ph.D. thesis, Universidade do Porto, Porto, Portugal.

Dias, S., Brito, P., 2015. Linear regression model with histogram-valued variables. Statistical Analysis and Data Mining 8 (2), 75–113.

Gil, M., Lubiano, M., Montenegro, M., López, M., 2002. Least squares fitting of an affine function and strength of association for interval-valued data. Metrika 56 (2), 97–111.

Giordani, P., 2014. Lasso-constrained regression analysis for interval-valued data. Advances in Data Analysis and Classification, 1–15.

González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A., 2007. Least squares estimation of linear regression models for convex compact random sets. Advances in Data Analysis and Classification 1 (1), 67–81.

Hofer, V., 2014. Adapting a classification rule to local and global shift when only unlabelled data are available. European Journal of Operational Research.

Irpino, A., Verde, R., 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Batagelj, V., Bock, H.-H., Ferligoj, A., Ziberna, A. (Eds.), Data Science and Classification, Proc. IFCS'06. Springer Berlin Heidelberg, pp. 185–192.

Irpino, A., Verde, R., 2015. Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification 9 (2), 143–175.

Lima Neto, E., Cordeiro, G., De Carvalho, F., 2011. Bivariate symbolic regression models for interval-valued variables. Journal of Statistical Computation and Simulation 81 (11), 1727–1744.

Lima Neto, E., De Carvalho, F., 2008. Centre and Range method for fitting a linear regression model to symbolic interval data. CSDA 52 (3), 1500–1515.

Lima Neto, E., De Carvalho, F., 2010. Constrained linear regression models for symbolic interval-valued variables. CSDA 54 (2), 333–347.

Mallows, C., 1972. A note on asymptotic joint normality. The Annual of Mathematical Statistics 43 (2), 508–515.

Noirhomme-Fraiture, M., Brito, P., 2011. Far beyond the classical data models: Symbolic Data Analysis. Statistical Analysis and Data Mining 4 (2), 157–170.

Pedrycz, W., 2013. Granular Computing: Analysis and Design of Intelligent Systems. CRC Press/Francis Taylor, Boca Raton.

Pedrycz, W., 2014. Allocation of information granularity in optimization and decision-making models: Towards building the foundations of granular computing. European Journal of Operational Research 232 (1), 137–145.

Winston, W., 1994. Operations Research. Applications and Algorithms, 3rd Edition. Wadsworth.

Supplementary Material

Proof of Proposition 2.3.

Before proving Proposition 2.3 it is necessary to consider two theorems (Winston, 1994) and to define the function to optimize, in matricial form.

Theorem A1. *Consider the optimization problem (5). If* $\boldsymbol{b}^* = (a^*, b^*, v^*)$ *is an optimal solution of this problem,* $\boldsymbol{b}^*$ *must satisfy the constraints of the optimization problem and the Kuhn Tucker conditions:*

- *Constraints:* $-a^* \leq 0$ *and* $-b^* \leq 0$
- *Kuhn Tucker conditions:*
  1. $\frac{\partial f}{\partial a}(\boldsymbol{b}^*) - \lambda = 0$
  2. $\frac{\partial f}{\partial b}(\boldsymbol{b}^*) - \delta = 0$
  3. $\frac{\partial f}{\partial v}(\boldsymbol{b}^*) = 0$
  4. $\lambda a^* = 0$
  5. $\delta b^* = 0$
  6. $\lambda, \delta \geq 0$.

Theorem A2. *Consider the minimization problem (5). If* $f(a, b, v)$, $g_1(a, b, v)$ *and* $g_2(a, b, v)$ *are convex functions, then any vector that satisfies the hypotheses of Theorem A1 is an optimal solution of the optimization problem in (5).*

In the particular case of the minimization problem (5), the optimization function $f(a, b, v)$ may be rewritten in matricial form as:

$$f(a, b, v) = \frac{1}{2}\mathbf{b}^T\mathbf{H_1}\mathbf{b} + \mathbf{w}_1^T\mathbf{b} + K \tag{10}$$

where the matrices and vectors involved are the following:

- $\mathbf{H}_1$ is the hessian matrix, a symmetric matrix of order 3,

$$\mathbf{H}_1 = \begin{bmatrix} \sum_{j=1}^{n} 2c_{X(j)}^2 + \frac{2}{3}r_{X(j)}^2 & \sum_{j=1}^{n} -2c_{X(j)}^2 + \frac{2}{3}r_{X(j)}^2 & \sum_{j=1}^{n} 2c_{X(j)} \\ \sum_{j=1}^{n} -2c_{X(j)}^2 + \frac{2}{3}r_{X(j)}^2 & \sum_{j=1}^{n} 2c_{X(j)}^2 + \frac{2}{3}r_{X(j)}^2 & \sum_{j=1}^{n} -2c_{X(j)} \\ \sum_{j=1}^{n} 2c_{X(j)} & \sum_{j=1}^{n} -2c_{X(j)} & 2n \end{bmatrix};$$

- $\mathbf{w}_1$ is the column vector of independent terms,

$$\mathbf{w}_1 = \begin{bmatrix} \sum_{j=1}^{n} -2c_{Y(j)}c_{X(j)} - \frac{2}{3}r_{Y(j)}r_{X(j)} \\ \sum_{j=1}^{n} 2c_{Y(j)}c_{X(j)} - \frac{2}{3}r_{Y(j)}r_{X(j)} \\ \sum_{j=1}^{n} -2c_{Y(j)} \end{bmatrix};$$

- $\mathbf{b}$ is the column vector of parameters $\mathbf{b} = (a, b, v)^T$;

- K is a real value, $K = \sum_{j=1}^{n} c_{Y(j)}^2 + \frac{1}{3}r_{Y(j)}^2$.

Proof. Consider the optimization problem (5) where:

- the functions $g_1(a, b, v)$ and $g_2(a, b, v)$ that define the non-negative constraints are convex, so that the feasible region of the optimization problem is a convex set;

- $f(a, b, v)$ is a convex function. Consider the matrix $\mathbf{X}$ defined in Expression (7) but now only for one explicative variable. In this particular case, we have:

$$\mathbf{X} = \begin{bmatrix} c_{X(1)} & -c_{X(1)} & 1 \\ \vdots & \vdots & \vdots \\ c_{X(n)} & -c_{X(n)} & 1 \\ \frac{1}{\sqrt{3}}r_{X(1)} & \frac{1}{\sqrt{3}}r_{X(1)} & 0 \\ \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{3}}r_{X(n)} & \frac{1}{\sqrt{3}}r_{X(n)} & 0 \end{bmatrix}$$

  As $\mathbf{H_1} = 2\mathbf{X}^T\mathbf{X}$, the matrix $\mathbf{H_1}$ is positive semi-definite, therefore $f(a, b, v)$ is a convex function.

- the intervals of the explicative variable X are not all degenerate ($\exists j \in \{1,\ldots,n\} : r_{X(j)} \neq 0$) or symmetric ($\exists j \in \{1,\ldots,n\} : \quad c_{X(j)} \neq 0$). In this situation, the columns of $\mathbf{X}$ are linearly independent, so $\mathbf{H_1}$ is positive definite and consequently the function $f(a,b,v)$ is strictly convex, and therefore the optimal solution is unique.

Since the function to optimize is convex and the feasible region too, it may be ensured by Theorem A2, that the vectors that verify the *Kuhn Tucker* conditions are optimal solutions.

Note that in this proof, the following simplifications were considered:

$$\sum_{j=1}^{n} \left(\left(c_{X(j)} - \overline{X} \right) c_{X(j)} \right) = \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X} \right)^2 ; \tag{11}$$

$$\sum_{j=1}^{n} \left(\left(c_{X(j)} - \overline{X} \right) c_{Y(j)} \right) = \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X} \right) \left(c_{Y(j)} - \overline{Y} \right) \tag{12}$$

with $\overline{X} = \dfrac{1}{n} \sum_{j=1}^{n} c_{X(j)}$ and $\overline{Y} = \dfrac{1}{n} \sum_{j=1}^{n} c_{Y(j)}$.

As the optimization problem (5) verifies the conditions of Theorem A1, it is possible to find the expressions of the parameters for the simple linear regression model.

The objective function $f(a,b,v)$ of the minimization problem (5) is,

$$f(a,b,v) = \sum_{j=1}^{n} \left[\left(c_{Y(j)} - (a-b)\, c_{X(j)} - v \right)^2 + \frac{1}{3} \left(r_{Y(j)} - (a+b)\, r_{X(j)} \right)^2 \right]$$

Consider the expressions of the first order partial derivatives of this function:

$$\begin{aligned} \frac{\partial f}{\partial a} &= 2a \sum_{j=1}^{n} (c_{X(j)}^2 + \tfrac{1}{3} r_{X(j)}^2) + 2b \sum_{j=1}^{n} (-c_{X(j)}^2 + \tfrac{1}{3} r_{X(j)}^2) + 2v \sum_{j=1}^{n} c_{X(j)} - \\ &\quad -2 \sum_{j=1}^{n} (c_{Y(j)} c_{X(j)} + \tfrac{1}{3} r_{Y(j)} r_{X(j)}); \end{aligned}$$

$$\begin{aligned} \frac{\partial f}{\partial b} &= 2a \sum_{j=1}^{n} (-c_{X(j)}^2 + \tfrac{1}{3} r_{X(j)}^2) + 2b \sum_{j=1}^{n} (c_{X(j)}^2 + \tfrac{1}{3} r_{X(j)}^2) - 2v \sum_{j=1}^{n} c_{X(j)} + \\ &\quad +2 \sum_{j=1}^{n} (c_{Y(j)} c_{X(j)} - \tfrac{1}{3} r_{Y(j)} r_{X(j)}); \end{aligned}$$

$$\frac{\partial f}{\partial v} = 2(a-b)\sum_{j=1}^{n} c_{X(j)} + 2nv - 2\sum_{j=1}^{n} c_{Y(j)}.$$

Consider the *Kuhn Tucker conditions 1* to *3* in Theorem A1. From condition *3* we have,

$$\frac{\partial f}{\partial v}(\mathbf{b}^*) = 0 \Leftrightarrow v = \overline{Y} - (a^* - b^*)\overline{X}. \tag{13}$$

Substituting Expression (13) in *Kuhn Tucker conditions 1* and *2*, and applying the Expressions (11) and (12), we obtain:

$$\begin{aligned}\frac{\partial f}{\partial a}(\mathbf{b}^*) = \lambda \quad \Leftrightarrow \quad & 2a^* \sum_{j=1}^{n}\left[\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] + 2b^* \sum_{j=1}^{n}\left[-\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] \\ & +2\sum_{j=1}^{n}\left[-\left(c_{X(j)} - \overline{X}\right)\left(c_{Y(j)} - \overline{Y}\right)\right] - 2\sum_{j=1}^{n}\tfrac{1}{3}r_{X(j)}r_{Y(j)} = \lambda;\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial f}{\partial b}(\mathbf{b}^*) \quad = \quad & \delta \Leftrightarrow 2a^* \sum_{j=1}^{n}\left[-\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] + 2b^* \sum_{j=1}^{n}\left[\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] \\ & +2\sum_{j=1}^{n}\left(c_{X(j)} - \overline{X}\right)\left(c_{Y(j)} - \overline{Y}\right) - 2\sum_{j=1}^{n}\tfrac{1}{3}r_{X(j)}r_{Y(j)} = \delta.\end{aligned} \tag{15}$$

From the *Kuhn Tucker conditions 4*: $\lambda a^* = 0$ and *5*: $\delta b^* = 0$, substituting λ and δ by the Expressions (14) and (15), respectively, it is possible to build the system:

$$\begin{cases} (a^*)^2 \sum_{j=1}^{n}\left[\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] + a^*b^* \sum_{j=1}^{n}\left[-\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] - \\ \quad -a^* \sum_{j=1}^{n}\left(c_{X(j)} - \overline{X}\right)\left(c_{Y(j)} - \overline{Y}\right) - a^* \sum_{j=1}^{n}\tfrac{1}{3}r_{X(j)}r_{Y(j)} = 0 \\ a^*b^* \sum_{j=1}^{n}\left[-\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] + (b^*)^2 \sum_{j=1}^{n}\left[\left(c_{X(j)} - \overline{X}\right)^2 + \tfrac{1}{3}r_{X(j)}^2\right] + \\ \quad +b^* \sum_{j=1}^{n}\left(c_{X(j)} - \overline{X}\right)\left(c_{Y(j)} - \overline{Y}\right) - b^* \sum_{j=1}^{n}\tfrac{1}{3}r_{X(j)}r_{Y(j)} = 0 \end{cases}$$

Solving this system, four possible solutions may occur:

Case I: $a^* = 0$ and $b^* = 0$. In this case the parameters are non-negative, as required.

However, the *Kuhn Tucker condition 6* has to be verified, i.e. $\lambda, \delta \geq 0$. Substituting in Expressions (14) and (15), $a^* = 0$ and $b^* = 0$ we have,

$$\begin{cases} -\sum_{j=1}^{n}(c_{X(j)} - \overline{X})(c_{Y(j)} - \overline{Y}) \geq \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)}) \\ \sum_{j=1}^{n}(c_{X(j)} - \overline{X})(c_{Y(j)}\overline{Y}) \geq \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)}) \end{cases}$$

So, $a^* = 0; \quad b^* = 0; \quad v^* = \overline{Y} \quad$ if

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} = \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) = 0.$$

Case II: $a^* = 0$ and $b^* = \dfrac{\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} - \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$.

From the conditions of Theorem A1, we have that $b^* \geq 0$ and $\lambda, \delta \geq 0$. As it is assumed that the intervals of the explicative variable X are not all degenerate ($\exists j \in \{1, \ldots, n\} : r_{X(j)} \neq 0$) and the centers of all these intervals are not all the same, we may ensure that

$$\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 > 0.$$

For the expression that defines b^* to be non-negative it is necessary that,

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \geq \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right).$$

As $\lambda, \delta \geq 0$, substituting the $a^* = 0$ and b^* for the expression above in Expressions (14) and (15), we conclude that

$$\begin{cases} \dfrac{-\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 - \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2} \geq 0 \\ \delta = 0 \end{cases}.$$

So, $a^* = 0$; $\quad b^* = \dfrac{\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} - \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$;

$v^* = \overline{Y} + b^* \overline{X}$ if

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \geq \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)$$

and

$$\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 \leq -\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 .$$

Case III: $a^* = \dfrac{\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} + \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$ and $b^* = 0$.

From the conditions of Theorem A1, we have that $a^* \geq 0$ and $\lambda, \delta \geq 0$. Analogously to case II, we have that the expression that defines a^* is non-negative if,

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \geq -\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right).$$

As $\lambda, \delta \geq 0$, substituting the $b^* = 0$ and a^* for the expression above in Expressions (14) and (15), we conclude that

$$\begin{cases} \lambda = 0 \\ \dfrac{\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right) \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2 - \sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2} \geq 0 \end{cases}$$

So, $a^* = \dfrac{\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} + \sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)}{\sum_{j=1}^{n} \left(c_{X(j)} - \overline{X}\right)^2 + \sum_{j=1}^{n} \frac{1}{3} r_{X(j)}^2}$; $\quad b^* = 0$;

$v^* = \overline{Y} - a^* \overline{X}$ if

$$\sum_{j=1}^{n} \frac{1}{3} r_{X(j)} r_{Y(j)} \geq -\sum_{j=1}^{n} \left(c_{Y(j)} - \overline{Y}\right)\left(c_{X(j)} - \overline{X}\right)$$

and

$$\sum_{j=1}^{n}\left(c_{Y(j)}-\overline{Y}\right)\left(c_{X(j)}-\overline{X}\right)\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}\geq\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}r_{Y(j)}\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}.$$

Case IV:

$$a^{*}=\frac{\sum_{j=1}^{n}\left(c_{Y(j)}-\overline{Y}\right)\left(c_{X(j)}-\overline{X}\right)\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}+\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}r_{Y(j)}\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}}{2\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}}$$

$$b^{*}=\frac{-\sum_{j=1}^{n}\left(c_{Y(j)}-\overline{Y}\right)\left(c_{X(j)}-\overline{X}\right)\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}+\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}r_{Y(j)}\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}}{2\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}}$$

To ensure the conditions $\lambda \geq 0$ and $\delta \geq 0$, the expressions that define a^{*} and b^{*} have to be substituted in Expressions (14) and (15). Performing these substitutions we obtain $\lambda = 0$ and $\delta = 0$.

For the reasons enumerated in the last points, it is ensured that the denominator of the expressions is non-negative.

As a^{*} and b^{*} are non-negative parameters, their expressions verify these conditions only if

$$-\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}r_{Y(j)}\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}\leq\sum_{j=1}^{n}\left(c_{Y(j)}-\overline{Y}\right)\left(c_{X(j)}-\overline{X}\right)\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}$$

and

$$\sum_{j=1}^{n}\left(c_{Y(j)}-\overline{Y}\right)\left(c_{X(j)}-\overline{X}\right)\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}^{2}\leq\sum_{j=1}^{n}\frac{1}{3}r_{X(j)}r_{Y(j)}\sum_{j=1}^{n}\left(c_{X(j)}-\overline{X}\right)^{2}.$$

□

Simulation study - Tables of results

Table 8: Results, in different conditions, of the *ID Model* with $a = 2$, $b = 1$ and $v = -1$ (Section 3.1).

Pattern of variability	*Error level*	n	*Estimated parameters*						*Goodness of fit measures*			
			$\overline{a^*}$ (s)	MSE(a)	$\overline{b^*}$ (s)	MSE(b)	$\overline{v^*}$ (s)	MSE(v)	$\overline{\Omega}$ (s)	$\overline{RMSE}_M$ (s)	$\overline{RMSE}_L$ (s)	$\overline{RMSE}_U$ (s)
Low variability	Level I	10	1.9958 (0.2408)	0.0580	1.0054 (0.2416)	0.0583	-0.8118 (9.6354)	92.7835	0.9838 (0.0053)	0.4271 (0.0732)	0.4586 (0.0949)	0.4534 (0.0933)
		30	2.0033 (0.0924)	0.0085	0.9964 (0.0921)	0.0085	-1.1385 (3.6848)	13.583	0.9833 (0.0028)	0.4661 (0.0396)	0.4982 (0.0530)	0.4970 (0.0512)
		100	2.0011 (0.0576)	0.0033	0.9993 (0.0581)	0.0034	-1.0356 (2.3105)	5.3342	0.9806 (0.0016)	0.5006 (0.0212)	0.5270 (0.0277)	0.5294 (0.0282)
		250	2.0007 (0.0416)	0.0017	0.9996 (0.0414)	0.0017	-1.0232 (1.6567)	2.7426	0.9795 (0.0011)	0.5072 (0.0142)	0.5285 (0.0174)	0.5295 (0.0175)
	Level II	10	1.8192 (1.3178)	1.7676	1.2908 (1.2670)	1.6884	8.4889 (51.2458)	2713.545	0.4115 (0.0971)	4.3285 (0.7012)	4.6161 (0.9161)	4.6246 (0.9109)
		30	1.9406 (0.8099)	0.6588	1.0658 (0.7870)	0.6230	1.4953 (31.8518)	1019.7514	0.3813 (0.0537)	4.6733 (0.3885)	4.9874 (0.5231)	4.9840 (0.5324)
		100	2.0133 (0.5921)	0.3504	0.9926 (0.5874)	0.3447	-1.4118 (23.5757)	555.4258	0.3412 (0.0255)	5.0121 (0.2165)	5.2991 (0.2792)	5.2776 (0.2880)
		250	1.9834 (0.4246)	0.1804	1.0171 (0.4237)	0.1797	-0.3207 (16.9244)	286.6094	0.3256 (0.0156)	5.0712 (0.1413)	5.2941 (0.1758)	5.2858 (0.1755)
High variability	Level I	10	1.9439 (0.7971)	0.6378	1.0560 (0.7945)	0.6337	1.2212 (31.7073)	1009.2841	0.9887 (0.0030)	2.7352 (0.3702)	3.3330 (0.6975)	3.2979 (0.6762)
		30	1.9730 (0.6034)	0.3645	1.0267 (0.6029)	0.3638	0.0859 (24.1892)	585.7105	0.9867 (0.0019)	2.9924 (0.2089)	3.5783 (0.3980)	3.5640 (0.4059)
		100	2.0174 (0.3350)	0.1124	0.9818 (0.3346)	0.1122	-1.7259 (13.4232)	180.5279	0.9863 (0.001)	3.0559 (0.1094)	3.6494 (0.2154)	3.6411 (0.2117)
		250	1.9886 (0.2355)	0.0555	1.0114 (0.2353)	0.0555	-0.5444 (9.3996)	88.4709	0.9862 (0.0006)	3.0809 (0.0706)	3.6627 (0.1358)	3.6568 (0.1356)
	Level II	10	1.6480 (1.5162)	2.4205	1.4053 (1.4975)	2.4044	14.1145 (59.4807)	3762.8658	0.4646 (0.1066)	28.1745 (3.6975)	33.7758 (7.0893)	33.7809 (7.0398)
		30	1.5917 (1.4276)	2.2027	1.4260 (1.4139)	2.1787	15.6896 (56.9147)	3514.5929	0.4253 (0.0592)	30.2256 (2.1111)	35.985 (3.8944)	36.1349 (4.0445)
		100	1.6689 (1.3222)	1.8562	1.3395 (1.3194)	1.8544	12.4257 (53.0652)	2993.3517	0.4209 (0.0314)	30.586 (1.1431)	36.5152 (2.1633)	36.4730 (2.1463)
		250	1.7359 (1.2114)	1.5359	1.2657 (1.2082)	1.5289	9.5994 (48.2692)	2439.9304	0.4172 (0.0199)	30.8052 (0.7197)	36.5458 (1.3842)	36.6275 (1.3552)
Mixed variability	Level I	10	2.000 (0.0877)	0.0077	1.0001 (0.0864)	0.0075	-1.0053 (2.8745)	8.2546	0.9822 (0.0064)	1.6609 (0.3134)	1.6988 (0.3325)	1.6949 (0.3490)
		30	1.9999 (0.0468)	0.0022	1.0000 (0.0469)	0.0022	-0.9942 (1.6390)	2.6837	0.9808 (0.0033)	1.8064 (0.1599)	1.8149 (0.1641)	1.8147 (0.1646)
		100	1.9999 (0.0242)	0.0006	1.0001 (0.0243)	0.0006	-1.0021 (0.7579)	0.5738	0.9760 (0.0022)	1.8592 (0.0864)	1.8684 (0.0886)	1.8672 (0.0891)
		250	1.9996 (0.0164)	0.0003	1.0004 (0.0165)	0.0003	-0.9860 (0.4972)	0.2472	0.9745 (0.0014)	1.8740 (0.0537)	1.8825 (0.0549)	1.8826 (0.0554)
	Level II	10	2.0225 (0.8497)	0.7217	1.0304 (0.7728)	0.5975	-1.1635 (26.9129)	723.6086	0.4084 (0.1161)	16.8083 (2.9515)	17.2367 (3.2065)	17.1714 (3.1907)
		30	1.9888 (0.4987)	0.2486	1.0135 (0.4911)	0.2411	-0.4504 (17.4941)	306.0391	0.3571 (0.0555)	18.0251 (1.5968)	18.1125 (1.6477)	18.1045 (1.6313)
		100	2.0086 (0.2464)	0.0607	0.9915 (0.2458)	0.0604	-1.3072 (7.8400)	61.4978	0.2945 (0.0291)	18.6183 (0.8325)	18.6968 (0.8512)	18.7105 (0.8644)
		250	2.0051 (0.1567)	0.0246	0.9950 (0.1572)	0.0247	-1.1080 (4.7367)	22.4253	0.2795 (0.0175)	18.7173 (0.5250)	18.8009 (0.5354)	18.8062 (0.5445)

Table 9: Results, in different conditions, of the *ID Model* with $a = 2$, $b = 8$ and $v = 3$ (Section 3.1).

Pattern of variability	*Error level*	n	*Estimated parameters*						*Goodness of fit measures*			
			$\overline{a^*}$ (s)	MSE(a)	$\overline{b^*}$ (s)	MSE(b)	$\overline{v^*}$ (s)	MSE(v)	$\overline{\Omega}$ (s)	$\overline{RMSE}_M$ (s)	$\overline{RMSE}_L$ (s)	$\overline{RMSE}_U$ (s)
Low variability	Level I	10	1.9744 (0.7856)	0.6172	8.0230 (0.7892)	0.6227	3.9889 (31.4686)	990.2581	0.9840 (0.0049)	1.4296 (0.2290)	1.5254 (0.3133)	1.5282 (0.3033)
		30	1.9897 (0.3039)	0.0923	8.0069 (0.3060)	0.0936	3.3496 (12.1628)	147.9081	0.9837 (0.0027)	1.5617 (0.1304)	1.6782 (0.1741)	1.6531 (0.1708)
		100	2.0046 (0.1968)	0.0387	7.9959 (0.1965)	0.0386	2.8259 (7.8441)	61.4992	0.9811 (0.0016)	1.6717 (0.0743)	1.7644 (0.0972)	1.7619 (0.0957)
		250	1.9945 (0.1417)	0.0201	8.0057 (0.1423)	0.0203	3.2272 (5.6708)	32.1774	0.9800 (0.0011)	1.6912 (0.0466)	1.7627 (0.0581)	1.7649 (0.0571)
	Level II	10	3.9997 (4.2095)	21.7009	6.4533 (4.4015)	21.7459	-68.1982 (170.5964)	34143.1937	0.4126 (0.0966)	14.5161 (2.3589)	15.5880 (2.9938)	15.4589 (3.1209)
		30	2.3295 (2.3499)	5.6248	7.8146 (2.5356)	6.4574	-7.3426 (97.2957)	9563.9654	0.3931 (0.0558)	15.5431 (1.3173)	16.6159 (1.7503)	16.5627 (1.8119)
		100	2.1335 (1.7408)	3.0453	7.8933 (1.7798)	3.1761	-1.8296 (70.2139)	4948.3897	0.3469 (0.0272)	16.6842 (0.7128)	17.5787 (0.9228)	17.6269 (0.9443)
		250	2.0207 (1.2861)	1.6527	7.977 (1.3048)	1.7013	2.1043 (51.6686)	2667.773	0.3307 (0.0170)	16.9069 (0.4726)	17.6291 (0.5899)	17.6423 (0.5788)
High variability	Level I	10	2.4174 (2.3327)	5.6104	7.5903 (2.3438)	5.6559	-13.5640 (93.0847)	8930.4700	0.9889 (0.0030)	9.0612 (1.2684)	10.8983 (2.2931)	11.0143 (2.3947)
		30	2.1783 (1.7720)	3.1687	7.8177 (1.7758)	3.1836	-4.2384 (71.0896)	5101.0672	0.9868 (0.0019)	9.9365 (0.7348)	11.9266 (1.3310)	11.8448 (1.3177)
		100	1.9844 (1.1238)	1.2618	8.0158 (1.1243)	1.2629	3.5973 (45.0618)	2028.8908	0.9863 (0.0010)	10.2033 (0.3860)	12.165 (0.7281)	12.1928 (0.6992)
		250	1.9982 (0.7601)	0.5772	7.9998 (0.7596)	0.5763	3.0737 (30.3151)	918.0925	0.9862 (0.0007)	10.2698 (0.2466)	12.2089 (0.4541)	12.1878 (0.4605)
	Level II	10	4.4850 (4.9604)	30.7561	5.6798 (4.9797)	30.1559	-92.9741 (197.1509)	48040.6333	0.4617 (0.1056)	94.4150 (12.5131)	113.0991 (24.2285)	112.4600 (23.7080)
		30	4.4098 (4.6370)	27.2870	5.6196 (4.6394)	27.1687	-92.2450 (185.8014)	43559.2486	0.4241 (0.0591)	100.662 (6.9163)	119.8776 (13.5412)	119.8337 (13.4980)
		100	4.2069 (4.3952)	24.1685	5.8154 (4.4684)	24.7185	-85.0827 (177.6946)	39302.3741	0.4201 (0.0318)	102.0593 (3.7356)	121.7733 (7.0518)	121.5274 (7.3249)
		250	3.6169 (3.9147)	17.9238	6.3974 (3.9491)	18.1480	-61.3240 (156.7565)	28685.5906	0.4179 (0.0203)	102.6302 (2.3045)	121.6439 (4.3301)	122.2160 (4.2979)
Mixed variability	Level I	10	1.9954 (0.2891)	0.0835	8.0031 (0.2856)	0.0815	3.3311 (9.5573)	91.3604	0.9848 (0.0051)	5.5460 (0.9764)	5.6462 (1.0752)	5.6910 (1.0768)
		30	1.9920 (0.1633)	0.0267	8.0074 (0.1634)	0.0267	3.2623 (5.7499)	33.0972	0.9835 (0.0028)	6.0181 (0.5258)	6.0455 (0.5436)	6.0461 (0.5350)
		100	2.0037 (0.0801)	0.0064	7.9961 (0.0806)	0.0065	2.9205 (2.5047)	6.2735	0.9803 (0.0017)	6.2168 (0.2780)	6.2436 (0.2847)	6.2465 (0.2878)
		250	2.0004 (0.0563)	0.0032	7.9995 (0.0564)	0.0032	2.9909 (1.7152)	2.9391	0.9789 (0.0012)	6.2440 (0.1818)	6.2706 (0.1881)	6.2749 (0.1848)
	Level II	10	2.4269 (2.3666)	5.7775	7.9477 (2.8481)	8.1064	-4.2571 (86.4180)	7513.2685	0.4407 (0.1248)	55.2868 (10.5882)	56.6721 (11.5276)	56.8957 (11.4630)
		30	2.0618 (1.4820)	2.1979	7.9969 (1.5923)	2.5328	0.8784 (53.9364)	2910.7254	0.3872 (0.0663)	60.2426 (5.4772)	60.5431 (5.6330)	60.5306 (5.5730)
		100	2.0237 (0.7845)	0.6154	7.9769 (0.7864)	0.6184	1.9162 (24.5723)	604.3678	0.3380 (0.0372)	61.8683 (2.8135)	62.1480 (2.8662)	62.1610 (2.9210)
		250	1.9947 (0.5489)	0.3011	8.0054 (0.5491)	0.3012	3.1769 (16.3967)	268.6154	0.3185 (0.0232)	62.5027 (1.7145)	62.7952 (1.7797)	62.7796 (1.7650)

Table 10: Results, in different conditions, of the *ID Model* with $a = 6$, $b = 0$ and $v = 2$ (Section 3.1).

Pattern of variability	*Error level*	n	*Estimated parameters*						*Goodness of fit measures*			
			$\overline{a^*}$ (s)	MSE(a)	$\overline{b^*}$ (s)	MSE(b)	$\overline{v^*}$ (s)	MSE(v)	$\overline{\Omega}$ (s)	$\overline{RMSE}_M$ (s)	$\overline{RMSE}_L$ (s)	$\overline{RMSE}_U$ (s)
Low variability	Level I	10	5.8281 (0.3094)	0.1252	0.1991 (0.2843)	0.1204	9.4333 (11.7926)	194.1803	0.984 (0.0049)	0.8778 (0.1414)	0.9365 (0.1759)	0.9389 (0.1887)
		30	5.9391 (0.1253)	0.0194	0.0787 (0.1067)	0.0176	4.7945 (4.5856)	28.8161	0.9851 (0.0023)	0.9361 (0.0757)	0.9988 (0.1001)	1.0009 (0.1057)
		100	5.9640 (0.0779)	0.0074	0.0456 (0.0682)	0.0067	3.6388 (2.8966)	11.0676	0.9823 (0.0015)	1.0045 (0.0431)	1.0601 (0.0544)	1.0598 (0.0584)
		250	5.9744 (0.0533)	0.0035	0.0327 (0.0482)	0.0034	3.1656 (2.0127)	5.4054	0.9812 (0.0010)	1.0156 (0.0275)	1.0596 (0.0333)	1.0591 (0.0346)
	Level II	10	4.6186 (2.5007)	8.1556	1.6785 (2.2743)	7.9849	63.4324 (94.3885)	12674.2153	0.4116 (0.1004)	8.7809 (1.4507)	9.3692 (1.8559)	9.3729 (1.9114)
		30	5.5001 (1.2298)	1.7607	0.7125 (1.0396)	1.5873	26.3856 (44.8360)	2602.9123	0.4037 (0.0548)	9.4358 (0.7806)	10.1200 (1.0415)	10.0236 (1.0578)
		100	5.6756 (0.7300)	0.6377	0.4307 (0.6435)	0.5991	17.1496 (27.1720)	967.0868	0.3605 (0.0274)	10.0465 (0.4176)	10.6077 (0.5525)	10.5961 (0.5535)
		250	5.7158 (0.5340)	0.3656	0.3405 (0.4845)	0.3505	14.4597 (20.1908)	562.5059	0.3431 (0.0176)	10.1550 (0.2669)	10.5856 (0.3330)	10.5972 (0.3371)
High variability	Level I	10	5.2533 (1.0660)	1.6928	0.7558 (1.0566)	1.6866	31.8842 (42.2034)	2672.4158	0.9883 (0.0031)	5.5875 (0.7578)	6.7037 (1.4363)	6.7304 (1.3595)
		30	5.4787 (0.7325)	0.8078	0.5253 (0.7299)	0.8081	23.0263 (29.3728)	1304.0056	0.9865 (0.0019)	6.0274 (0.4344)	7.1849 (0.8463)	7.178 (0.8145)
		100	5.6924 (0.4304)	0.2796	0.3091 (0.4298)	0.2801	14.3622 (17.2180)	448.9896	0.9863 (0.0010)	6.1236 (0.2209)	7.3080 (0.4367)	7.3026 (0.4204)
		250	5.8084 (0.2795)	0.1147	0.1931 (0.2790)	0.1150	9.6567 (11.1427)	182.6611	0.9862 (0.0006)	6.1697 (0.1443)	7.3287 (0.2712)	7.3300 (0.2817)
	Level II	10	3.4040 (2.9642)	15.517	2.6602 (2.9627)	15.8453	106.3097 (117.0232)	24561.2529	0.4607 (0.1061)	56.4884 (7.6711)	67.0016 (13.9810)	68.2117 (14.6922)
		30	3.6571 (2.8469)	13.5859	2.4070 (2.8403)	13.8527	97.225 (114.5795)	22183.138	0.4284 (0.0613)	60.3564 (4.3267)	71.9955 (7.8317)	72.1277 (7.9684)
		100	4.0286 (2.5207)	10.2343	1.9796 (2.5089)	10.2069	80.9981 (100.7132)	16373.6965	0.4204 (0.0312)	61.1465 (2.2593)	72.7731 (4.2428)	72.9899 (4.4683)
		250	4.3594 (2.2199)	7.6145	1.6548 (2.2169)	7.6480	67.4541 (88.4495)	12099.7297	0.4176 (0.0203)	61.6769 (1.3986)	73.2487 (2.7104)	73.2659 (2.6078)
Mixed variability	Level I	10	5.9930 (0.1589)	0.0253	0.0693 (0.0976)	0.0143	3.2610 (4.2224)	19.4008	0.9885 (0.0038)	3.3896 (0.5893)	3.5214 (0.6577)	3.5000 (0.6571)
		30	5.9929 (0.0886)	0.0079	0.0392 (0.0563)	0.0047	2.7649 (2.5163)	6.9105	0.9877 (0.002)	3.6344 (0.3090)	3.6684 (0.3216)	3.6638 (0.3186)
		100	6.0021 (0.0498)	0.0025	0.0185 (0.0277)	0.0011	2.2481 (1.1952)	1.4885	0.9868 (0.0012)	3.7288 (0.1722)	3.7512 (0.1776)	3.7506 (0.1774)
		250	5.9991 (0.0331)	0.0011	0.0136 (0.0196)	0.0006	2.2241 (0.7850)	0.6659	0.9854 (0.0008)	3.7474 (0.1074)	3.7657 (0.1100)	3.7672 (0.1111)
	Level II	10	5.9723 (1.6399)	2.6873	0.6722 (1.0149)	1.4807	13.6752 (43.4882)	2025.6412	0.4811 (0.1241)	34.2931 (5.9513)	35.5889 (6.6460)	35.4847 (6.5394)
		30	5.9011 (0.8650)	0.7573	0.3908 (0.5481)	0.4529	10.0442 (24.8389)	681.0626	0.4480 (0.0703)	36.3191 (3.2645)	36.6272 (3.4213)	36.6183 (3.3194)
		100	5.9984 (0.4841)	0.2341	0.1953 (0.2743)	0.1133	5.0318 (11.6671)	145.1775	0.4281 (0.0443)	37.2317 (1.7218)	37.4426 (1.7479)	37.448 (1.7858)
		250	5.9999 (0.3291)	0.1082	0.1335 (0.1950)	0.0558	3.8653 (7.9642)	66.8448	0.4034 (0.0287)	37.4815 (1.0899)	37.6624 (1.1198)	37.6833 (1.1194)

Table 11: Results, in different conditions, of the *ID Model* with $a_1 = 2$, $b_1 = 1$, $a_2 = 0.5$, $b_2 = 3$, $a_3 = 1.5$, $b_3 = 1$ and $v = -1$ (Section 3.1).

Pattern of variability	*Error level*	n	*Estimated parameters*											
			$\overline{a_1^*}$ (s)	MSE(a_1)	$\overline{b_1^*}$ (s)	MSE(b_1)	$\overline{a_2^*}$ (s)	MSE(a_2)	$\overline{b_2^*}$ (s)	MSE(b_2)	$\overline{a_3^*}$ (s)	MSE(a_3)	$\overline{b_3^*}$ (s)	MSE(b_3)
Low variability	Level I	10	1.9873 (0.4518)	0.2041	0.9984 (0.4534)	0.2054	0.5738 (0.4980)	0.2532	3.0049 (0.5911)	0.3491	1.4843 (0.3928)	0.1544	0.9984 (0.3736)	0.1396
		30	1.9951 (0.2029)	0.0411	0.9924 (0.1994)	0.0398	0.5044 (0.3249)	0.1055	3.0135 (0.3587)	0.1287	1.4990 (0.2226)	0.0495	0.9924 (0.2241)	0.0504
		100	2.0047 (0.1131)	0.0128	0.9970 (0.1135)	0.0129	0.5005 (0.1613)	0.0260	2.9994 (0.1668)	0.0278	1.4960 (0.1237)	0.0153	0.997 (0.1224)	0.0150
		250	1.9992 (0.0810)	0.0065	0.9999 (0.0798)	0.0064	0.5028 (0.0964)	0.0093	2.9962 (0.0971)	0.0094	1.5025 (0.0713)	0.0051	0.9999 (0.0697)	0.0049
	Level II	10	1.4417 (1.7511)	3.3748	1.0320 (1.5241)	2.3215	1.7273 (2.8339)	9.5294	2.9513 (3.5121)	12.3250	1.7189 (2.3606)	5.6146	1.0320 (2.2744)	5.3968
		30	1.6504 (1.5217)	2.4355	0.8252 (1.1404)	1.3299	1.3117 (1.9717)	4.5428	3.2305 (3.0010)	9.0502	1.7122 (1.8313)	3.3953	0.8252 (1.5840)	2.5971
		100	1.866 (1.0200)	1.0574	0.9544 (0.8841)	0.7829	0.9098 (1.1354)	1.4558	3.0012 (1.7182)	2.9494	1.4883 (1.1622)	1.3495	0.9544 (1.0368)	1.0913
		250	1.9615 (0.7566)	0.5734	0.9972 (0.6878)	0.4726	0.6546 (0.7092)	0.5264	3.0507 (1.0669)	1.1397	1.4608 (0.7547)	0.5705	0.9972 (0.6562)	0.4302
High variability	Level I	10	1.4505 (1.6350)	2.9725	0.9302 (1.3798)	1.9067	1.1367 (1.6882)	3.2525	2.5831 (2.2883)	5.4050	1.7284 (2.1426)	4.6383	0.9302 (2.2020)	5.2768
		30	1.8592 (1.3425)	1.8204	0.9666 (1.0198)	1.0400	0.6821 (0.9699)	0.9729	2.8864 (1.7835)	3.1906	1.5942 (1.4060)	1.9837	0.9666 (1.3151)	1.7497
		100	1.8815 (0.8346)	0.7099	1.0027 (0.7737)	0.5980	0.6530 (0.7210)	0.5427	2.9920 (0.9866)	0.9725	1.5263 (0.9976)	0.9949	1.0027 (0.8538)	0.7283
		250	1.9798 (0.5608)	0.3146	1.0029 (0.5696)	0.3241	0.5507 (0.4925)	0.2449	2.9946(0.6085)	0.3700	1.4980 (0.5904)	0.3483	1.0029 (0.5763)	0.3323
	Level II	10	0.9794 (2.2653)	6.1683	0.8710 (2.1482)	4.6267	1.7534 (3.6735)	15.0522	2.0358 (3.9030)	16.1476	2.3045 (4.6732)	22.4642	0.8710 (4.6136)	22.8594
		30	0.8893 (2.0019)	5.2373	0.7165 (1.8213)	3.3942	1.4617 (3.0957)	10.4984	1.8802 (3.5180)	13.6179	2.8646 (4.6586)	23.5427	0.7165 (4.4953)	22.6418
		100	1.0368 (1.9756)	4.8268	0.8714 (1.8375)	3.3896	1.4222 (2.7718)	8.5259	2.1092 (3.3489)	11.9975	2.3599 (3.9488)	16.3168	0.8714 (3.9425)	17.0221
		250	1.1869 (1.9021)	4.2756	0.968 (1.7092)	2.9195	1.2531 (2.3406)	6.0400	2.1792 (2.8793)	8.9560	2.0913 (3.2809)	11.1030	0.9680 (3.1708)	11.2915
Mixed variability	Level I	10	1.9773 (0.2828)	0.0804	0.9629 (0.2803)	0.0798	0.5807 (0.5451)	0.3034	2.9720 (0.6038)	0.3650	1.5604 (0.6877)	0.4761	0.9629 (0.6278)	0.3964
		30	1.9957 (0.1081)	0.0117	0.9960 (0.1097)	0.0120	0.5013 (0.2892)	0.0835	3.0038 (0.2869)	0.0822	1.5026 (0.3038)	0.0922	0.9960 (0.3091)	0.0954
		100	1.9966 (0.0889)	0.0079	1.0035 (0.0842)	0.0071	0.5069 (0.1342)	0.0180	2.9959 (0.1362)	0.0185	1.5002 (0.1304)	0.0170	1.0035 (0.1298)	0.0168
		250	1.9993 (0.0563)	0.0032	1.0028 (0.0527)	0.0028	0.5057 (0.1078)	0.0116	2.9926 (0.1065)	0.0114	1.5044 (0.1095)	0.0120	1.0028 (0.1107)	0.0123
	Level II	10	2.0474 (2.3355)	5.4515	1.4200 (2.0369)	4.3212	0.9010 (1.5054)	2.4248	2.5413 (2.4635)	6.2730	1.7274 (2.1937)	4.8592	1.4200 (1.5418)	2.3843
		30	1.8009 (1.6091)	2.6261	0.9287 (1.2111)	1.4703	0.8858 (1.3049)	1.8498	2.7054 (2.2389)	5.0945	1.9305 (2.1158)	4.6572	0.9287 (1.4019)	1.9668
		100	1.6986 (1.1177)	1.3388	0.9543 (1.0156)	1.0324	0.8015 (1.0768)	1.2494	2.9489 (1.6981)	2.8832	1.6305 (1.5321)	2.3620	0.9543 (1.1345)	1.2893
		250	1.8465 (0.7867)	0.6419	0.9258 (0.7165)	0.5184	0.7316 (0.8022)	0.6966	2.8978 (1.1353)	1.2981	1.6809 (1.1057)	1.2541	0.9258 (0.8674)	0.7622

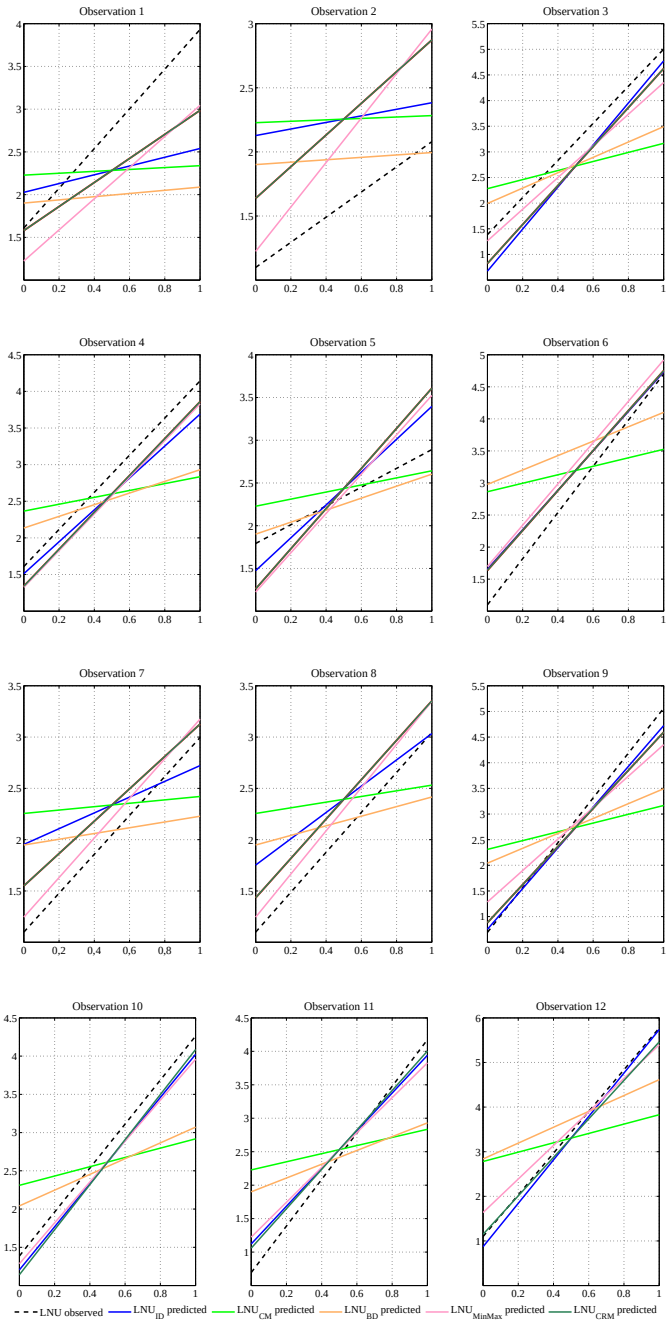
Table 12: Results, in different conditions, of the *ID Model* with $a_1 = 2$, $b_1 = 1$, $a_2 = 0.5$, $b_2 = 3$, $a_3 = 1.5$, $b_3 = 1$ and $v = -1$ (continuation of the *Table 11*).

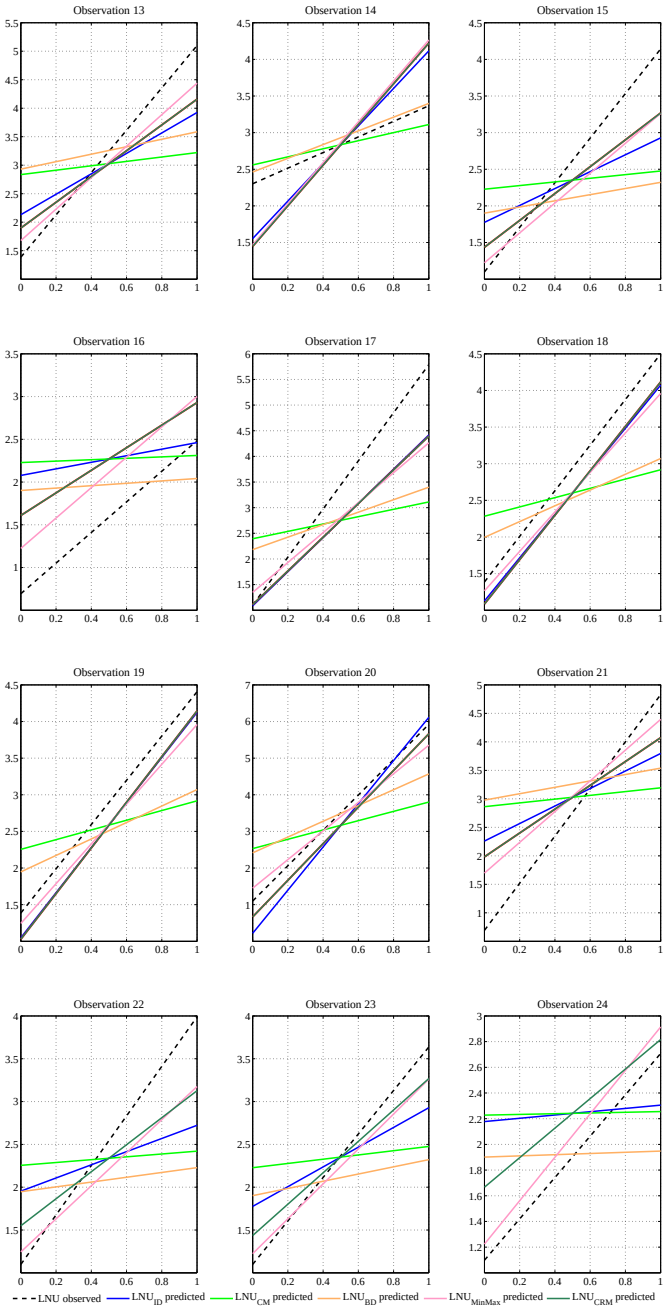
Pattern of variability	*Error level*	n	*Estimated parameter*		*Goodness-of-fit measures*			
			$\overline{v^*}$ (s)	MSE(v)	$\overline{\Omega}$ (s)	$\overline{RMSE}_M$ (s)	$\overline{RMSE}_L$ (s)	$\overline{RMSE}_U$ (s)
Low variability	Level I	10	-1.7823 (26.9084)	723.9481	0.9904 (0.0036)	0.6469 (0.1265)	0.7359 (0.1702)	0.7380 (0.1677)
		30	-0.7590 (13.6114)	185.1425	0.9839 (0.0027)	0.8641 (0.0742)	0.9634 (0.1123)	0.9650 (0.1081)
		100	-1.1333 (6.5151)	42.422	0.9809 (0.0017)	0.9154 (0.0404)	0.9648 (0.0529)	0.9651 (0.0530)
		250	-1.1204 (4.5715)	20.892	0.9801 (0.0011)	0.9258 (0.0264)	0.9654 (0.0314)	0.9641 (0.0319)
	Level II	10	-5.6925 (108.7766)	11842.5294	0.5283 (0.1060)	7.1313 (1.1988)	8.1398 (1.7480)	8.2489 (1.7593)
		30	-5.2635 (88.5241)	7846.8634	0.409 (0.0558)	8.7306 (0.6910)	9.7753 (1.0901)	9.7876 (1.0603)
		100	-3.924 (57.1169)	3267.633	0.3523 (0.0267)	9.1408 (0.3946)	9.6334 (0.5109)	9.6679 (0.5199)
		250	-1.3949 (41.1234)	1689.6024	0.3365 (0.0168)	9.2555 (0.2607)	9.6354 (0.3160)	9.6554 (0.3211)
High variability	Level I	10	-2.8910 (73.6315)	5419.7462	0.9905 (0.0029)	5.5707 (0.8689)	6.8057 (1.4393)	6.8558 (1.4628)
		30	-2.7308 (63.6582)	4051.3082	0.9877 (0.0018)	6.3919 (0.4626)	7.7205 (0.9202)	7.6404 (0.8660)
		100	-1.1470 (36.6992)	1345.5062	0.9867 (0.0010)	6.7295 (0.2554)	8.0298 (0.4690)	8.0232 (0.4782)
		250	-1.5652 (26.0025)	675.7751	0.9866 (0.0006)	6.7441 (0.1610)	8.0588 (0.3082)	8.046 (0.3045)
	Level II	10	-12.7210 (132.6605)	17718.5841	0.4737 (0.1090)	61.667 (8.3808)	73.751 (15.1705)	73.7701 (15.6579)
		30	-13.222 (124.0239)	15515.9344	0.4392 (0.0604)	65.7387 (4.6680)	78.8249 (8.8622)	78.4996 (8.9895)
		100	-7.7275 (113.5824)	12933.3227	0.4266 (0.0316)	67.7501 (2.5506)	81.0198 (4.8083)	80.6425 (4.8974)
		250	-3.6319 (100.6135)	10119.8815	0.4234 (0.0203)	67.7029 (1.6031)	80.8972 (3.0755)	80.7315 (3.0851)
Mixed variability	Level I	10	-4.5079 (36.0235)	1308.6972	0.9936 (0.0023)	3.0175 (0.5736)	3.4427 (0.7839)	3.4204 (0.8027)
		30	-0.9733 (17.5595)	308.0289	0.9916 (0.0015)	3.1782 (0.2837)	3.3886 (0.3657)	3.3939 (0.3649)
		100	-1.1383 (7.4955)	56.146	0.9902 (0.0008)	3.3377 (0.1339)	3.7223 (0.2179)	3.7222 (0.2142)
		250	-1.2755 (6.2929)	39.6373	0.9904 (0.0005)	3.7374 (0.0965)	4.0525 (0.1407)	4.0456 (0.1416)
	Level II	10	-12.7191 (96.75)	9488.5425	0.5064 (0.1183)	43.4851 (8.1285)	46.8738 (10.0865)	47.4534 (10.6710)
		30	-15.5361 (96.7261)	9557.8811	0.3645 (0.0586)	49.9148 (4.4089)	51.5413 (5.0236)	51.513 (5.1361)
		100	-3.5017 (74.2281)	5510.5629	0.3006 (0.0282)	52.9952 (2.3928)	55.6267 (2.9831)	55.5293 (2.9814)
		250	-9.0616 (55.8534)	3181.472	0.3175 (0.0172)	56.2304 (1.5351)	58.3002 (1.9127)	58.4289 (1.8492)

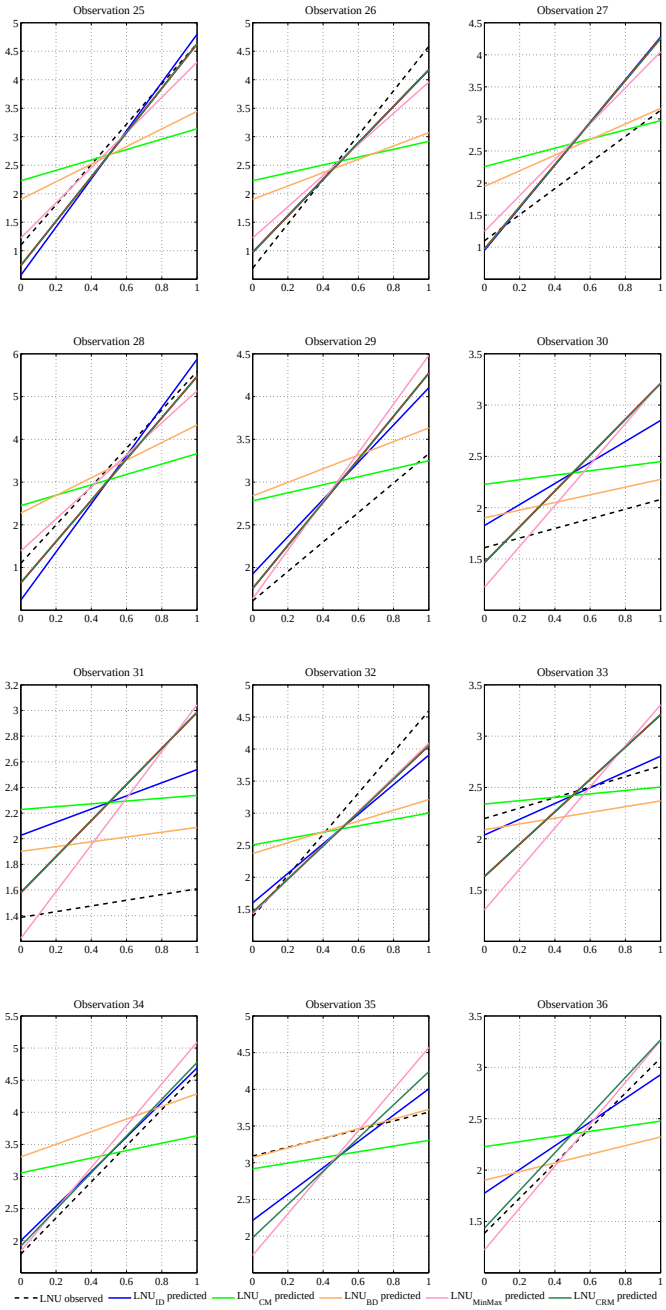
Table 13: Performance of the symbolic linear regression models considering the different cases described in Section 3.2.

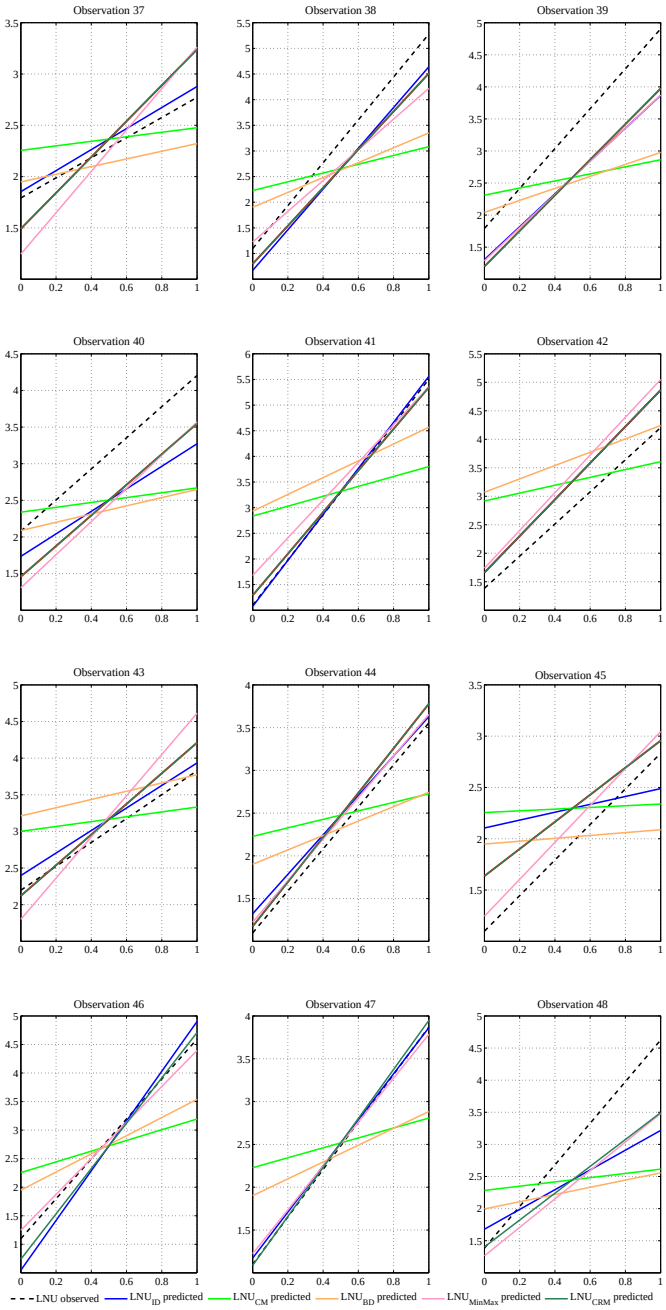
n	Case	$\overline{RMSE_M}(s)$			
		ID Model	ID_T *Model*	*CRM*	*CCRM*
30	*A*	3.0214 (0.0791)	2.9525 (0.0814)	2.8852 (0.0836)	2.8852 (0.0836)
	B	3.1414 (0.0677)	3.0435 (0.0689)	3.0261 (0.0696)	3.0261 (0.0696)
	C	11.8054 (0.3209)	11.7901 (0.3236)	11.7757 (0.2771)	11.7757 (0.2771)
	D	11.9001 (0.3486)	11.8744 (0.3494)	11.8690 (0.3495)	11.8690 (0.3495)
	E	2.9824 (0.1022)	2.9681 (0.1031)	2.9608 (0.1035)	2.9829 (0.1034)
	F	3.2803 (0.0690)	3.1869 (0.0721)	3.1973 (0.0747)	3.2086 (0.0739)
	G	11.8197 (0.3445)	11.8194 (0.3460)	11.8158 (0.3476)	11.9131 (0.3592)
	H	11.8728 (0.3297)	11.8495 (0.3326)	11.8535 (0.3349)	11.8564 (0.3348)
	I	3.6560 (0.1636)	3.4304 (0.1593)	3.1172 (0.1958)	3.1267 (0.1954)
	J	3.4058 (0.1643)	3.2511 (0.1671)	3.1845 (0.2168)	3.1899 (0.2163)
	K	12.5075 (0.5631)	12.5554 (0.6596)	12.8183 (1.0025)	12.8205 (1.0023)
	L	12.3471 (0.7177)	12.3497 (0.7423)	12.4626 (0.8404)	12.4645 (0.8404)
250	*A*	2.9646 (0.0163)	2.8936 (0.0167)	2.8240 (0.0173)	2.8240 (0.0173)
	B	3.0782 (0.0089)	2.9800 (0.0091)	2.9591 (0.0091)	2.9591 (0.0091)
	C	11.5258 (0.0371)	11.5081 (0.0373)	11.4911 (0.0376)	11.4911 (0.0376)
	D	11.5885 (0.0480)	11.5622 (0.0481)	11.5556 (0.0481)	11.5556 (0.0481)
	E	2.9081 (0.0106)	2.8937 (0.0107)	2.8856 (0.0107)	2.9081 (0.0106)
	F	3.2266 (0.0101)	3.1310 (0.0106)	3.1381 (0.0123)	3.1450 (0.0108)
	G	11.5447 (0.0654)	11.5409 (0.0656)	11.5387 (0.0658)	11.5448 (0.0658)
	H	11.5470 (0.0359)	11.5209 (0.0363)	11.5226 (0.0366)	11.5242 (0.0366)
	I	3.7344 (0.1214)	3.3619 (0.1027)	2.9091 (0.0716)	2.9186 (0.0714)
	J	3.2677 (0.0674)	3.0845 (0.0649)	2.9616 (0.0668)	2.9696 (0.0666)
	K	11.9848 (0.1858)	11.9280 (0.1943)	11.8938 (0.2538)	11.8961 (0.2538)
	L	11.6344 (0.2703)	11.5933 (0.2742)	11.5841 (0.2789)	11.5842 (0.2788)

Observed and Predicted quantile functions of the LNU in Section 4.1.




Observation 13
Observation 14
Observation 15
Observation 16
Observation 17
Observation 18
Observation 19
Observation 20
Observation 21
Observation 22
Observation 23
Observation 24
LNU observed
LNU$_{ID}$ predicted
LNU$_{CM}$ predicted
LNU$_{BD}$ predicted
LNU$_{MinMax}$ predicted
LNU$_{CRM}$ predicted


Observation 25
Observation 26
Observation 27
Observation 28
Observation 29
Observation 30
Observation 31
Observation 32
Observation 33
Observation 34
Observation 35
Observation 36
LNU observed
LNU$_{ID}$ predicted
LNU$_{CM}$ predicted
LNU$_{BD}$ predicted
LNU$_{MinMax}$ predicted
LNU$_{CRM}$ predicted


Observation 37
Observation 38
Observation 39
Observation 40
Observation 41
Observation 42
Observation 43
Observation 44
Observation 45
Observation 46
Observation 47
Observation 48
LNU observed
LNU_ID predicted
LNU_CM predicted
LNU_BD predicted
LNU_MinMax predicted
LNU_CRM predicted

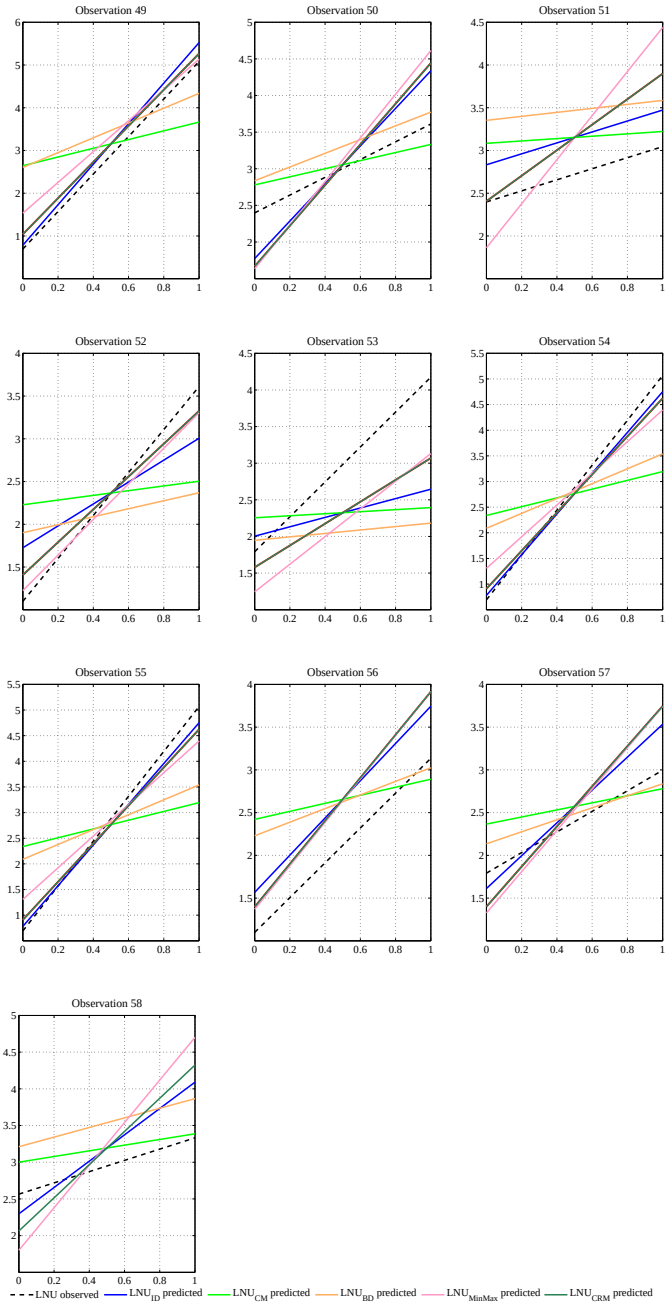


Figure 7: Observed and predicted quantile functions considering all methods presented in Table 3.