Linear Regression Model with Histogram-Valued Variables

Sónia Dias[1]* and Paula Brito[2]

[1]*INESC TEC - INESC Technology and Science and ESTG/IPVC - School of Technology and Management, Polytechnic Institute of Viana do Castelo, Viana do Castelo 4900-347, Portugal*

[2]*INESC TEC - INESC Technology and Science and FEP - Faculty of Economy, University of Porto, Porto 4099-002, Portugal*

Received 6 May 2014; revised 30 January 2015; accepted 14 February 2015
DOI:10.1002/sam.11260
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Histogram-valued variables are a particular kind of variables studied in *Symbolic Data Analysis* where to each entity under analysis corresponds a distribution that may be represented by a histogram or by a quantile function. Linear regression models for this type of data are necessarily more complex than a simple generalization of the classical model: the parameters cannot be negative; still the linear relation between the variables must be allowed to be either direct or inverse. In this work, we propose a new linear regression model for histogram-valued variables that solves this problem, named *Distribution and Symmetric Distribution Regression Model*. To determine the parameters of this model, it is necessary to solve a quadratic optimization problem, subject to non-negativity constraints on the unknowns; the error measure between the predicted and observed distributions uses the Mallows distance. As in classical analysis, the model is associated with a goodness-of-fit measure whose values range between 0 and 1. Using the proposed model, applications with real and simulated data are presented. © 2015 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2015

Keywords: data with variability; linear regression; symbolic data analysis; quantile functions; Mallows distance

1. INTRODUCTION

Classical multivariate statistics studies data tables that summarize observations made on "statistical units" (individuals); each row of the table represents one individual and each of these individuals is characterized by different variables (in columns). The "values" attained by the variables may be real values if the variable represents the measurement of a quantity (quantitative variables) or a category if the variable is qualitative. As an example, let us have classical quantitative variables such as the age, weight, or height of a particular football player. The observations of these data are typically represented in classical data tables. But how can we represent the result of the weight of the football player if we do not know his exact weight? And what if we are interested in studying the age, weight, and height not of one single player but of a football team? In the first situation, the individuals are described by attributes whose associated values are quantitative values that cannot be "measured" with precision. In cases like this, we are in the presence of imprecise data. In the second situation, we are interested in describing one class of individuals. The "best values" attained by the variables that characterize each class are not real values or categories but sets of "values", intervals, or distributions. Even though data with variability or uncertainty may be represented by the same type of elements, the meaning of these elements is different. For example, the interval $[80, 82]$ may mean that the weight of one football player is between 80 and 82 Kg. On the other hand, the interval $[75, 80]$ may represent the weights of all players of a given football team. In the first situation, the interval represents the imprecision of the weight value, whereas in the second situation, the interval considers the variability of weight values in the football team.

In this research, we will focus on situations where variability in data description occurs. The classical solution to analyze these data is to reduce the collection of records associated to each individual or class of individuals to one value, which may be the mean, mode, or maximum/minimum; however, with this option, the variability

Correspondence to: Sónia Dias (sdias@estg.ipvc.pt)

© 2015 Wiley Periodicals, Inc.

Table 1. Data for three healthcare centers.

Healthcare centers	Age	Waiting time for consultation (minutes)
A	[25, 53]	{[0, 15) , 0; [15, 30) , 0.25; [30, 45) , 0.5; [45, 60) , 0; ≥ 60, 0.25}
B	[33, 68]	{[0, 15) , 0.25; [15, 30) , 0.25; [30, 45) , 0.25; [45, 60) , 0.25; ≥ 60, 0}
C	[20, 75]	{[0, 15) , 0.33; [15, 30) , 0; [30, 45) , 0.33; [45, 60) , 0; ≥ 60, 0.33}

across the records is lost. As an alternative to applying the classical analysis to these kinds of data, Diday [1] introduced *Symbolic Data Analysis*, where the term *symbolic data* refers precisely to data with variability. To understand the concept of symbolic data, it is important to assess where variability comes from. The variability of the data might emerge due to the aggregation of observations [2] that can be contemporary, if the records are collected in the same temporal instant or the temporal instant is not relevant, and temporal if the time is the aggregation criterion, and if the records are grouped along one unit of time, for example, one day. In both situations, the initial data or microdata are organized in classical data tables where each individual, termed first-level unit, is described by classical variables. Depending on the type of aggregation, the construction of the symbolic data table is different. When the aggregation is temporal, the entities under analysis are the original first-level units, now characterized by sets of values originating from the records collected over a unit of time. In situations where the aggregation is contemporary, the entities - higher-level units - are classes of individuals (sets of first-level units) grouped according to specific characteristics. In this situation, the variables describing both the higher-level and the respective first-level units are the same; however, the "values" that the variables take for each higher-level unit are now sets of values or functions obtained from the respective first-level units.

Similar to the classical case, symbolic variables can also be classified as quantitative or qualitative. For quantitative symbolic variables, each unit is allowed to take a single value (single-valued variables); a finite set of values (multi-valued variables); an interval (interval-valued variables); or a mapping that can be a probability/frequency/weight distribution (modal-valued variables). In this paper, we will be dealing with a particular type of modal-valued variables, the *histogram-valued variables*.

As an example, consider a symbolic data table containing information about patients (adults) attending healthcare centers, during a fixed period of time. In healthcare center A, the age of patients ranged from 25 to 53 years; in healthcare center B, it ranged from 33 to 68 years; and in healthcare center C, the age of patients ranged from 20 to 75 years, so that the age is an interval-valued variable. Now consider another variable that records the waiting time for consultations. In this case, information is recorded with respect to five intervals of time: 0−15 minutes; 15−30 minutes; 30−45 minutes; 45−60 minutes; and > 60 minutes, with associated frequencies of the waiting time in each healthcare center. Each entity is a histogram and the waiting time for consultation is a histogram-valued variable (see Table 1). Notice that in this example, the entities under analysis are the healthcare centers (higher-level units), for each of which we have aggregated information (contemporary aggregation), and NOT the individual patients attending each center (first-level units).

Since the eighties of the last century, Symbolic Data Analysis has achieved considerable development of new statistical and (multivariate) data analysis techniques to analyze multi-valued data (see, for instance, [3-7]). Recently, there has been a growing interest in the analysis of histogram-valued variables, although still more research is developed for interval-valued variables. The methods proposed so far for the former are indeed, frequently, a generalization of their counterparts for the latter. The main definitions of descriptive statistics for one, two, or more histogram-valued variables have already been studied. Billard and Diday [4] defined mean; observed and relative frequency; empirical density function and empirical joint density function. For variance and covariance, two definitions were proposed [3,4,8]; Arroyo [2] defined distribution functions and joint distribution functions.

The first definitions and methods for histogram-valued variables are generally obtained from the application of the classic concepts to the midpoints of the histograms' subintervals, using the respective weights. Furthermore, although the symbolic variables' values are distributions and not real numbers, the results of the application of these concepts are real numbers. For example, the mean of m observations of a histogram-valued variable, proposed by Billard and Diday [4], is a real number. It should be noticed, however, that in recent years, other works have been put forward where the "results" are already distributions. For example, Irpino and Verde [9] present an alternative definition of mean for histogram-valued variables, which produces a mean distribution, that they termed by *barycentric histogram*.

Work with histogram-valued variables has been recently reported in different domains, such as Principal Component Analysis [10,11]; Cluster Analysis [9,12,13]; Time series [14]; and Linear Regression [8,15].

The first linear regression model for histogram-valued variables was a generalization of the first model proposed for interval-valued variables by Billard and Diday [3,16]. Other models have also been proposed for interval-valued variables [17,18]; however, these models present some limitations: firstly, they are based on differences between real values and do not appropriately quantify the closeness between intervals; then, the elements predicted by the models may fail to build an interval; the most recent model imposes non-negativity constraints on the coefficients, therefore forcing a direct linear relation. These limitations prevent a generalization of the models to histogram-valued variables, so that alternative models are being developed (see, e.g., [15,19]). Our goal is to propose a linear regression model for histogram-valued variables allowing predicting distributions from other distributions, without forcing the linear relation to be direct.

The development of nondescriptive methods in Symbolic Data Analysis is still an open research topic for almost all kinds of symbolic variables. Notice, however, papers are recently being published proposing probabilistic models for interval-valued variables [20,21].

The remaining of the paper is organized as follows. Section 2 introduces histogram-valued variables and presents a short study about the space of the quantile functions. In Section 3, the problem of defining a linear regression model for histogram-valued variables is addressed. A model and a respective goodness-of-fit measure are also proposed. Section 4 reports results of a simulation study and two examples that illustrate the application of the model. Finally, Section 5 concludes the paper, pointing out directions for future research.

2. SYMBOLIC DATA ANALYSIS: HISTOGRAM DATA

2.1. Histogram-valued Variables

According to the formal definition presented in Chapter 3 of the book [5], a symbolic variable may be defined as follows:

DEFINITION 1: A symbolic variable Y is a mapping

$$\begin{aligned} Y: \quad & E \rightarrow \mathbb{B} \\ & j \mapsto Y(j) = \xi_j \end{aligned}$$

defined on a set E of statistical entities.

We have $\Omega = E = \{1, 2, \ldots, m\}$ when the individuals are first-level units or $E = \{C_1, C_2, \ldots\}$ with $C_j \subseteq \Omega$ when the individuals are higher-level units (classes/concepts or categories). Each unit j in E takes its "values" in $\mathbb{B}$. According to the type of realization of the symbolic variables, the set $\mathbb{B}$ will be: $\mathbb{B} = \mathcal{Y}$ (classical variables); $\mathbb{B} = \{\mathcal{D} : \mathcal{D} \subseteq \mathcal{Y}, \mathcal{D} \neq \emptyset\}$; $\mathbb{B}$ a set of intervals in $\mathcal{Y} \subseteq \mathbb{R}$ or $\mathbb{B}$ a family of distributions on $\mathcal{Y}$.

The histogram-valued variables are a particular case of modal-valued variables [4,5].

DEFINITION 2: When $\mathbb{B}$ is a set of distributions on $\mathcal{Y}$, a particular outcome in modal-valued variables takes the form:

$$Y(j) = \left\{\eta_i, p_i; i = 1, \ldots, n_j\right\}$$

where p_i is a nonnegative measure (weight, probability, relative frequency) associated with $\eta_i \in \mathcal{Y}$ and n_j is the number of η_i taken by Y for each element j; η_i can be finite or countably infinite in number and categorical or quantitative in value.

If the "values" η_i with $i \in \{1, \ldots, n_j\}$ are ordered and disjoint intervals of values in $\mathcal{Y} \subseteq \mathbb{R}$ and $\sum_{i=1}^{n_j} p_{ij} = 1$, the symbolic variable Y is a histogram-valued variable.

Each realization j of the histogram-valued variable may be represented by the histogram

$$\begin{aligned} H_{Y(j)} = \Big\{ & \left[\underline{I}_{Y(j)_1}, \overline{I}_{Y(j)_1}\right), p_{j1}; \left[\underline{I}_{Y(j)_2}, \overline{I}_{Y(j)_2}\right), p_{j2}; \ldots \\ & \left[\underline{I}_{Y(j)_{n_j}}, \overline{I}_{Y(j)_{n_j}}\right], p_{jn_j}\Big\} \end{aligned} \quad (1)$$

where $\underline{I}_{Y(j)_i}$ and $\overline{I}_{Y(j)_i}$ represent the lower and upper bounds of the subinterval i, respectively; p_{ji} is the frequency associated to the subinterval $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i}\right)$ with $i \in \left\{1, 2, \ldots, n_j\right\}$; n_j is the number of subintervals for the j^{th} unit, $j \in \{1, \ldots, m\}$, $\sum_{i=1}^{n_j} p_{ij} = 1$, $\underline{I}_{Y(j)_i} \leq \overline{I}_{Y(j)_i}$ and $\overline{I}_{Y(j)_i} \leq \underline{I}_{Y(j)_{i+1}}$. Furthermore, it is also assumed that within each subinterval $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i}\right)$, the values of the variable Y for each unit $j \in \{1, \ldots, m\}$, are uniformly distributed.

Each realization $Y(j)$, of the histogram-valued variable Y can be represented by the cumulative empirical distribution function, as well as by its inverse, also called quantile function [9]:

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} \underline{I}_{Y(j)_1} + \frac{t}{w_{j1}} a_{Y(j)_1} & if \;\; 0 \leq t < w_{j1} \\ \underline{I}_{Y(j)_2} + \frac{t - w_{j1}}{w_{j2} - w_{j1}} a_{Y(j)_2} & if \;\; w_{j1} \leq t < w_{j2} \\ \vdots & \\ \underline{I}_{Y(j)_{n_j}} + \frac{t - w_{jn_j - 1}}{1 - w_{jn_j - 1}} a_{Y(j)_{n_j}} & if \;\; w_{jn_j - 1} \leq t \leq 1 \end{cases} \quad (2)$$

where $w_{jl} = \begin{cases} 0 & if \quad l = 0 \\ \sum_{h=1}^{l} p_{jh} & if \quad l = 1, \ldots, n_j \end{cases}$ and $a_{Y(j)_i} = \overline{I}_{Y(j)_i} - \underline{I}_{Y(j)_i}$ with $i \in \{1, \ldots, n_j\}$; n_j is the number of subintervals in $Y(j)$.

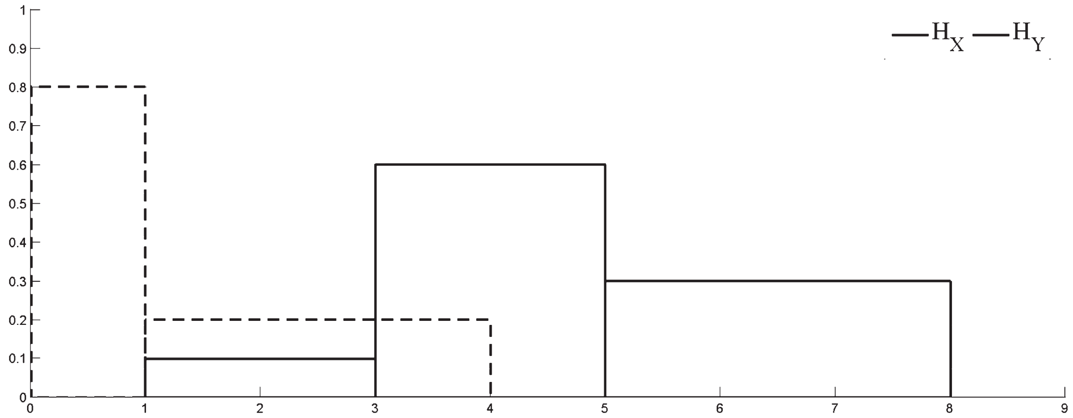

Fig. 1 Representation of the histograms H_X and H_Y in Example 1.

Or, considering the subintervals of the histograms defined by their centers $c_{Y(j)_i}$ and half-ranges $r_{Y(j)_i}$, the representation of $Y(j)$ can be given by

$$H_{Y(j)} = \left\{\left[c_{Y(j)_1} - r_{Y(j)_1}, c_{Y(j)_1} + r_{Y(j)_1}\right), p_{j1}; \ldots; \left[c_{Y(j)n_j} - r_{Y(j)n_j}, c_{Y(j)n_j} + r_{Y(j)n_j}\right], p_{jn_j}\right\} \quad (3)$$

or

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} c_{Y(j)_1} + \left(\frac{2t}{w_{j1}} - 1\right) r_{Y(j)_1} & if \quad 0 \leq t < w_{j1} \\ c_{Y(j)_2} + \left(\frac{2(t-w_{j1})}{w_{j2}-w_{j1}} - 1\right) r_{Y(j)_2} & if \quad w_{j1} \leq t < w_{j2} \\ \vdots & \\ c_{Y(j)n_j} & if \quad w_{jn_j-1} \leq t \leq 1 \\ \; + \left(\frac{2(t-w_{jn_j-1})}{1-w_{jn_j-1}} - 1\right) r_{Y(j)n_j} & \end{cases} \quad (4)$$

Any of these representations of the empirical distribution that each unit takes can be termed *histogram value*. Henceforth, when we use the term distribution, we are referring to an empirical distribution of a continuous variable.

If any of the weights p_{ji} with $i > 1$ is null, the function $\Psi_{Y(j)}$ does not have inverse with domain between 0 and 1. Consequently, the function $\Psi^{-1}_{Y(j)}$ is not continuous and has $n_j - 1$ pieces. In this case, it is not possible to calculate the value of $\Psi^{-1}_{Y(j)}(w_{ji-1})$ but only $\lim_{t \to w^-_{ji-1}} \Psi^{-1}_{Y(j)}(t)$ and $\lim_{t \to w^+_{ji-1}} \Psi^{-1}_{Y(j)}(t)$.

When we work with histogram-valued variables, it is important to note that for different observations, the number of subintervals in the histograms or the pieces in functions have to be the same. In addition, the subintervals of histograms $H_{Y(j)}$ are considered ordered and disjoint, and if this is not the case, it must be possible to rewrite them in the required form [2,22].

EXAMPLE 1: Consider the histograms

$$H_X = \{[1, 3)\,, 0.1; [3, 5)\,, 0.6; [5, 8]\,, 0.3\}$$

and

$$H_Y = \{[0, 1)\,, 0.8; [1, 4]\,, 0.2\}$$

that characterize a unit for the histogram-valued variables X and Y, respectively. These histograms are represented in Fig. 1.

Alternatively, these histograms can be represented by their quantile functions (see Fig. 2):

$$\Psi^{-1}_X(t) = \begin{cases} 1 + \frac{t}{0.1} \times 2 & if \quad 0 \leq t < 0.1 \\ 3 + \frac{t-0.1}{0.6} \times 2 & if \quad 0.1 \leq t < 0.7 \\ 5 + \frac{t-0.7}{0.3} \times 3 & if \quad 0.7 \leq t \leq 1 \end{cases}$$

$$\Psi^{-1}_Y(t) = \begin{cases} \frac{t}{0.8} & if \quad 0 \leq t < 0.8 \\ 1 + \frac{t-0.8}{0.2} \times 3 & if \quad 0.8 \leq t \leq 1 \end{cases}$$

It is important to bear in mind that in a histogram $\underline{I}_{Y(j)i} \leq \overline{I}_{Y(j)i}$ and $\overline{I}_{Y(j)i} \leq \underline{I}_{Y(j)i+1}$; consequently, the quantile function that represents the empirical distribution is always a nondecreasing function in the domain [0, 1].

Many concepts and methods for histogram-valued variables have been defined using the representation of their realizations in the form of histograms [3,4]. Only in more recent studies have the values of these variables been represented as quantile functions [9,14,23,24]. When the distributions are represented as histograms, the choice of the arithmetic becomes crucial. The complexity of the arithmetics [22,25] that have been proposed so far for

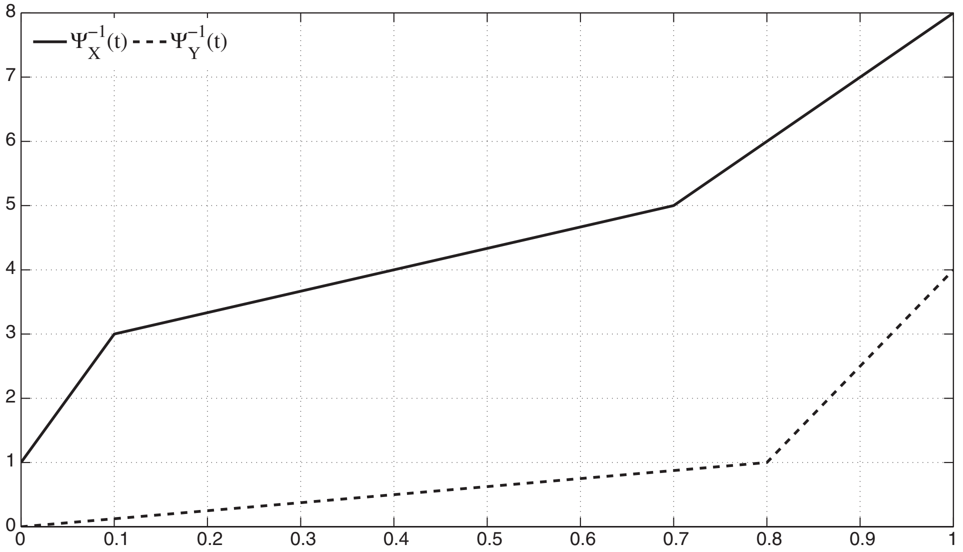

Fig. 2 Representation of the quantile functions Ψ_X^{-1} and Ψ_Y^{-1} in Example 1.

histograms was arguably the reason why the distributions began to be represented as quantile functions. If we represent the distribution that each unit takes on a histogram-valued variable by a quantile function, then operations are simplified because, as quantile functions are piecewise functions, the adequate arithmetic for them is a function arithmetic. In this work, the option is to represent the distributions by quantile functions. However, this representation raises other questions.

To operate with quantile functions, it is necessary to define all functions involved with an equal number of pieces and the domain of each piece has to be the same for all functions. In other words, it is necessary to rewrite all correspondent histograms with the same number of subintervals and the weight associated to each subinterval has to be the same in all units but not in all subintervals of each unit, i.e. the histogram associated to each unit is not necessarily an equiprobable histogram. For this, it may be necessary to apply the procedure defined by Irpino and Verde [9]. In addition, it is important to avoid that the number of subintervals for each histogram becomes "too" large (which could happen by applying the referred process), in which case the distributions that represent the data would be meaningless. To prevent the situation mentioned above and when the microdata are known, we may consider the option of Colombo and Jaarsma [25], which encountered similar problems when operating with histograms and has been considered to be advantageous to work with equiprobable histograms (histograms with equal probability subintervals). In their study, Colombo and Jaarsma [25] refer that the use of equal probability intervals offers many advantages: the distributions are reasonably well approximated by equiprobable histograms, the subintervals into which a distribution is subdivided are small when the frequency is high and large when the frequency is low; operations/combinations of equal frequency subintervals form again equal frequency subintervals.

2.2. The Space of Quantile Functions

Quantile functions are a particular kind of functions. If we consider the set of the functions defined from $\mathbb{R}$ in $\mathbb{R}$, $\mathcal{F}(\mathbb{R},\mathbb{R})$ and the usual operations defined in $\mathcal{F}$: addition $(f+g)(x)=f(x)+g(x), \forall x \in \mathbb{R}$ and product of a function by a real number $(\lambda f)(x)=\lambda f(x), \forall x \in \mathbb{R}$, and $\lambda \in \mathbb{R}$, it follows that $(\mathcal{F},+,.)$ is a vector space. However, if we consider the particular case of the set $\mathcal{E}$ of the quantile functions from $[0,1]$ in $\mathbb{R}$, $\mathcal{E}([0,1],\mathbb{R})$ is not a subspace of the vector space $(\mathcal{F},+,.)$. Analyzing the behavior of these operations, it is possible to understand why the space $(\mathcal{E},+,.)$ with the usual operations does not verify the vector space definition.

The usual addition between two quantile functions is a nondecreasing function; however, when a quantile function is multiplied by a real number, different behaviors may be observed. If the real number is positive, we will have a nondecreasing function, but if the real number is negative, we will obtain a decreasing function that cannot be a quantile function, because quantile functions must always be nondecreasing functions. It is for this reason that $(\mathcal{E},+,.)$ is a semi-vector (or semi-linear) space. The following example illustrates this situation.

EXAMPLE 2: Consider the distribution represented by the quantile function $\Psi_X^{-1}(t)$ presented in *Example 1*. If we

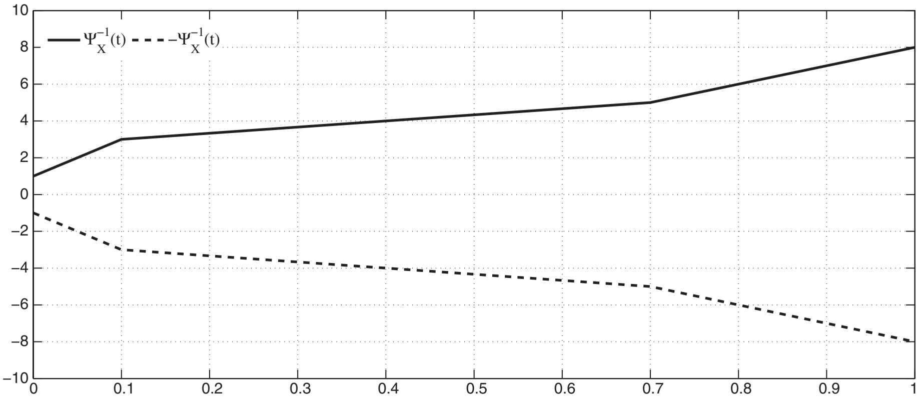

Fig. 3 Representation of the functions $\Psi_X^{-1}(t)$ and $-\Psi_X^{-1}(t)$ in Example 2.

multiply the quantile function $\Psi_X^{-1}(t)$ by the negative real number -1, the resulting function is not a nondecreasing function. The representations in Fig. 3 illustrate this case.

In conclusion, $(\mathcal{E}, +, .)$ is not a vector space. If we have a quantile function $\Psi_X(t)$, the function $-\Psi_X^{-1}(t)$ is not a nondecreasing function and consequently cannot be a quantile function. However, if we consider the distributions represented by histograms and use the histograms arithmetic proposed by Colombo and Jaarsma [25] and afterwards by Case [26], it is possible to obtain a new histogram, that is, the symmetric of the histogram H_X. The histogram $-H_X$ is the symmetric of the histogram H_X if $-H_X$ and H_X are symmetric in relation to the $yy-$axis.

As an example of the situation above, Fig. 4 represents the histogram H_X in Example 1 and the respective symmetric histogram.

It is obviously possible to define the quantile function that represents the distribution of the histogram $-H_X$. This quantile function is $-\Psi_X^{-1}(1-t)$ with $t \in [0, 1]$ and is not the function obtained by multiplying the quantile function $\Psi_X^{-1}(t)$ by -1. Fig. 5 shows that the function $-\Psi_X^{-1}(t)$ in Example 2 is different from the quantile function $-\Psi_X^{-1}(1-t)$ that corresponds to the histogram $-H_X$.

To conclude this section, it is important to underline some conclusions about the function $-\Psi_X^{-1}(1-t)$, $t \in [0, 1]$:

- As it is required for quantile functions, $-\Psi_X^{-1}(1-t)$ is a nondecreasing function;
- $\Psi_X^{-1}(t) - \Psi_X^{-1}(1-t)$ is not a null function, as expected, but is a quantile function with null (symbolic) mean [4];
- the functions $-\Psi_X^{-1}(1-t)$ and $\Psi_X^{-1}(t)$ are linearly independent, provided that $-\Psi_X^{-1}(1-t) \neq \Psi_X^{-1}(t)$;
- $-\Psi_X^{-1}(1-t) = \Psi_X^{-1}(t)$ only when the histogram H_X is symmetric with respect to the $yy-$axis.

3. LINEAR REGRESSION MODEL FOR HISTOGRAM-VALUED VARIABLES

The first linear regression models for histogram-valued variables were proposed by Billard and Diday [3]. These models are an extension of the first models proposed by the authors for interval-valued variables [3,8]. The models proposed by Billard and Diday [3] for histogram-valued variables present some limitations. The main method consists in the fact that those models are a simple adaptation of the classical linear regression model. The estimation parameters are not deduced from the model but are an adaptation of the solution obtained by the Least Squares estimation method for the classical model where the variance and covariance symbolic definitions are applied. Moreover, the process to build the predicted histograms is not clear and when the estimated parameters are negative, we may obtain predictions that are not histograms, because subintervals where the lower bound is greater than the upper bound may occur. The authors do not present a solution to this problem.

An alternative method was proposed by Irpino and Verde [15,19]. This model is defined taking into account the entire distributions that are represented by quantile functions and relies on the exploitation of the properties of a decomposition of the Mallows distance [27] (that the authors name Wasserstein distance). Using a particular decomposition of this distance, the authors propose the Least Squares method where the quantile functions of the predictors may be obtained by a linear combination of the averages and the centered quantile functions of the explicative distributions. Because the space of the quantile functions is not a vector space, non-negative constraints are

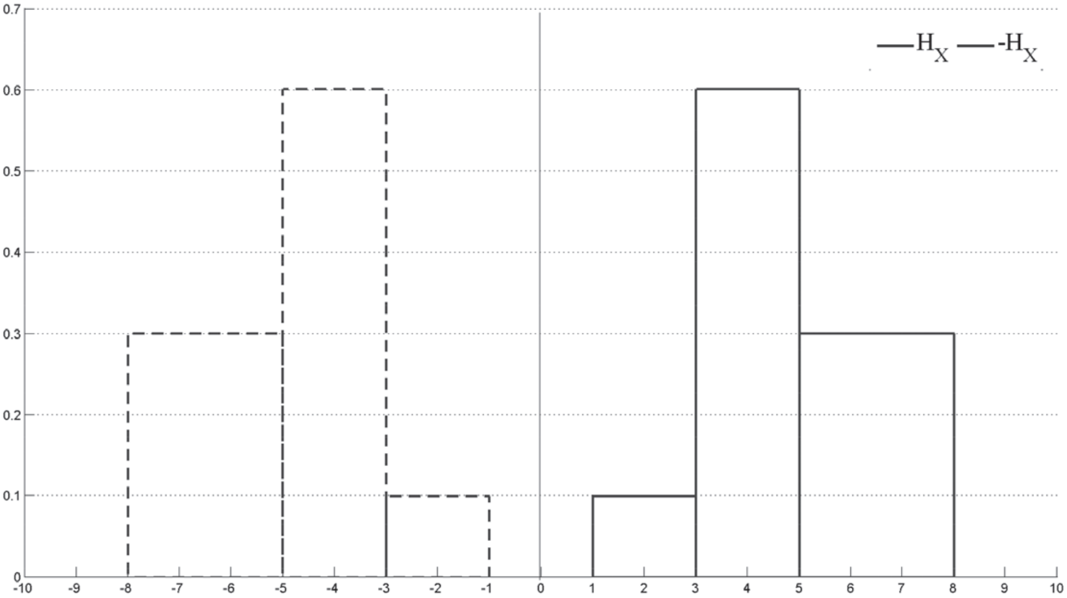

Fig. 4 Representation of the histogram H_X in Example 1 and the respective symmetric histogram $-H_X$.

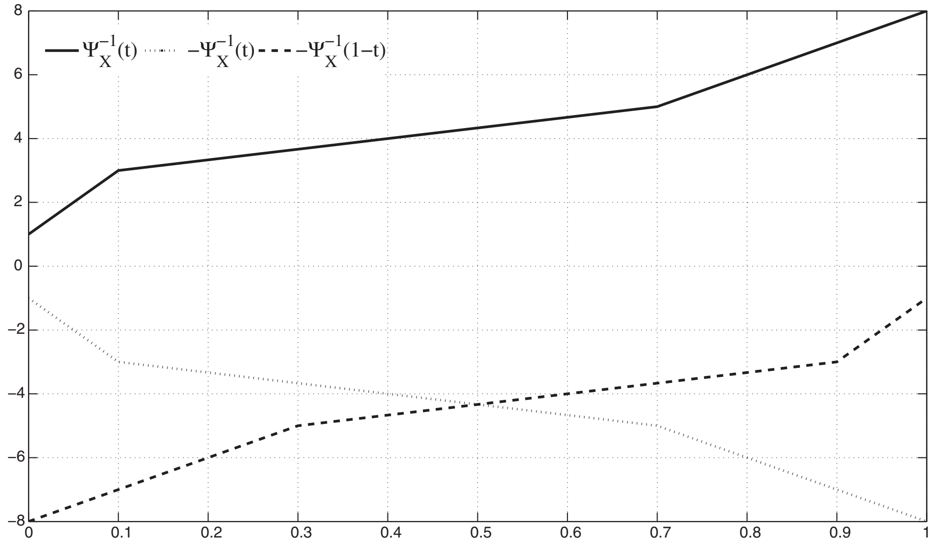

Fig. 5 Representation of the functions $\Psi_X^{-1}(t)$, $-\Psi_X^{-1}(t)$, and $-\Psi_X^{-1}(1-t)$, in Example 2.

imposed to the parameters of the model associated with the centered quantile functions. From the model of Irpino and Verde, a goodness-of-fit measure, named "Pseudo-R^2", was deduced.

The main goal in this work is to propose an alternative linear regression model for histogram-valued variables. More precisely, to provide a linear regression model that considers data with variability and allows predicting histogram values, without forcing a direct linear relation. To address this latter point, it is necessary to solve the problem raised by the semi-linearity of the space of the quantile functions. Moreover, it is important to underline that the methods used to find the parameters of the model are simple and that it is possible to deduce a goodness-of-fit measure, analogous to what happens in classical linear regression.

3.1. Error Measure

In classical linear regression, to quantify the error between the observed values y_j and the predicted values $\widehat{y}_j$, the difference between two real numbers, $e_j = y_j - \widehat{y}_j$, is used. In this case, the model to estimate the values $\widehat{y}_j$ minimizes the quantity $\sum_{j=1}^{m}(y_j - \widehat{y}_j)^2$. However, due to the complexity of histogram-valued variables, the error between the observed and predicted distributions requires a different approach.

In their work on forecasting time series, applied to histogram-valued variables, Arroyo and Maté [2,14] also needed to measure the error between the observed and forcasted distributions. Therefore, they sought for a good measure to analyze the similarity between two distributions. Firstly, they considered the possibility of computing the difference between two distributions represented by their respective histograms using histograms arithmetic. However, this option turned out to be of little use. It is not easy to operate with histograms arithmetic and some results are not as expected. This shows that it is not adequate to analyze the similarity between distributions with this concept. The options of those authors were to use dissimilarity measures for distributions and they opted for the Wasserstein and Mallows distances [14,28] to measure the difference between the observed and predicted distributions. The justification for the choice of the Wasserstein and Mallows distance was the fact that they are distances and thus present interesting properties for error measurement: positive definiteness, symmetry, and triangle inequality condition. On the other hand, for Arroyo and Maté [2,14], the Mallows distance is the one that better adjusts to the concept of distance as assessed by the human eye. This distance was also used in other works such as those by Irpino and Verde [9], where the Mallows distance was used to determine the *barycentric histogram* and then successfully applied to cluster histogram data. The same authors used this distance in their linear regression model for histogram-valued variables [15,19].

When using the Wasserstein and Mallows distances, the distributions taken by the histogram-valued variables are represented by their quantile functions. These distances are defined as follows:

DEFINITION 3: Given two quantile functions $\Psi^{-1}_{X(j)}(t)$ and $\Psi^{-1}_{Y(j)}(t)$ that represent the distributions that the histogram-valued variables X and Y take at unit j, the Wasserstein distance is defined as:

$$D_W(\Psi^{-1}_{X(j)}(t), \Psi^{-1}_{Y(j)}(t)) = \int_0^1 \left| \Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t) \right| dt \tag{5}$$

and the Mallows distance:

$$D_M(\Psi^{-1}_{X(j)}(t), \Psi^{-1}_{Y(j)}(t)) = \sqrt{\int_0^1 (\Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t))^2 dt} \tag{6}$$

It seems therefore appropriate to choose the Wasserstein or the Mallows distance to measure the similarity between the observed and predicted distributions in the linear regression model. Because of the properties of the absolute value function, we choose to define the error measure between two distributions with the Mallows distance.

DEFINITION 4: Consider, for each unit j, $\Psi^{-1}_{Y(j)}(t)$ the quantile function of the observed distribution $Y(j)$ and $\Psi^{-1}_{\widehat{Y}(j)}(t)$ the quantile function that represents the predicted distribution $\widehat{Y}(j)$. The error between $Y(j)$ and $\widehat{Y}(j)$ is defined by:

$$SSE(j) = D^2_M(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t)) \tag{7}$$

Irpino and Verde [9] rewrote the Mallows distance using the center and half-range of the subintervals that compose the histograms. According to this result, the total error may be written as follows:

$$\begin{aligned} SSE &= \sum_{j=1}^{m} SSE(j) = \sum_{j=1}^{m} D^2_M(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t)) \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n_j} p_{ji} \left[(c_{Y(j)_i} - c_{\widehat{Y}(j)_i})^2 + \frac{1}{3}(r_{Y(j)_i} - r_{\widehat{Y}(j)_i})^2 \right] \end{aligned} \tag{8}$$

3.2. The DSD Regression Model

The first option to define the functional linear relation between histogram data was to adapt the classical model to these kinds of data. Consider that we want to predict the distributions of histogram-valued variable Y from p histogram-valued variables X_k with $k \in \{1, \ldots, p\}$. At each unit j, $j \in \{1, \ldots, m\}$, the predicted distribution $\widehat{Y}(j)$ would then be obtained as follows:

$$\widehat{Y}(j) = v + a_1 X_1(j) + a_2 X_2(j) + \ldots + a_p X_p(j).$$

As already mentioned, in this work, we chose to represent the distributions by quantile functions. However, when we multiply a quantile function by a negative number, we do not obtain a nondecreasing function. Therefore, it is necessary to impose non-negativity constraints on the parameters of the model. As such, a functional linear relation between the observations of the histogram-valued variables, represented by the respective quantile functions, may be defined as follows:

$$\begin{aligned} \Psi^{-1}_{\widehat{Y}(j)}(t) &= v + a_1 \Psi^{-1}_{X_1(j)}(t) + a_2 \Psi^{-1}_{X_2(j)}(t) \\ &\quad + \ldots + a_p \Psi^{-1}_{X_p(j)}(t) \end{aligned} \tag{9}$$

with $a_k \geq 0$ and $k \in \{1, 2, \ldots, p\}$.

The non-negativity constraints imposed on the coefficients force a direct linear relation, and limitations similar to those present in linear regression models defined for interval-valued variables occur (see, e.g., [17]). Although we did not generalize the model of interval-valued variables to histogram-valued variables, by defining a model that allows predicting a quantile function from other quantile functions, we obtain a model with similar limitations as observed before.

It is not possible to have negative parameters in the previous model. Nevertheless, it is fundamental to allow for the possibility of a direct and an inverse linear relation between the variable Y and the variables X_k. For this reason, our proposal is to include in the linear regression model both the quantile functions $\Psi^{-1}_{X_k(j)}(t)$, that represent the distributions that the histogram-valued variables X_k take for each unit j, and the quantile functions that represent the respective symmetric histograms $-\Psi^{-1}_{X_k(j)}(1-t)$ (see Section 2.2). Therefore, despite non-negativity constraints being imposed on the coefficients of the model, the linear relation may be direct or inverse because both the quantile functions that represent the distributions $X_k(j)$ and the quantile functions that represent the respective symmetric histogram will be in the model proposed next.

DEFINITION 5: Consider the histogram-valued variables $X_1; X_2; \ldots; X_p$. The quantile functions that represent the distribution that these histogram-valued variables take for each unit j are denoted $\Psi^{-1}_{X_1(j)}(t)$, $\Psi^{-1}_{X_2(j)}(t), \ldots,$ $\Psi^{-1}_{X_p(j)}(t)$ and the quantile functions that represent the respective symmetric histograms associated to each unit of the referred variables are $-\Psi^{-1}_{X_1(j)}(1-t), -\Psi^{-1}_{X_2(j)}(1-t),$ $\ldots, -\Psi^{-1}_{X_p(j)}(1-t)$, with $t \in [0, 1]$. Each quantile function $\Psi^{-1}_{Y(j)}$ can be expressed as follows:

$$\Psi^{-1}_{Y(j)}(t) = \Psi^{-1}_{\widehat{Y}(j)}(t) + e_j(t).$$

where $\Psi^{-1}_{\widehat{Y}(j)}(t)$ is the predicted quantile function for unit j, obtained from

$$\begin{aligned}\Psi^{-1}_{\widehat{Y}(j)}(t) = v + a_1\Psi^{-1}_{X_1(j)}(t) - b_1\Psi^{-1}_{X_1(j)}(1-t)\\ + a_2\Psi^{-1}_{X_2(j)}(t) - b_2\Psi^{-1}_{X_2(j)}(1-t)\\ + \ldots + a_p\Psi^{-1}_{X_p(j)}(t) + b_p\Psi^{-1}_{X_p(j)}(1-t).\end{aligned}$$

with $t \in [0, 1]$; $a_k, b_k \geq 0$, $k \in \{1, 2, \ldots, p\}$ and $v \in \mathbb{R}$.

The error, for each unit j, is the piecewise function given by $e_j(t) = \Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)$.

It should be noted that $\Psi^{-1}_{\widehat{Y}(j)}(t)$ is always a quantile function since it is a linear combination of quantile functions where the coefficients are always nonnegative real values.

For each unit j, the predicted distribution $\widehat{Y}(j)$ can be represented by the quantile function $\Psi^{-1}_{\widehat{Y}(j)}$ or by the respective histogram $H_{\widehat{Y}(j)}$. This linear regression model will be named **Distribution and Symmetric Distribution (DSD) Regression Model**.

Consider the particular case of the linear regression model where there is only one explicative histogram-valued variable X. In this case, we can obtain the quantile function $\Psi^{-1}_{Y(j)}(t)$, for each unit j, by the model:

$$\Psi^{-1}_{Y(j)}(t) = v + a\Psi^{-1}_{X(j)}(t) - b\Psi^{-1}_{X(j)}(1-t) + e_j(t) \quad (10)$$

with $a, b \geq 0$, and $v \in \mathbb{R}$.

To define the *DSD Regression Model*, it is necessary to take into account that:

1. For none of the histogram-valued variables, all m observations present a histogram, which is symmetric as it relates the yy-axis, because in this case $\Psi^{-1}_{X(j)}(t)$ and $-\Psi^{-1}_{X(j)}(1-t)$ would be colinear.
2. For all observations of each variable, the histograms are assumed to be defined with the same number n of subintervals, and to each subinterval i of each observation, and for all variables, is associated the same weight p_i, that verifies the condition $p_i = p_{n-i+1}$, $i \in \{1, ..., n\}$.

If the histograms do not follow the conditions referred in 2, it is necessary to apply the process proposed by Irpino and Verde [9]. Using this process, it is possible to rewrite all distributions associated with each histogram-valued variable X_k, $k \in \{1, 2, \ldots, m\}$, the distributions that represent the respective symmetric histograms and the distributions associated with the response variable Y, with the same number of subintervals and weights. When we rewrite the histograms and respective symmetric with the same number of subintervals, the condition $p_i = p_{n-i+1}$, with $i \in \{1, 2, \ldots, n\}$ is verified. To define a linear regression model, we consider also the distributions associated with the response variable but not the distributions that represent the respective symmetric. Because of this, in some situations, the condition $p_i = p_{n-i+1}$ may not occur. When this happens, we consider the symmetric of the histograms that are the observations of the response variable Y but only with the goal of defining the weights of the subintervals such that $p_i = p_{n-i+1}$, $i \in \{1, 2, \ldots, n\}$.

As an alternative to rewriting all distributions using the process of Irpino and Verde [9] and when the microdata are known, we may organize all histograms as equiprobable and use the respective distributions that represent them.

In this work, we consider that all distributions of all variables and the respective symmetric distributions are defined with n subintervals and with the set of cumulative weights $\{0, w_1, \ldots, w_{n-1}, 1\}$.

3.3. Parameters of the DSD Regression Model

In classical statistics, the parameters of the linear regression model are estimated by solving the minimization problem $\sum_{j=1}^{m}(y_j - \widehat{y}_j)^2$, where y_j are the observed values and $\widehat{y}_j$ the predicted values, with $j \in \{1, \ldots, m\}$. To solve this problem, the least squares method is used.

For histogram-valued variables, the parameters of the *DSD Model*, in Definition 5, are estimated by solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns.

DEFINITION 6: Consider $\Psi^{-1}_{\widehat{Y}(j)}(t)$ obtained by the *DSD Model*. The quadratic optimization problem is written as:

$$\textit{Minimize} \qquad SSE = \sum_{j=1}^{m} D_M^2(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t))$$

with $a_k, b_k \geq 0,\ k \in \{1, 2, \ldots, p\}$ and $v \in \mathbb{R}$.

To present more specifically the function to minimize, it is important to define all the quantile functions involved in this expression considering the conditions referred to in Section 3.2. The quantile functions that represent the distributions taken by X_k and the respective symmetric, for a given unit j, are, respectively:

$$\Psi^{-1}_{X_k(j)}(t) = \begin{cases} c_{X_k(j)_1} + \left(\frac{2t}{w_1} - 1\right) r_{X_k(j)_1} & if \quad 0 \leq t < w_1 \\ c_{X_k(j)_2} + \left(\frac{2(t-w_1)}{w_2-w_1} - 1\right) r_{X_k(j)_2} & if \quad w_1 \leq t < w_2 \\ \vdots & \\ c_{X_k(j)_n} & if \quad w_{n-1} \leq t \leq 1 \\ \quad + \left(\frac{2(t-w_{(n-1)})}{1-w_{(n-1)}} - 1\right) r_{X_k(j)_n} & \end{cases} \tag{11}$$

$$\begin{aligned} & - \Psi^{-1}_{X_k(j)}(1-t) \\ & = \begin{cases} - c_{X_k(j)_n} + \left(\frac{2t}{w_1} - 1\right) r_{X_k(j)_n} & if \quad 0 \leq t < w_1 \\ -c_{X_k(j)_{n-1}} & if \quad w_1 \leq t < w_2 \\ \quad + \left(\frac{2(t-w_1)}{w_2-w_1} - 1\right) r_{X_k(j)_{n-1}} & \\ \vdots & \\ - c_{X_k(j)_1} & if \quad w_{n-1} \leq t \leq 1 \\ \quad + \left(\frac{2(t-w_{n-1})}{1-w_{n-1}} - 1\right) r_{X_k(j)_1} & \end{cases} \end{aligned} \tag{12}$$

Similarly, the quantile function that represents the distribution taken by the histogram-valued variable, Y, may be given by expression (4).

According to the *DSD Model*, the quantile function that represents the distribution taken by the predicted histogram-valued variable $\widehat{Y}$, for a given unit j is:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \begin{cases} \sum_{k=1}^{p} \left(a_k c_{X_k(j)_1} - b_k c_{X_k(j)_n}\right) & \\ \quad + v + \left(\frac{2t}{w_1} - 1\right) & if \quad 0 \leq t < w_1 \\ \quad \times \sum_{k=1}^{p} \left(a_k r_{X_k(j)_1} + b_k r_{X_k(j)_n}\right) & \\ \sum_{k=1}^{p} \left(a_k c_{X_k(j)_2} - b_k c_{X_k(j)_{n-1}}\right) & \\ \quad + v + \left(\frac{2(t-w_1)}{w_2-w_1} - 1\right) & if \quad w_1 \leq t < w_2 \\ \quad \times \sum_{k=1}^{p} \left(a_k r_{X_k(j)_2} + b_k r_{X_k(j)_{n-1}}\right) & \\ \vdots & \\ \sum_{k=1}^{p} \left(a_k c_{X_k(j)_n} - b_k c_{X_k(j)_1}\right) & \\ \quad + v + \left(\frac{2(t-w_{n-1})}{1-w_{n-1}} - 1\right) & if \quad w_{n-1} \leq t \leq 1 \\ \quad \times \sum_{k=1}^{p} \left(a_k r_{X_k(j)_n} + b_k r_{X_k(j)_1}\right) & \end{cases} \tag{13}$$

Consider these quantile functions and the Definition 4. The quadratic optimization problem presented in Definition 6 can then be rewritten as follows:

$$\begin{aligned} \textit{Minimize} \quad SSE = & \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \\ & \times \left[\left(c_{Y(j)_i} - \sum_{k=1}^{p} \left(a_k c_{X_k(j)_i} - b_k c_{X_k(j)_{n-i+1}}\right) - v \right)^2 \right. \\ & \left. + \tfrac{1}{3} \left(r_{Y(j)_i} - \sum_{k=1}^{p} \left(a_k r_{X_k(j)_i} + b_k r_{X_k(j)_{n-i+1}}\right) \right)^2 \right] \end{aligned} \tag{14}$$

subject to $a_k, b_k \geq 0,\ k \in \{1, 2, \ldots, p\}$ and $v \in \mathbb{R}$.

The quadratic optimization problem that allows estimating the parameters of the *DSD Model* may be rewritten in matricial form as a constraint quadratic problem or as a constraint least squares problem[1]. In this paper, we will

[1] In practical examples of this work, the optimization problems to estimate the parameters of the *DSD Model* are solved using the *Matlab* function *quadprog* if we treat the problem as a constraint quadratic problem and the *Matlab* function *lsqlin* when we write the problem as a constraint least squares problem.

only consider the matricial form of the *DSD Model* written as a constraint quadratic problem:

$$Minimize \quad SEE = \frac{1}{2}B^T HB + F^T B + C \tag{15}$$

subject to $-a_k, -b_k \leq 0$; $k \in \{1, 2, \ldots, p\}$ and $v \in \mathbb{R}$.

In this case, $H = \left[h_{lq}\right]$ is the hessian matrix, a symmetric matrix of order $2p+1$, with p the number of variables X_k. The elements of the symmetric matrix H are defined as follows:

$$h_{lq} = \begin{cases} \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big(2c_{X_{\frac{l+1}{2}}(j)_i} c_{X_{\frac{q+1}{2}}(j)_i} & if \quad l,q \text{ are odd} \\ \quad + \frac{2}{3} r_{X_{\frac{l+1}{2}}(j)_i} r_{X_{\frac{q+1}{2}}(j)_i} \Big) & \quad \text{and } l,q \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big(2c_{X_{\frac{l}{2}}(j)_{n-i+1}} c_{X_{\frac{q}{2}}(j)_{n-i+1}} & if \quad l,q \text{ are even} \\ \quad + \frac{2}{3} r_{X_{\frac{l}{2}}(j)_{n-i+1}} r_{X_{\frac{q}{2}}(j)_{n-i+1}} \Big) & \quad \text{and } l,q \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big(-2c_{X_{\frac{l}{2}}(j)_{n-i+1}} c_{X_{\frac{q+1}{2}}(j)_i} & if \quad l \text{ is even, } q \text{ is odd} \\ \quad + \frac{2}{3} r_{X_{\frac{l}{2}}(j)_{n-i+1}} r_{X_{\frac{q+1}{2}}(j)_i} \Big) & \quad \text{and } l,q \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} 2p_i c_{X_{\frac{q+1}{2}}(j)_i} & if \quad q \text{ is odd and} \\ & \quad l = 2p+1 \\ \sum_{j=1}^{m}\sum_{i=1}^{n} -2p_i c_{X_{\frac{q}{2}}(j)_{n-i+1}} & if \quad q \text{ is even and} \\ & \quad l = 2p+1 \end{cases}$$

The vector column of independent terms, $F = \left[f_l\right]$ with $2p+1$ rows is given by:

$$f_l = \begin{cases} \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big(-2c_{Y(j)_i} c_{X_{\frac{l+1}{2}}(j)_i} & if \quad l \text{ is odd} \\ \quad - \frac{2}{3} r_{Y(j)_i} r_{X_{\frac{l+1}{2}}(j)_i} \Big) & \quad \text{and } l \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big(2c_{Y(j)_i} c_{X_{\frac{l+1}{2}}(j)_{n-i+1}} & if \quad l \text{ is even} \\ \quad - \frac{2}{3} r_{Y(j)_i} r_{X_{\frac{l+1}{2}}(j)_{n-i+1}} \Big) & \quad \text{and } l \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} -2p_i c_{Y(j)_i} & if \quad l = 2p+1 \end{cases}$$

The elements of the matrices H and F are computed from the first order partial derivatives of the function SEE in (14). These derivatives are presented in Appendix A. Finally, the vector column of the parameters, B, and the real value C, are defined as follows:

$$B = \begin{bmatrix} a_1 & b_1 & a_2 & b_2 & \ldots & a_p & b_p & v \end{bmatrix}^T$$

and

$$C = \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(c^2_{Y(j)_i} + \frac{1}{3} r^2_{Y(j)_i} \right).$$

For each particular situation, it is possible to solve this quadratic optimization problem, subject to non-negativity on the constraints, and find the optimal solution. Consider the optimal solution for this optimization problem,

$$B^* = \begin{bmatrix} a_1^* & b_1^* & a_2^* & b_2^* & \cdots & a_n^* & b_n^* & v^* \end{bmatrix}^T.$$

It is then possible to predict the distributions $\widehat{Y}(j)$, for each $j \in \{1, \ldots, m\}$, considering the obtained matrix B^*. Each predicted distribution may be represented by the quantile function as in (13) or by the respective histogram

$$H_{\widehat{Y}(j)} = \Bigg\{ \Bigg[\sum_{k=1}^{p} \left(a_k^* \underline{I}_{X_k(j)_1} - b_k^* \overline{I}_{X_k(j)_n} \right) + v^*, \\ \sum_{k=1}^{p} \left(a_k^* \overline{I}_{X_k(j)_1} - b_k^* \underline{I}_{X_k(j)_n} \right) + v^* \Bigg], p_1; \ \ldots; \\ \Bigg[\sum_{k=1}^{p} \left(a_k^* \underline{I}_{X_k(j)_n} - b_k^* \overline{I}_{X_k(j)_1} \right) + v^*, \\ \sum_{k=1}^{p} \left(a_k^* \overline{I}_{X_k(j)_n} - b_k^* \underline{I}_{X_k(j)_1} \right) + v^* \Bigg], p_n \Bigg\}$$

Consider the minimization problem defined in (14) or matricially in (15). The optimal solution of the quadratic optimization problem, subject to non-negativity constraints, verifies the Kuhn Tucker conditions [29]. Therefore, the optimal solution B^* for this optimization problem, for all $k \in \{1, \ldots, p\}$ verifies the following conditions:

- $-a_k^*, -b_k^* \leq 0$;
- $\frac{\partial SEE}{\partial a_k}(B^*) \geq 0$; $\frac{\partial SEE}{\partial b_k}(B^*) \geq 0$; $\frac{\partial SEE}{\partial v}(B^*) = 0$;
  $a_k^* \frac{\partial SEE}{\partial a_k}(B^*) = 0$; $b_k^* \frac{\partial SEE}{\partial b_k}(B^*) = 0$;

From the Kuhn Tucker conditions, it is possible to prove some properties associated with the predicted distribution. Some of these are the counterparts of the corresponding properties in classical statistics, and will allow defining a measure to evaluate the goodness-of-fit of the model. Before describing these properties, it is necessary to present two important definitions of the concept of mean for histogram-valued variables.

DEFINITION 7 [4]: Consider the histogram-valued variable Y. For each unit j, with $j \in \{1, \ldots, m\}$, $Y(j)$ may be represented by the histogram defined in (4). The mean of variable Y is defined as follows:

$$\overline{Y} = \frac{1}{m} \sum_{j=1}^{m} \left(\sum_{i=1}^{n_j} c_{Y(j)_i} p_{ji} \right).$$

where n_j is the number of subintervals for the j^{th} unit.

Irpino and Verde [9] defined the *barycentric histogram* as the histogram that is at a minimum distance - in the sense of the Mallows distance - of the m distributions. In this case, a mean distribution is obtained instead of a mean that is a real number.

The quantile function of the *barycentric histogram* is the same as the mean quantile function that is computed from the average of the m quantile functions that represent the m given distributions. The mean quantile function is defined as follows:

DEFINITION 8: Consider the m quantile functions $\Psi^{-1}_{Y(j)}(t)$, $j \in \{1, \ldots, m\}$, all defined with n pieces. The mean quantile function $\overline{\Psi^{-1}_{Y}}(t)$ is the function where each piece is the mean of the corresponding m pieces involved. The function is then,

$$\overline{\Psi^{-1}_{Y}}(t) = \begin{cases} \sum\limits_{j=1}^{m} \dfrac{c_{Y(j)_1}}{m} + \left(\dfrac{2t}{w_1} - 1\right) \dfrac{r_{Y(j)_1}}{m} & if \quad 0 \le t < w_1 \\ \sum\limits_{j=1}^{m} \dfrac{c_{Y(j)_2}}{m} + \left(\dfrac{2(t - w_1)}{w_2 - w_1} - 1\right) \dfrac{r_{Y_{j2}}}{m} & if \quad w_1 \le t < w_2 \\ \vdots & \\ \sum\limits_{j=1}^{m} \dfrac{c_{Y(j)_n}}{m} + \left(\dfrac{2(t - w_{n-1})}{1 - w_{n-1}} - 1\right) \dfrac{r_{Y_{jn}}}{m} & if \quad w_{n-1} \le t \le 1 \end{cases}$$

So, we have $\overline{\Psi^{-1}_{Y}}(t) = \frac{1}{m} \sum_{j=1}^{m} \Psi^{-1}_{Y(j)}(t)$.

These two concepts of mean for histogram-valued variables are related as we can see in the following proposition.

PROPOSITION 1: Considering the mean quantile function $\overline{\Psi^{-1}_{Y}}(t)$ of the histogram-valued variable Y and its mean $\overline{Y}$, we have

$$\overline{Y} = \int_0^1 \overline{\Psi^{-1}_{Y}}(t)dt.$$

This result is due to Irpino and Verde [23] and may easily be proved considering Definitions 7 and 8.

Now, considering the previous results and the Kuhn Tucker conditions, we may prove the following propositions.

PROPOSITION 2: For each unit j, let $\widehat{Y}(j)$ be the distribution predicted by the *DSD Model* and consider the parameters obtained for the optimal solution $B^* = \begin{bmatrix} a_1^* & b_1^* & a_2^* & b_2^* & \cdots & a_n^* & b_n^* & v^* \end{bmatrix}^T$. The mean of the predicted histogram-valued variable $\overline{\widehat{Y}}$ is given by:

$$\overline{\widehat{Y}} = \sum_{k=1}^{p} \left(a_k^* - b_k^*\right) \overline{X_k} + v^*.$$

Proof: Each observation j, of the predicted histogram-valued variable $\widehat{Y}(j)$, can be represented by the quantile function as in (13) considering for parameters the optimal solution B^*, of the quadratic optimization problem in (14). As such, the mean quantile function $\Psi^{-1}_{\widehat{Y}}$ can be calculated by Definition 8. So, applying Proposition 1, we can prove that $\overline{\widehat{Y}} = \sum_{k=1}^{p} \left(a_k^* - b_k^*\right) \overline{X_k} + v^*$. □

When including in the *DSD Model*, both the distribution of the explicative histogram-valued variables and the respective symmetric distributions, the restrictions on the parameters are imposed; however, this does not imply a direct linear relation. In the particular case of single regression (10), we consider that the linear regression is direct if $a > b$ and inverse if $a < b$. To have a better insight of this behavior, it is necessary to consider Proposition 2.

EXAMPLE 3: In a first situation, consider a symbolic dataset where 10 units are described by two symbolic variables: Y the response histogram-valued variable and X the explicative histogram-valued variable. All observations of the histogram-valued variables are rewritten as histograms with six subintervals and for all units, the weights associated to each subinterval i are the same.

In a second situation, the explicative histogram-valued variable is the symmetric of the histogram-valued variable X, denoted $-X$, Y as in the first case.

The scatter plots of both situations are represented in Fig. 6.

Comparing the expressions of the *DSD Models,* in both situations, we can observe that in the second, as expected, the values of the parameters a and b change relatively to the first.

DSD Model - Situation 1:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = -1.95 + 3.56\Psi^{-1}_{X(j)}(t) - 0.41\Psi^{-1}_{X(j)}(1 - t)$$

DSD Model - Situation 2:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = -1.95 + 0.41\Psi^{-1}_{X(j)}(t) - 3.56\Psi^{-1}_{X(j)}(1 - t)$$

Observing the behavior of the scatter plots, it is important to underline that two orientations can be distinguished: 1) The orientation of the subintervals of each histogram, which obviously is always direct (when the histograms are represented by quantile functions, which are nondecreasing functions) and 2) the orientation of each subinterval i for all units j; it is this latter orientation, and consequently the orientation of the mean values of the histograms, that induces the direct or inverse relation between the histogram-valued variables.

In the first situation, $a > b$ so, according to Proposition 2 and having $\overline{X}_k(j) = \sum_{i=1}^{n} c_{X_k(j)_i} p_{ji}$, $\overline{\widehat{Y}}(j) =$

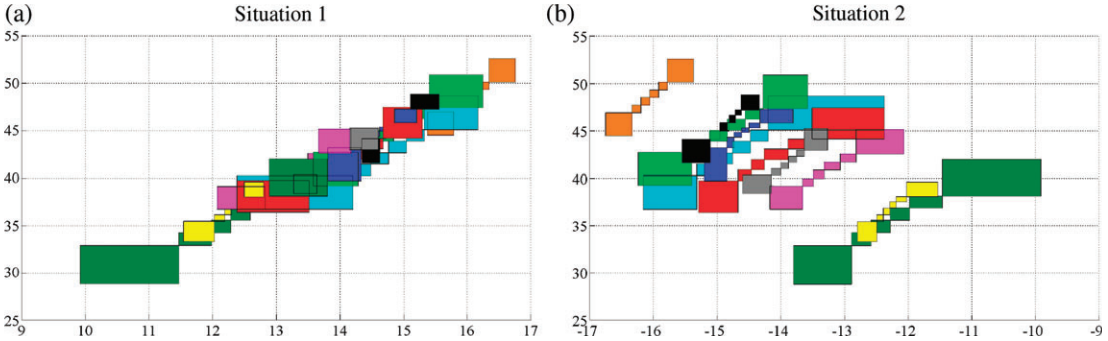

Fig. 6 Scatter plots (projection in $z = 0$) of the observations of the histogram-valued variables X and Y in (a); $-X$ and Y in (b).

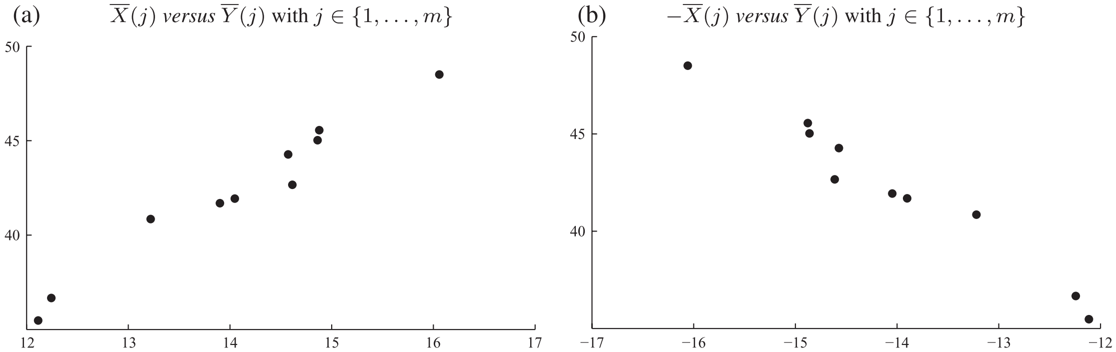

Fig. 7 Scatter plots considering the mean values of the observations of the histogram-valued variables X and Y in (a); $-X$ and Y in (b).

$\sum_{i=1}^{n} c_{\widehat{Y}(j)_i} p_{ji}$, the classical linear relation between the mean values of the histograms, which are the observations of the histogram-valued variables, is a direct linear relation, as illustrated in Fig. 7(a). This behavior means that the relation between histogram-valued variables is classified as direct. On the other hand, we consider that the linear relation between the histogram-valued variables Y and $-X$ is inverse because the parameter a is lower than b. As we can observe in Fig. 7(b), the classical linear relation between the mean values of the histograms $Y(j)$ and $X_k(j)$ is inverse.

From the above propositions, it is still possible to prove other results.

PROPOSITION 3: The mean of the predicted histogram-valued variable $\overline{\widehat{Y}}$ is equal to the mean of the observed histogram-valued variable $\overline{Y}$.

Proof: Consider the function to minimize in (14),

$$SEE = \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \times \left[\left(c_{Y(j)_i} - \sum_{k=1}^{p} (a_k c_{X_{k(j)i}} - \beta_k c_{X_{k(j)n-i+1}}) - v \right)^2 + \tfrac{1}{3} \left(r_{Y(j)_i} - \sum_{k=1}^{p} (a_k r_{X_{k(j)i}} + b_k r_{X_{k(j)n-i+1}}) \right)^2 \right]$$

For the optimal solution B^*, we have $\frac{\partial SEE}{\partial v}(B^*) = 0$. Consequently,

$$2 \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \left(\sum_{k=1}^{p} a_k^* c_{X_{k(j)i}} \right) - 2 \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \times \left(\sum_{k=1}^{p} b_k^* c_{X_{k(j)(n-i+1)}} \right) + 2mv^* - 2 \sum_{j=1}^{m} \sum_{i=1}^{n} p_i c_{Y(j)_i} = 0$$

$$\Longleftrightarrow \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \sum_{k=1}^{p} a_k^* \frac{c_{X_{k(j)i}}}{m} - \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \times \sum_{k=1}^{p} b_k^* \frac{c_{X_{k(j)(n-i+1)}}}{m} + v^* = \sum_{j=1}^{m} \sum_{i=1}^{n} p_i \frac{c_{Y(j)_i}}{m}$$

$$\Longleftrightarrow \sum_{k=1}^{p} \left(a_k^* \overline{X_k} - b_k^* \overline{X_k} \right) + v^* = \overline{Y}$$

From Proposition 2, it follows that

$$\overline{\widehat{Y}} = \sum_{k=1}^{p} \left[\left(a_k^* - b_k^* \right) \overline{X_k} \right] + v^*,$$

so $\overline{\widehat{Y}} = \overline{Y}$. □

PROPOSITION 4: For each unit j, the quantile function of the distribution $\widehat{Y}(j)$ predicted by the *DSD Model* can be rewritten as follows:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y} = \sum_{k=1}^{p}\Big[a_k^*\left(\Psi^{-1}_{X_k(j)}(t) - \overline{X_k}\right) + b_k^*\left(-\Psi^{-1}_{X_k(j)}(1-t) + \overline{X_k}\right)\Big].$$

Proof: In Proposition 3, we proved that

$$\overline{Y} = \sum_{k=1}^{p}\left[\left(a_k^* - b_k^*\right)\overline{X_k}\right] + v^* \Longleftrightarrow v^* = \overline{Y} - \sum_{k=1}^{p}\left(a_k^* - b_k^*\right)\overline{X_k}.$$

For the optimal solution B^*, for each unit j, the quantile function predicted by the linear regression model *DSD*, in Definition 5, is given by

$$\Psi_{\widehat{Y}(j)}(t) = \sum_{k=1}^{p}\left(a_k^*\Psi^{-1}_{X_k(j)}(t) - b_k^*\Psi^{-1}_{X_k(j)}(1-t)\right) + v^*$$

which may then be rewritten as

$$\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y} = \sum_{k=1}^{p}\Big[a_k^*\left(\Psi^{-1}_{X_k(j)}(t) - \overline{X_k}\right) + b_k^*\left(-\Psi^{-1}_{X_k(j)}(1-t) + \overline{X_k}\right)\Big]. \qquad \square$$

PROPOSITION 5: For the observed and predicted distributions $Y(j)$ and $\widehat{Y}(j)$, respectively, with $j \in \{1, \ldots, m\}$, of the variable Y, we have

$$\sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)dt = 0.$$

Proof: The proof is given in Appendix B.

3.4. Goodness-of-fit measure

To complete the investigation of the linear regression model for histogram-valued variables, a goodness-of-fit measure remains to be deduced. We define this measure in a similar way as in the classical model for real data.

PROPOSITION 6: The sum of the square of the Mallows distance between each observed distribution j, $j \in \{1, \ldots, m\}$, of the histogram-valued variable Y, and the mean of the histogram-valued variable Y, $\overline{Y}$, can be decomposed as follows:

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

Proof: Consider each observation j of the histogram-valued variable Y, represented by its quantile function $\Psi^{-1}_{Y(j)}(t)$, and the mean of this histogram-valued variable, $\overline{Y}$. We have,

$$\begin{aligned}\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right) &= \sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \overline{Y}\right)^2 dt\\ &= \sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t) + \Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt\\ &= \sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt\\ &\quad + \sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt\\ &\quad + 2\sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)dt\end{aligned}$$

From Proposition 5 we have,

$$\sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)dt = 0.$$

So, we may write

$$\begin{aligned}&\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)\\ &\quad = \sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt\\ &\qquad + \sum_{j=1}^{m}\int_0^1\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt. \qquad \square\end{aligned}$$

Therefore, similar to the classical model, it is possible to define the goodness-of-fit measure of the *DSD Model*.

DEFINITION 9: Consider the observed and predicted distributions of the histogram-valued variable Y and $\widehat{Y}$ represented, respectively, by their quantile functions $\Psi_{Y(j)}(t)$ and $\Psi^{-1}_{\widehat{Y}(j)}(t)$, and the mean of the histogram-valued variable Y, $\overline{Y}$. The goodness-of-fit measure is given by

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

In classical linear regression, the coefficient of determination R^2 ranges from 0 to 1. In this case, the goodness-of-fit measure, Ω, also ranges from 0 to 1.

PROPOSITION 7: The goodness-of-fit measure Ω ranges from 0 to 1.

Proof: Consider the goodness-of-fit measure $\Omega = \frac{\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right)}{\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)}$. This measure is nonnegative. So, $\Omega \geq 0$.

From Proposition 6, we have

$$\begin{aligned}
&\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right) \\
&\quad = \sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt \\
&\qquad + \sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt \\
&\quad \Longleftrightarrow 1 = \frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt}{\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)} \\
&\qquad + \frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt}{\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)} \\
&\quad \Longleftrightarrow \Omega = 1 - \frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt}{\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)}
\end{aligned}$$

Since the term $\frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt}{\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)}$ is nonnegative, the value of Ω is always less than or equal to 1. So, we have $0 \leq \Omega \leq 1$.

Let us now analyze the extreme situations.

Suppose $\Omega = 0$. In this case,

$$\begin{aligned}
&\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right) = 0 \\
&\quad \Longleftrightarrow \sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt = 0.
\end{aligned}$$

So, for all $j \in \{1, \ldots, m\}$, we have

$$\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y} = 0 \Longleftrightarrow \Psi^{-1}_{\widehat{Y}(j)}(t) = \overline{Y}.$$

In this case, the predicted function for all observations j is a constant function.

Suppose now that $\Omega = 1$. In this case,

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

From the decomposition obtained in Proposition 6, we have,

$$\begin{aligned}
&\sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right) \\
&\quad = \sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right) + \sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \Psi^{-1}_{Y(j)}(t)\right) \\
&\quad \Longleftrightarrow \sum_{j=1}^{m} D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \Psi^{-1}_{Y(j)}(t)\right) = 0.
\end{aligned}$$

So, for all $j \in \{1, \ldots, m\}$,

$$\begin{aligned}
&D_M^2\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \Psi^{-1}_{Y(j)}(t)\right) = 0 \\
&\quad \Longleftrightarrow \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \Psi^{-1}_{Y(j)}(t)\right)^2 dt = 0 \\
&\quad \Longrightarrow \Psi^{-1}_{\widehat{Y}(j)}(t) = \Psi^{-1}_{Y(j)}(t).
\end{aligned}$$

In this case, for each observation j, the predicted and observed quantile functions are coincident.

In conclusion, $0 \leq \Omega \leq 1$. If $\Omega = 0$, there is no linear relation between the histogram-valued variable Y and the histogram-valued variables X_k. If $\Omega = 1$, the linear relation is perfect, so the relation between the histogram-valued variable Y and histogram-valued variables X_k, with $k \in \{1, \ldots, p\}$, is exactly the relation defined by the linear regression model. □

The goodness-of-fit measure, Ω, deduced from the models is computed with respect to the symbolic mean of the histogram-valued response variable, that is, a real value and not an average distribution, the barycentric histogram. This option is due to the apparent impossibility in obtaining the decomposition of the total sum of squares (Proposition 6), when the barycentric histogram is considered.

4. EXPERIMENTS

To illustrate and analyze the *DSD Model*, we performed a simulation study and applied the method to real datasets.

4.1. Simulation Study

To analyze the behavior of the parameter estimation and the performance of the *DSD Model* in different situations, we performed a simulation study. The first step was to generate the observations of the histogram-valued variables X_k, $k = \{1, \dots, p\}$ and Y, where Y is the variable to be modelized from X_k by the linear relation. Next, the parameters were estimated by the *DSD Model* and goodness-of-fit measures computed, considering symbolic-simulated data tables covering different situations. From these results, it was possible to analyze the behavior of the model and draw some meaningful conclusions.

4.1.1. Building symbolic-simulated data tables

The observations of the explicative and response histogram-valued variables X_k and Y were generated in different ways.

- The observations of each histogram-valued variable X_k are created.
  According to the concept of symbolic variables, to obtain the m observations associated to a histogram-valued variable X_k, we started by simulating 5000 real values corresponding to each unit. These values are then organized in histograms, which represent the empirical distribution for each unit. It was considered that in all observations, the subintervals of each histogram have the same weight (equiprobable) with frequency 0.10. This option is supported by the work of Colombo and Jaarsma [25]. If equiprobable histograms with the same weight distributions in all observations were not considered, we would have obtained a large number of different weights and consequently the subintervals would have very low frequencies. It is possible that histograms are not equiprobable; however, the weight in each subinterval has to be the same in all observations (see Section 2.1).

- The observations of the histogram-valued variable Y are created.
  The histograms that are the observations of the histogram-valued variable Y are obtained in three steps. First, we consider the perfect linear regression, without error, given by

$$\Psi^{-1}_{Y^*(j)}(t) = v + \sum_{k=1}^{p} a_k \Psi^{-1}_{X_k(j)}(t) - \sum_{k=1}^{p} b_k \Psi^{-1}_{X_k(j)}(1-t),$$

for particular values of the parameters. The histogram-valued variables X_k and Y^* are in a perfect linear relation; this is, however, not what is intended to simulate for the symbolic data table. Then, we disturb the perfect linear relation by introducing an error function in the model $\Psi^{-1}_{Y(j)}(t) = \Psi^{-1}_{Y^*(j)}(t) + \varepsilon_j(t)$. The error function is a piecewise linear function (but not necessarily a quantile function) defined by:

$$e_j(t) = \begin{cases} \widetilde{c}(j)_1 + \left(\frac{2t}{w_1} - 1\right)\widetilde{r}(j)_1 & if \quad 0 \leq t < w_1 \\ \widetilde{c}(j)_1 + \widetilde{r}(j)_1 + \widetilde{r}(j)_2 & if \quad w_1 \leq t < w_2 \\ +\left(\frac{2(t-w_1)}{w_2-w_1} - 1\right)\widetilde{r}(j)_2 & \\ \vdots & \\ \widetilde{c}(j)_1 + \widetilde{r}(j)_1 + \sum\limits_{i=2}^{n-1} 2\widetilde{r}(j)_i & if \quad w_{n-1} \leq t \leq 1 \\ +\widetilde{r}(j)_n + \left(\frac{2\left(t-w_{(n-1)}\right)}{1-w_{(n-1)}} - 1\right)\widetilde{r}(j)_n & \end{cases} \tag{16}$$

The values of $\widetilde{c}(j)_1$ and $\widetilde{r}(j)_i$, with $i \in \{1, \dots, n\}$ that compose the error function $e_j(t)$ are randomly selected from intervals with low or high variation depending on whether we want the linear regression between the variables to be better or worse. However, the values of $\widetilde{r}(j)_i$ have a limitation. Each half-range $r_{Y(j)_i}$ in the quantile function $\Psi^{-1}_{Y(j)}(t)$, which results from the perturbation of $\Psi^{-1}_{Y^*(j)}(t)$ by the error function $e_j(t)$, is obtained by $r_{Y(j)_i} = r_{Y^*(j)_i} + \widetilde{r}(j)_i$, for each unit j and subinterval i. As it is not imposed that the error function is a quantile function, the values of $\widetilde{r}(j)_i$ may be negative but cannot be lower than $-r_{Y^*(j)_i}$, else for this unit j and subinterval i, the half-range $r_{Y(j)_i}$ would be negative. The expression of the error function in (16) and the constraints imposed to the values that compose it ensure that when we add the quantile function $\Psi^{-1}_{Y^*(j)}(t)$ with the error function $e_j(t)$, we obtain a quantile function. The disturbance induced in the centers of the subintervals of the histograms is difficult to control because the building of $\widetilde{c}(j)_i$ with $i \in \{2, \ldots, n\}$ is recursive and depends on the values of $\widetilde{c}(j)_1$ and $\widetilde{r}(j)_i$.

4.1.2. Description of the simulation study

To perform the simulation study, symbolic data tables that illustrate different situations were created. In this study, a full factorial design was employed, with the following factors:

- Sample size: $m = 10$; 30; 100; 250.
- Number of explicative histogram-valued variables: $p = 1$ and $p = 3$.
- Parameters of the *DSD Model*.
  - For $p = 1$:
    - **i)** $a = 2$; $b = 1$; $v = -1$; (a and b are close)
    - **ii)** $a = 2$; $b = 8$; $v = 3$; (a is lower than b)
    - **iii)** $a = 8$; $b = 0$; $v = 4$; (a is larger than b)
  - For $p = 3$:
    - **i)** $a_1 = 2$; $b_1 = 1$; $a_2 = 0.5$; $b_2 = 3$; $a_3 = 1.5$; $b_3 = 1$; $v = -1$; (the values of a_k and b_k, $k \in \{1, 2, 3\}$ are close)
    - **ii)** $a_1 = 6$; $b_1 = 0$; $a_2 = 2$; $b_2 = 8$; $a_3 = 10$; $b_3 = 5$; $v = 3$; (the values of a_k and $b_{k,}$ $k \in \{1, 2, 3\}$ are apart)
- Distribution of the microdata (real values $x_{jk}(w)$, with $w \in \{1, \ldots, 5000\},$) that allows generating the histograms corresponding to each observation of the variables X_k, with $k = \{1, 2, 3\}$.
  - **i)** Uniform distribution: $x_{jk}(w) \sim \mathcal{U}(\delta_1(j), \delta_2(j))$ are randomly generated considering for each $j \in \{1, \ldots, m\}$ and $k \in \{1, 2, 3\}$.
    - $k = 1$: $\delta_1(j) \sim \mathcal{U}(-1.5, 0.5)$ and $\delta_2(j) \sim \mathcal{U}(2.5, 4.5)$;
    - $k = 2$: $\delta_1(j) \sim \mathcal{U}(3, 5)$ and $\delta_2(j) \sim \mathcal{U}(11, 13)$;
    - $k = 3$: $\delta_1(j) \sim \mathcal{U}(-5, -3)$ and $\delta_2(j) \sim \mathcal{U}(9, 11)$;
  - **ii)** Normal distribution: $x_{jk}(w) \sim \mathcal{N}(\delta_3(j), \delta_4(j))$ are randomly generated considering for each $j \in \{1, \ldots, m\}$ and $k \in \{1, 2, 3\}$.
    - $k = 1$: $\delta_3(j) \sim \mathcal{U}(1, 2)$ and $\delta_4(j) \sim \mathcal{U}(0.5, 1.5)$;
    - $k = 2$: $\delta_3(j) \sim \mathcal{U}(7.5, 8.5)$ and $\delta_4(j) \sim \mathcal{U}(1.5, 2.5)$;
    - $k = 3$: $\delta_3(j) \sim \mathcal{U}(2.5, 3.5)$ and $\delta_4(j) \sim \mathcal{U}(3.5, 4.5)$;
  - **iii)** Log-Normal distribution: $x_{jk}(w) \sim Log\mathcal{N}(\delta_5(j), \delta_6(j))$ are randomly generated considering for each $j \in \{1, \ldots, m\}$ and $k \in \{1, 2, 3\}$.
    - $k = 1$: $\delta_5(j) \sim \mathcal{U}(-0.5, 0.5)$ and $\delta_6\ (j) \sim \mathcal{U}(0.5, 1)$;
    - $k = 2$: $\delta_5(j) \sim \mathcal{U}(1.5, 2.5)$ and $\delta_6(j) \sim \mathcal{U}(0, 0.5)$;
    - $k = 3$: $\delta_5(j) \sim \mathcal{U}(0, 1)$ and $\delta_6(j) \sim \mathcal{U}(0.75, 1.25)$;
  - **iv)** Mixture of distributions: $x_{jk}(w)$ are randomly generated considering for each $j \in \left\{1, \ldots, m\right\}$ and $k \in \{1, 2, 3\}$.
    - $k = 1$: randomly selected from $\mathcal{U}\ (-0.5, 3.5)$; $\mathcal{N}(1.5, 1)$; $\mathcal{X}^2(1)$; $Log\mathcal{N}(0,\ 0.75)$; $-Log\mathcal{N}(0, 0.75)$.
    - $k = 2$: randomly selected from $\mathcal{U}(4,\ 12)$; $\mathcal{N}(8, 2)$; $Log\mathcal{N}(2, 0.25)$; $-Log\ \mathcal{N}(2, 0.25)$; $\mathcal{X}^2(8)$.
    - $k = 3$: randomly selected from $\mathcal{U}\ (-4, 10)$; $\mathcal{N}(3, 4)$; $Log\mathcal{N}(0.5, 1)$; $-Log\mathcal{N}(0.5, 1)$; $\mathcal{X}^2(3)$.

  With the exception of the situation of the mixture of distributions, and as the goal is to study the effect of the type of distribution of the explicative variables, for each k, the histogram-valued variable X_k is built from different distributions but those distributions have similar values for the mean and standard deviation.

1. Error level: in the error function $e_j(t)$, the values of $\widetilde{c}_{(j)_1}$ and $\widetilde{r}(j)_i$ are randomly selected.

   i) The values of $\widetilde{c}_{(j)_1}$ are randomly generated in

   - $\mathcal{U}_{c1} = 0.1 * \mathcal{U}(-ma_H, ma_H)$;
   - $\mathcal{U}_{c2} = 0.5 * \mathcal{U}(-ma_H, ma_H)$;
   - $\mathcal{U}_{c3} = \mathcal{U}(-ma_H, ma_H)$

   with $ma_H = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{2}\left(\overline{I}_{Y^*(j)_{10}} - \underline{I}_{Y^*(j)_1}\right)$;

   ii) The values of $\widetilde{r}(j)_i$ are randomly generated in

   - $\mathcal{U}_{r1} = 0.1 * \mathcal{U}(-mr_i, mr_i)$;
   - $\mathcal{U}_{r2} = 0.5 * \mathcal{U}(-mr_i, mr_i)$;
   - $\mathcal{U}_{r3} = \mathcal{U}(-mr_i, mr_i)$.

   with, for each $i \in \{1, \ldots, n\}$, $mr_i = \min\limits_{j\in\{1,\ldots,m\}}\left\{r_{Y^*(j)_i}\right\}$.

   The level I error corresponds to the case where the linearity is slightly disturbed (considering the variability of values in the distributions of the response variable), which implies that the values of $\widetilde{c}_{(j)_1}$ and $\widetilde{r}(j)_i$ in the error function $e_j(t)$ are randomly generated in $\mathcal{U}_{c1} = 0.1 * \mathcal{U}(-ma_H, ma_H)$ and $\mathcal{U}_{r1} = 0.1 * \mathcal{U}(-mr_i, mr_i)$, respectively. The level II error is considered when the values of $\widetilde{c}_{(j)_1}$ and $\widetilde{r}(j)_i$ are randomly generated in intervals with larger variability, $\mathcal{U}_{c3} = \mathcal{U}(-ma_H, ma_H)$ and $\mathcal{U}_{r3} = \mathcal{U}(-mr_i, mr_i)$, respectively. Considering Ω as the measure to quantify the goodness-of-fit, it was observed that the disturbance of a linear relation between distributions must take into account the variability of the values in the distributions of the Y^*. This is the reason that led us to consider the range of the distributions of Y^* in the disturbance of the centers.

It is important to underline that in this simulation study, it was only possible to control the type of distributions of the observations of the explicative histogram-valued variables. This simulation does not allow selecting the distributions of the observations of the response variable. The distribution of the response variable depends on the distribution of the variable Y^* and/or the disturbance applied to the histograms $Y^*(j)$. In the studied cases, the variability of the values in distribution of the variable Y, when $p = 1$, is higher when $a = 2$; $b = 8$; $v = 3$.

In addition to the selection of parameters considered in the factorial design, other choices were analyzed. However, as the results were similar, we chose to present only the cases enumerated above.

The simulated symbolic data tables include the observations of the histogram-valued variables X_k and Y, according to the previous description and factors. For these tables, we computed the estimated parameters for the *DSD Model* and the goodness-of-fit measures. As we considered 1000 replications for each situation, the values presented are the means of the obtained values and the respective standard deviation values (represented by s).

The goodness-of-fit measures considered in this study are:

- $\overline{\Omega}$, where Ω is the measure deduced from the *DSD Model* (see Section 3.4);
- Root-mean-square error ($RMSE_M$), a measure defined using the Mallows distance (also used in the *DSD Model*), proposed by Irpino and Verde [19]; it is defined by

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

- Adaptations of the lower ($RMSE_L$) and the upper bound ($RMSE_U$) root-mean-square that Neto and Carvalho [17,18], use to study the performance of the linear regression models defined for interval-valued variables. For histogram-valued variables, the $RMSE_L$ and the $RMSE_U$ are given by:

$$RMSE_L = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}(\underline{I}_{\widehat{Y}(j)_i} - \underline{I}_{Y(j)_i})^2 p_i}$$

$$RMSE_U = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}(\overline{I}_{\widehat{Y}(j)_i} - \overline{I}_{Y(j)_i})^2 p_i}$$

with $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i}\right)$ and $\left[\underline{I}_{\widehat{Y}(j)_i}, \overline{I}_{\widehat{Y}(j)_i}\right)$ the subintervals $i \in \{1, \ldots, n\}$ of the observed and predicted histograms, for each unit j.

4.1.3. *Results and conclusions*

In Appendix C, three tables are presented, each of which contains the results obtained when applying the DSD

Model, with $p = 1$, in different conditions and considering the following three selections of the parameters: $a = 2$; $b = 1$; $v = -1$ (Table C1); $a = 2$; $b = 8$; $v = 3$ (Table C2); and $a = 6$; $b = 0$; $v = 2$ (Table C3). From Tables C4 to C7, similar results are presented for the cases obtained by applying the *DSD Model* with $p = 3$.

The main goals of this study are to verify if the goodness-of-fit measures are in accordance with the considered error levels and to analyze the performance of the *DSD Model* applied to histogram-valued variables evaluating the behavior of the parameters' estimation. For $p = 1$, it is also our goal to analyze how the symmetry/asymmetry of the distributions of the observations of the explicative histogram-valued variable affect the symmetry/asymmetry of the distributions in the observations of the response variable.

Concerning the goodness-of-fit measures versus level of linearity

To evaluate if the linear relation between distributions is strong or weak, we used the coefficient of determination Ω, deduced from the *DSD Model*. However, when we delineated this simulation study with the goal of analyzing the performance of the *DSD Model* applied to histogram-valued variables, it was necessary to define what is the meaning of a high or low disturbance. As Ω was the selected measure to quantify the goodness-of-fit, the levels of disturbance were defined taking into account its expected behavior.

In all studied situations (see tables in Appendix C), when we consider a disturbance with error level I, Ω presents values close to one, i.e. it indicates that the linear relation between the distributions is strong. When a higher disturbance (error level II) is considered, the values of Ω are, as expected, more distant from one and approach zero. As the error functions are defined considering the range of values in distributions of Y^*, the behavior of Ω shows that the disturbance takes into account the variability of the values in these distributions. Therefore, in order to obtain a similar value of Ω, the linear relations when the explicative variables have low variability need to be less disturbed than in the case when the explicative variables have higher variability.

The tables in Appendix C record only the values of the goodness-of-fit measures considering two levels of variability for the error function. However, to analyze more comprehensively the level of sensitivity of Ω, for different kinds of error functions, we must consider some cases where the error functions affect more the half-range of the subintervals of the histograms and other cases where the centers are more affected. Tables 2 and 3 illustrate the results that were obtained for samples with 10 and 100 observations, for *DSD Model* with $a = 2$; $b = 8$; $v = 3$. The values of Ω were determined considering different error functions that use three levels of variability for the values of $\widetilde{c}_{(j)_1} : \mathcal{U}_{c1}$, $\mathcal{U}_{c2}$, $\mathcal{U}_{c3}$ and, for each one, three levels of variability for $\widetilde{r}(j)_i : \mathcal{U}_{r1}$, $\mathcal{U}_{r2}$, $\mathcal{U}_{r3}$, as defined in Section 4.1.2.

Based on these results, we can say that the linearity between histogram-valued variables is more affected by disturbances in the centers of the subintervals than in the half-ranges. This behavior is not surprising because the distance associated with this model is the Mallows distance, and as we have observed, the contribution of the centers of

Table 2. Mean values of Ω considering different levels of linearity, when the distributions generating observations of X are Uniform $\left(\overline{\Omega}_{\mathcal{U}}\right)$ and Normal $\left(\overline{\Omega}_{\mathcal{N}}\right)$.

		$\overline{\Omega}_{\mathcal{U}}$ (s)			$\overline{\Omega}_{\mathcal{N}}$ (s)		
	m	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r1}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r2}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r3}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r1}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r2}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r3}$
$\widetilde{c}_{(j)_1} \sim \mathcal{U}_{c1}$	10	0.9924 (0.0027)	0.9798 (0.0075)	0.9434 (0.0212)	0.9801 (0.0068)	0.9628 (0.0139)	0.9130 (0.0282)
	100	0.9907 ($8.4E-4$)	0.9809 (0.0020)	0.9521 (0.0053)	0.9762 (0.0021)	0.9620 (0.0039)	0.9201 (0.0080)
$\widetilde{c}_{(j)_1} \sim \mathcal{U}_{c2}$	10	0.8508 (0.0453)	0.8410 (0.0480)	0.8144 (0.0551)	0.6804 (0.0753)	0.6713 (0.0769)	0.6487 (0.0798)
	100	0.8163 (0.0144)	0.8102 (0.0148)	0.7901 (0.0166)	0.6285 (0.0214)	0.6241 (0.0227)	0.6061 (0.0236)
$\widetilde{c}_{(j)_1} \sim \mathcal{U}_{c3}$	10	0.6028 (0.0911)	0.5919 (0.0896)	0.5856 (0.0885)	0.3661 (0.0863)	0.3567 (0.0798)	0.3563 (0.0844)
	100	0.5315 (0.0250)	0.5273 (0.0252)	0.5183 (0.0268)	0.3023 (0.0209)	0.3001 (0.0208)	0.2966 (0.0204)

Table 3. Mean values of Ω considering different levels of linearity, when the distributions generating observations of X are Log-Normal $\left(\overline{\Omega}_{Log\mathcal{N}}\right)$ and a mixture of distributions $\left(\overline{\Omega}_{Mix}\right)$.

		$\overline{\Omega}_{Log\mathcal{N}}$ (s)			$\overline{\Omega}_{Mix}$ (s)		
	m	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r1}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r2}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r3}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r1}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r2}$	$\widetilde{r}_{(j)_i} \sim \mathcal{U}_{r3}$
$\widetilde{c}_{(j)_1} \sim \mathcal{U}_{c1}$	10	0.9779 (0.0074)	0.9424 (0.0209)	0.8459 (0.0460)	0.9800 (0.0066)	0.9773 (0.0077)	0.9699 (0.0104)
	100	0.9784 (0.0020)	0.9719 (0.0029)	0.9512 (0.0050)	0.9788 (0.0019)	0.9761 (0.0023)	0.9677 (0.0033)
$\widetilde{c}_{(j)_1} \sim \mathcal{U}_{c2}$	10	0.6692 (0.0736)	0.6450 (0.0804)	0.6108 (0.0939)	0.6707 (0.0767)	0.6702 (0.0811)	0.6648 (0.0813)
	100	0.6483 (0.0206)	0.6464 (0.0212)	0.6361 (0.0227)	0.6505 (0.0234)	0.6501 (0.0254)	0.6461 (0.0247)
$\widetilde{c}_{(j)_1} \sim \mathcal{U}_{c3}$	10	0.3439 (0.0867)	0.3416 (0.0829)	0.3305 (0.0896)	0.3520 (0.0983)	0.3533 (0.0970)	0.3568 (0.0958)
	100	0.3165 (0.0224)	0.3160 (0.0218)	0.3150 (0.0225)	0.3198 (0.0329)	0.3208 (0.0319)	0.3191 (0.0319)

the subintervals is three times larger than that of the half-ranges (see Definition 4). Moreover, the limitation imposed on the values of $\widetilde{r}(j)_i$ prevents larger disturbances in the half-ranges of the subintervals of Y_j.

Comparing all situations where the considered error levels are the same, we may conclude that when the observations of the explicative variables follow a Uniform distribution, the linearity is less disturbed. This behavior could be influenced by the non-variability of the half-ranges associated with each unit j.

An in-depth analysis of the values of the goodness-of-fit measures (see all tables of Appendix C) shows that the values of the root mean square error decrease in the same proportion as the levels of linearity. The mean values associated with the measures $RMSE_M$; $RMSE_L$ and $RMSE_U$ increase approximately ten times when we pass from high (error level I) to moderate/low (error level II) linearity. This increase is an exact reflection of the range of variability tested in this study for the error function (ten times from level I to level II). As the measures of the root mean square errors quantify the differences between the observed and predicted distributions in ”absolute terms,” these are not adequate to compare situations with different selections of parameters in the *DSD Model*. They are also not adequate when the selection of parameters is the same, but the distributions of the explicative variables are different.

Concerning the analysis of the parameters’ estimation

The results obtained using the *DSD Model* with one (Tables C1 to C3) or three (Tables C4 to C7) explicative variables are in general similar and as such in this section, we will analyze in detail the results obtained when $p = 1$.

Comparing the obtained results, we can see that the behavior of the parameters’ estimation is not very different for all distributions used to generate the microdata of the explicative variables and is similar for the three selections of the parameters. For the situations where the level I error is considered, we observe that the mean value of the estimated parameters is close to the true parameter values; both the standard deviation associated with the mean values of the estimated parameters and the values of MSE get closer to zero when the number of observations increases. This result confirms the empirical consistency of the estimation and is expected when the linear regression models are only slightly disturbed.

In Figs. 8, 9 and 10, we may observe the behavior described above. The figures illustrate only the situation where $a = 2$; $b = 8$; $v = 3$ (Table C2), but the behavior for the other selections of parameters is similar. For the different distributions used to generate the histogram values of X, the boxes reduce their ranges around the true values of the respective parameters as the number m of observations increases. It may also be observed that it is for the Normal distribution that the diversity of the estimated values of the parameters is higher.

The behavior of the independent parameter is always more unstable than the behavior of the other parameters of the model. Observing the obtained results, it is when the variability distributions of the response variable is higher that the values/quantile functions estimated for the independent parameters are more apart from the original values/quantile functions. In Fig. 10, we may observe that the values of the MSE correspondent to the independent parameter will be closer to zero when sizes of the samples increase. It is when the distribution of the explicative variables is Normal that these values are higher.

When we consider the level II error, the mean values associated with the estimated parameters a and b are distant from the original ones, essentially when the distributions of the explicative variables are Uniform or Normal and the number of observations is lower. In all situations, the estimated parameters have higher values of standard deviation and MSE than in the analogous situations when level I error is considered.

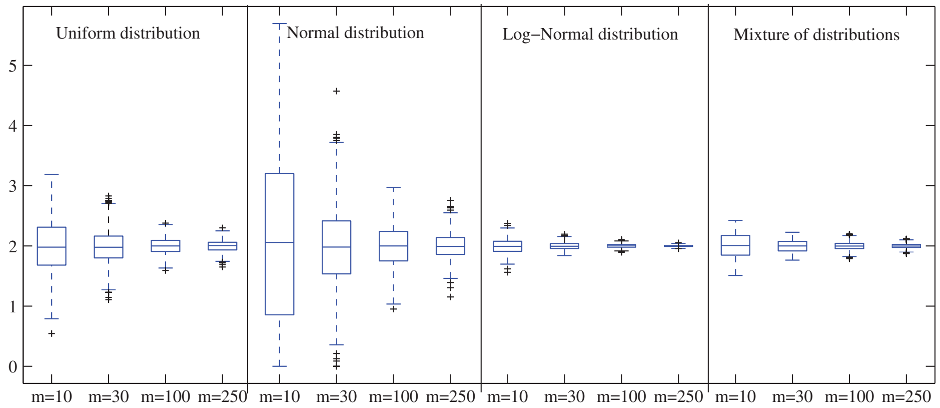

Fig. 8 Boxplots of the values estimated for parameter a, under different conditions, when *DSD Model* ($a = 2$, $b = 8$, $v = 3$) is applied to histogram-valued variables and when level I error is considered.

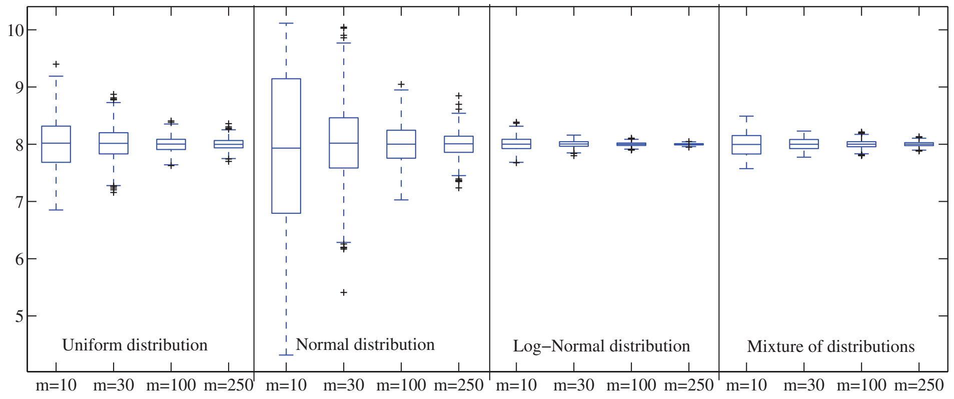

Fig. 9 Boxplots of the values estimated for parameter b, under different conditions, when *DSD Model* ($a = 2$, $b = 8$, $v = 3$) is applied to histogram-valued variables and when level I error is considered.

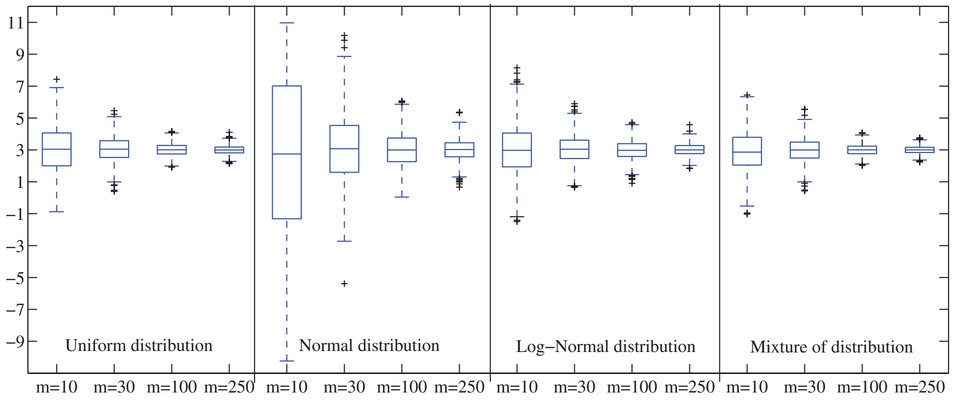

Fig. 10 Boxplots of the values estimated for parameter v, under different conditions, when *DSD Model* ($a = 2$, $b = 8$, $v = 3$) is applied to histogram-valued variables and when level I error is considered.

Concerning symmetry/asymmetry of $\widehat{Y}(j)$.

In this simulation study, it was possible to analyze the symmetry/assymetry of the predicted distributions obtained by the simple *DSD Model*, taking into consideration the symmetry/asymmetry of the distributions in the observations of the histogram-valued variables X and the values of the parameters of the models. When the observations of the histogram-valued variable X are symmetric histograms, represented by $\Psi^{-1}_{X(j)}(t)$, the respective symmetric histogram represented by $-\Psi^{-1}_{X(j)}(1-t)$ is also symmetric; but when the histogram represented by $\Psi^{-1}_{X(j)}(t)$ is asymmetric positive (negative) (Log-Normal, for example), the respective symmetric histogram represented by $-\Psi^{-1}_{X(j)}(1-t)$ is asymmetric negative (positive). In the *DSD Model*, the predicted distributions are obtained from $\Psi^{-1}_{\widehat{Y}(j)}(t) = v + a\Psi^{-1}_{X(j)}(t) - b\Psi^{-1}_{X(j)}(1-t)$. Therefore, if the distribution $\Psi^{-1}_{X(j)}(t)$ is symmetric, the distribution of $\Psi^{-1}_{\widehat{Y}(j)}(t)$ also tends to be symmetric. If the distribution $\Psi^{-1}_{X(j)}(t)$ is asymmetric, the distribution of $\Psi^{-1}_{\widehat{Y}(j)}(t)$ tends to be symmetric when the values of a and b are similar and asymmetric negative (resp. positive) when the value of a is lower (resp. higher) than the value of b. These conclusions are illustrated in Fig. 11 considering all predicted distributions in the simulation study with *DSD Model* for $p = 1$ and for samples with 10 observations.

In conclusion, when the distributions of observations $X(j)$ are symmetric, asymmetric positive, or asymmetric negative, it is possible, in several cases, to forecast whether the distributions of $\widehat{Y}(j)$ will be asymmetric. The value of the independent parameter does not influence the symmetry/asymmetry of $\widehat{Y}(j)$.

4.2. Applied examples

4.2.1. The relation between the hematocrit values and hemoglobin values

This first example was presented in Billard and Diday [3] to illustrate their linear regression model for histogram-valued variables. In this case, we have the symbolic data in Table 4, where 10 units are described by two symbolic variables, the hematocrit and the hemoglobin.

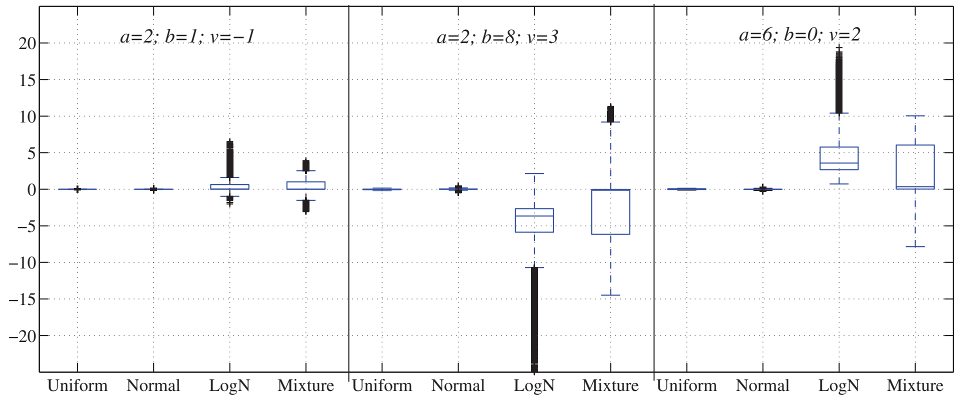

Fig. 11 Boxplots that represent the "skewness"[2] of the distributions estimated with *DSD Model*.

Table 4. Example of symbolic data table where the two variables hematocrit and hemoglobin are histogram-valued variables.

Obs.	Hematocrit (Y)	Hemoglobin (X)
1	{[33.29; 37.52) , 0.6; [37.52; 39.61] , 0.4}	{[11.54; 12.19) , 0.4; [12.19; 12.8] , 0.6}
2	{[36.69; 39.11) , 0.3; [39.11; 45.12] , 0.7}	{[12.07; 13.32) , 0.5; [13.32; 14.17] , 0.5}
3	{[36.69; 42.64) , 0.5; [42.64; 48.68] , 0.5}	{[12.38; 14.2) , 0.3; [14.2; 16.16] , 0.7}
4	{[36.38; 40.87) , 0.4; [40.87; 47.41] , 0.6}	{[12.38; 14.26) , 0.5; [14.26; 15.29] , 0.5}
5	{[39.19; 50.86] , 1}	{[13.58; 14.28) , 0.3; [14.28; 16.24] , 0.7}
6	{[39.7; 44.32) , 0.4; [44.32; 47.24] , 0.6}	{[13.81; 14.5) , 0.4; [14.5; 15.2] , 0.6}
7	{[41.56; 46.65) , 0.6; [46.65; 48.81] , 0.4}	{[14.34; 14.81) , 0.5; [14.81; 15.55] , 0.5}
8	{[38.4; 42.93) , 0.7; [42.93; 45.22] , 0.3}	{[13.27; 14.0) , 0.6; [14.0; 14.6] , 0.4}
9	{[28.83; 35.55) , 0.5; [35.55; 41.98] , 0.5}	{[9.92; 11.98) , 0.4; [11.98; 13.8] , 0.6}
10	{[44.48; 52.53] , 1}	{[15.37; 15.78) , 0.3; [15.78; 16.75] , 0.7}

We predicted the quantile function representing the distribution taken by the histogram-valued variable Y from the *DSD Model*, and obtained:

$$\Psi^{-1}_{\hat{Y}(j)}(t) = -1.953 + 3.5598\Psi^{-1}_{X(j)}(t) - 0.4128\Psi^{-1}_{X(j)}(1-t)$$

The value of the goodness-of-fit measure is, for this case, $\Omega = 0.9631$.

In Fig. 12, we may compare the quantile functions of the observed and predicted distributions of the histogram-valued variable Y. As it may be observed, the distributions are very similar, in agreement with the value of the coefficient of determination, Ω. The observed and predicted histograms of each observation are presented in Appendix D.

When we predict a histogram value, we have always associated an error function defined according to Definition 5. For this example, in Fig. 13 we can observe the error function for observations 1 and 3.

The relation between the histogram-valued variables in Table 4 may be visualized in the scatter plot for histograms in Fig. 14. In this graphic, each of the distributions is represented by a histogram with a different color. These graphics show that a strong linear relation between the histogram-valued variables hematocrit and hemoglobin is observed.

From Proposition 2, we may conclude that for the set of patients to which the data refer, the symbolic mean of hematocrit increases $a - b = 3.1470$ for each unit of increase of the symbolic mean of hemoglobin. As this value is positive, we may consider that the relation between the histogram-valued variables is direct.

For this example, we also predicted the hematocrit distributions using the linear regression models proposed by Billard and Diday [3] (the *Center Model (CM)* and *Billard and Diday Model (BD)*) and Irpino and Verde [15,19] (the *Verde and Irpino Model (VI)*). The hematocrit distributions obtained by these methods are presented in Appendix D. Even though a solution to build the predicted histograms is not proposed by Billard and Diday [3], the predicted distributions may be built if we consider the process of Irpino and Verde [9]. In fact, we may rewrite all histograms (of all variables involved) with the same number n of subintervals and the weight associated to each subinterval $i = 1, \ldots, n$ in all units, is the same for each i. After this process, the histograms may be predicted by multiplying the

[2] The "skewness" in this context is measured by the difference between the symbolic mean and the symbolic median. A distribution is considered to be asymmetric positive (negative) when this difference is positive (negative).

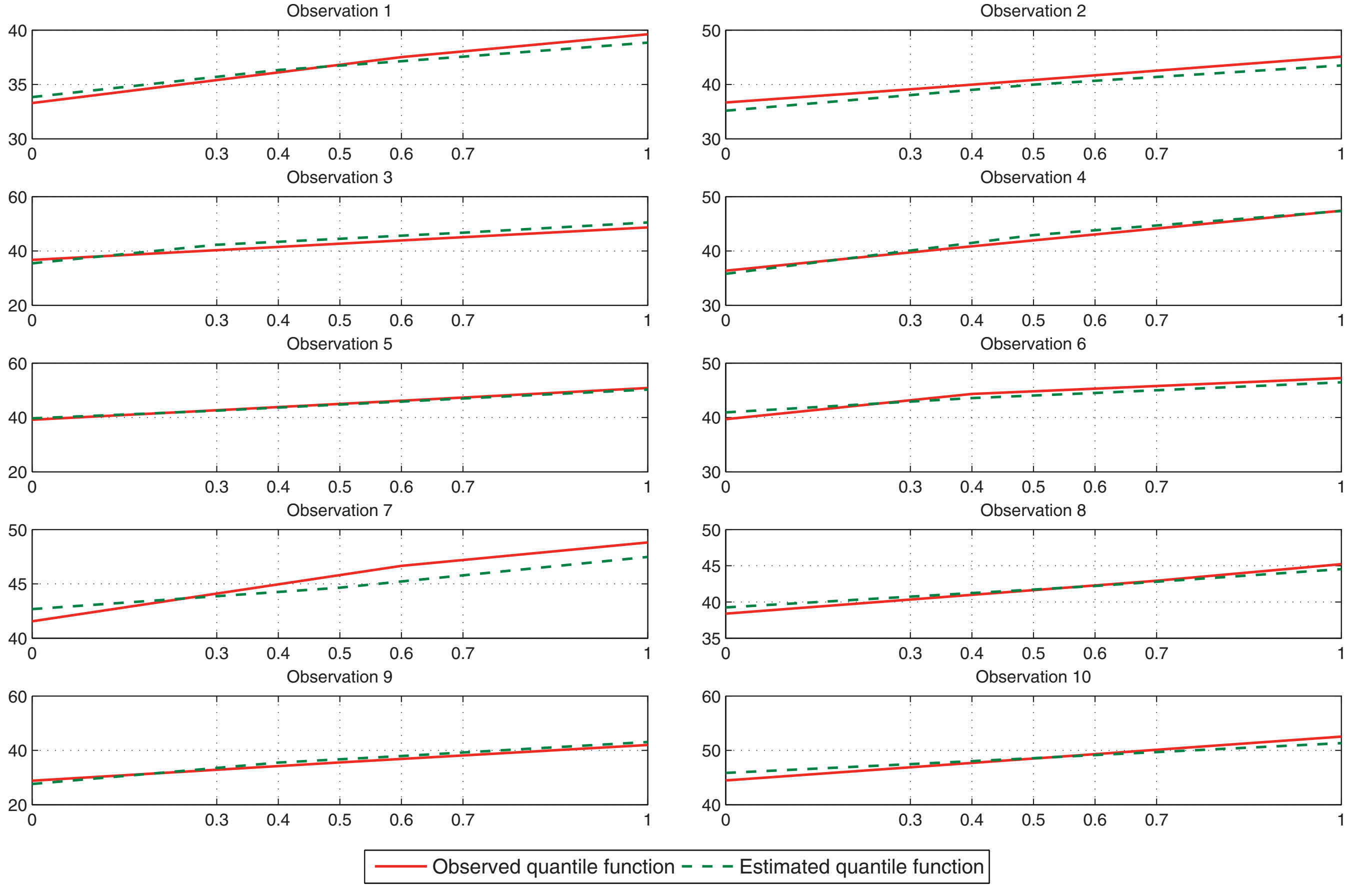

Fig. 12 Observed and predicted quantile functions for each observation in Table 4.

subintervals that compose the histogram associated to each unit j by the respective parameter. The predicted histogram $Y(j)$ is composed by n subintervals and each subinterval i with $i = \{1, \ldots, n\}$ is obtained by adding the subintervals i that compose the distributions of the explicative variables to the unit j.

To compare the performance of the methods, the measures $RMSE_M$, $RMSE_L$, $RMSE_U$, and Ω (see Section 4.1.2) were used (see Table 5).

4.2.2. *Distributions of Crimes in USA*

In this example, we consider a real data table (microdata) [30] where we have records related with communities in the USA. The original data combine socioeconomic data from the '90 Census and crime data from 1995. For this study, we selected the response variable *violent crimes* (total number of violent crimes per 100 000 habitants) and four explicative variables: X_1 (percentage of people aged 25 years and above with less than 9th grade education); X_2 (percentage of people aged 16 years and above who are employed); X_3 (percentage of population who are divorced); and X_4 (percentage of immigrants who immigrated within the last 10 years). To build the symbolic data table, we aggregated the information (contemporary aggregation) for each state. The units (higher units) of this study are the states of USA and their observations for each selected variable are the distributions of the records of the communities of the respective state. To build the initial data table, we considered only the states for which the number of records for the variables selected was higher than 30. Using this criterion, only 20 states were included (AL, CA, CT, FL, GA, IN, MA, MO, NC, NJ, NY, OH, OK, OR, PA, TN, TX, VA, WA, WI). Similar to the simulation study, we consider that in all observations, the subintervals of each histogram have the same weight (equiprobable) with frequency 0.20. Furthermore, as the response variable *violent crimes* admits only positive values and the distributions of these values are asymmetric, we will consider as response histogram-valued variable the variable LVC whose observations are the distributions of the logarithm of the number of violent crimes in each USA state. Considering these conditions, the model that allows predicting the distribution of LVC from the distributions of the explicative variables X_1, X_2, X_3 and

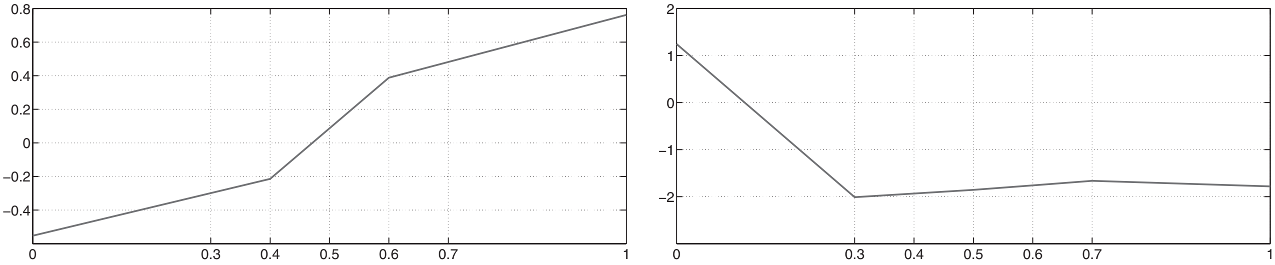

Fig. 13 Error function for the observations 1 and 3.

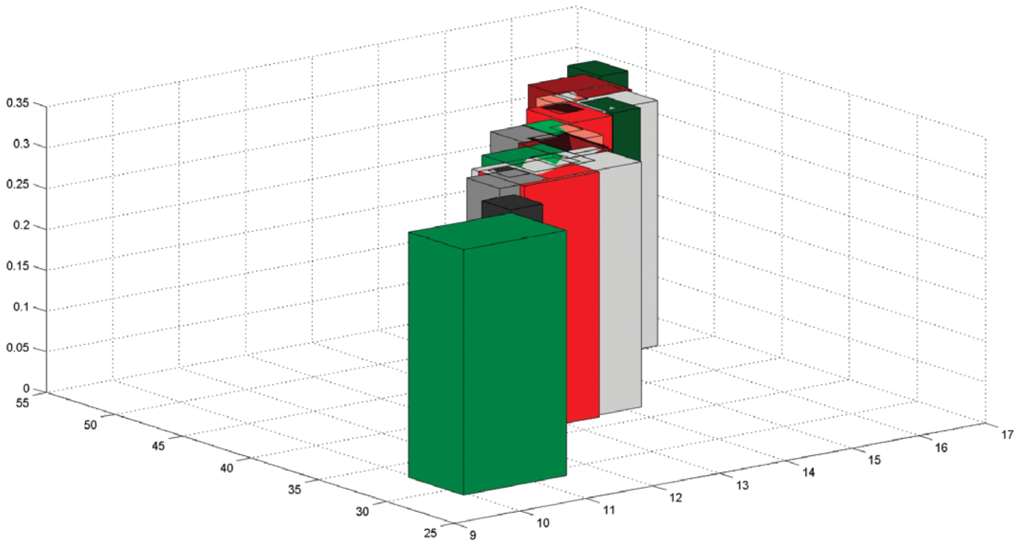

Fig. 14 Scatter plot of the data in Table 4.

Table 5. Comparison of the expressions and performance of the symbolic linear regression models for histogram-valued variables in Table 4.

Models	Expressions that allow predicting the distributions	$RMSE_L$	$RMSE_U$	$RMSE_M$	Ω
DSD	$\Psi^{-1}_{\widehat{Y}(j)}(t) = -1.95 + 3.56\Psi^{-1}_{X(j)}(t) - 0.41\Psi^{-1}_{X(j)}(1-t)$	0.9621	0.9496	0.8946	0.9631
CM	$\widehat{\overline{Y}}(j) = -2.16 + 3.16\overline{X}(j)$	1.0636	1.1501	1.0507	0.8460
BD	$\widehat{Y}(j) = 2.28 + 2.85X(j)$	1.1291	1.3480	1.2292	0.6853
VI	$\Psi^{-1}_{\widehat{Y}(j)}(t) = -2.16 + 3.16\overline{X}(j) + 3.92\Psi^{c^{-1}}_{X(j)}(t)$	1.0072	0.9633	0.9145	0.9613

X_4, for each USA state j is as follows:

$$\begin{aligned}\Psi^{-1}_{\widehat{LVC(j)}}(t) = 3.9321 &+ 0.0009\Psi^{-1}_{X_1(j)}(t)\\ &- 0.0123\Psi^{-1}_{X_2(j)}(1-t) + 0.2073\Psi^{-1}_{X_3(j)}(t)\\ &- 0.0353\Psi^{-1}_{X_3(j)}(1-t) + 0.0187\Psi^{-1}_{X_4(j)}(t)\end{aligned} \tag{17}$$

with $t \in [0, 1]$.

The values of the parameters estimated for this situation allow concluding that the variables X_1, X_3, and X_4 have a direct influence in the logarithm of the number of violent crimes and the percentage of employed people has an opposite effect. From Proposition 2, we may conclude that, for the set of states to which the data refer, when the symbolic mean of the percentage of population divorced increases 1% and the other variables remain constant, the symbolic mean of the LVC increases 0.1720. The percentage of divorced population is the one that influences the most the predicted histogram-valued variable. This

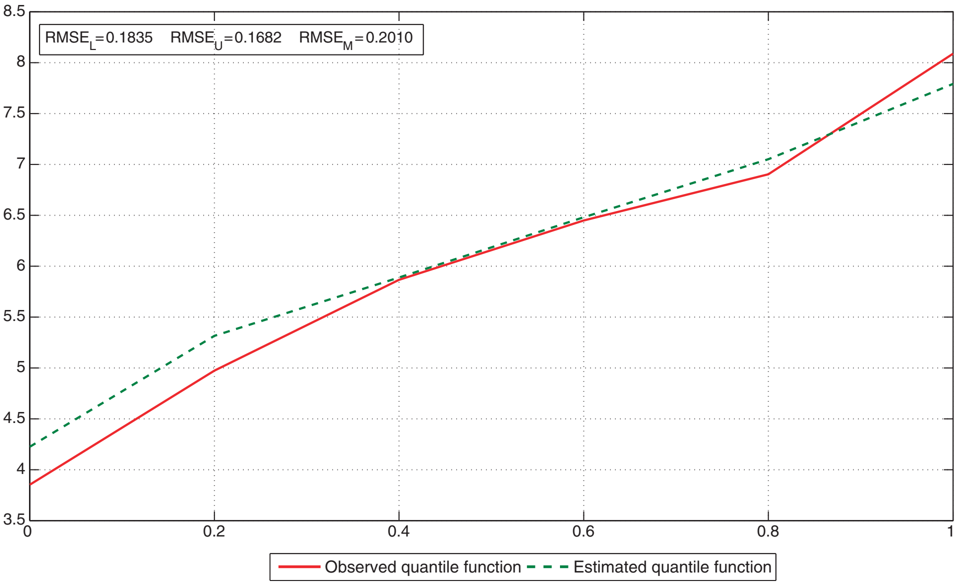

Fig. 15 Observed and estimated quantile function of the variable LVC in the state of Arkansas

interpretation can be extrapolated for the values of the associated parameter of all other explicative variables.

Consider one state that was not used to build the model, the state of Arkansas (AR). It is possible to predict the distribution of LVC if the distributions of the explicative variables for this state are known. The histogram predicted by the *DSD Model* (17) for the state Arkansas is

$$H_{LVC}(AR) = \{[4.2250, 5.3158), 0.2; [5.3158, 5.8887), 0.2; \\ [5.8887, 6.4802), 0.2; [6.4802, 7.0509), 0.2; \\ [7.0509, 7.7913], 0.2\}$$

Fig. 15 illustrates the estimated and observed quantile function for this state and the values of the measures $RMSE_M$, $RMSE_L$, $RMSE_U$ (see *Section 4.1.2*). The values of the goodness-of-fit measures are in accordance with the closeness between the observed and estimated quantile function that we may see in the figure.

Analyzing the predicted distribution, we may conclude that in the state of Arkansas, the estimated distribution tends to a uniform behavior with the values of LVC ranging between 4.23 and 7.79.

For this example, we also predicted the logarithm of the number of violent crimes using the linear regression models proposed by Billard and Diday [3] and Irpino and Verde [15,19]. In the case of the *CM*, the predicted histograms were built using the process described in *Section 4.2.1*. However, we obtain results that are not histograms because in some subintervals, the lower bound is greater than the upper bound. For each case, to build the subintervals, the lowest obtained value should be used for the lower bound and the highest for the upper bound. In this way, we obtain histograms where the subintervals are neither ordered nor disjoint, but may be rewritten according to the process of Williamson [22].

Table 6. Comparison of the expressions of the symbolic linear regression models that predict the number of violent crimes in USA states.

Models	Expressions that allow predicting the distributions
DSD	$\Psi^{-1}_{\widehat{LVC}(j)}(t) = 3.93 + 0.001\Psi^{-1}_{X_1(j)}(t) - 0.01\Psi^{-1}_{X_2(j)}(1-t) + 0.21\Psi^{-1}_{X_3(j)}(t) - 0.04\Psi^{-1}_{X_3(j)}(1-t) + 0.02\Psi^{-1}_{X_4(j)}(t)$
CM	$\widehat{\overline{LVC}}(j) = 6.01 + 0.09\overline{X}_1(j) - 0.05\overline{X}_2(j) + 0.11\overline{X}_3(j) + 0.01\overline{X}_4(j)$
BD	$\widehat{LVC}(j) = 4.40 + 0.03X_1(j) - 0.02X_2(j) + 0.11X_3(j) + 0.02X_4(j)$
VI	$\Psi^{-1}_{\widehat{LVC}(j)}(t) = 6.01 + 0.09\overline{X}_1(j) - 0.05\overline{X}_2(j) + 0.11\overline{X}_3(j) + 0.01\overline{X}_4(j) + 0.01\Psi^{c^{-1}}_{X_2(j)}(t) + 0.32\Psi^{c^{-1}}_{X_3(j)}(t) + 0.01\Psi^{c^{-1}}_{X_4(j)}(t)$

In Tables 6 and 7, it is possible to compare results obtained by different methods and their respective performance. The LVC distributions predicted by these methods are presented in Appendix D.

As we observed in the previous example, the linear regression models proposed by Billard and Diday present a weak performance. The behavior of the *DSD Model* and of the *Verde and Irpino Model* are similar in both examples. A similar performance was also

Table 7. Performance of the symbolic linear regression models that predict the number of violent crimes in USA states.

Models	$RMSE_L$	$RMSE_U$	$RMSE_M$	Ω
DSD	0.5571	0.4233	0.4477	0.8680
CM	0.9182	0.5617	0.6717	0.4585
BD	0.7927	0.4665	0.5801	0.4415
VI	0.5214	0.3444	0.3933	0.8982

verified for the example studied in the work of Irpino and Verde [19].

The advantage of studying a linear relation between data with variability is the possibility of predicting the distribution of the values of the response variable instead of only one real value as in a classical study. The classical alternative to study the logarithm of the number of violent crimes in each USA state would be to reduce the records of all communities of each state, for example to the mean value, and apply classical linear regression. In this case, the variability of the records would be lost and the predicted results would be less informative. Considering the mean of the records associated to each community, the classical model is the following:

$$\begin{aligned}\widehat{LVC}(j) &= 6.5817 + 0.0705\overline{X}_1(j) - 0.0503\overline{X}_2(j)\\ &\quad + 0.0933\overline{X}_3(j) + 0.0177\overline{X}_4(j) \qquad (18)\end{aligned}$$

For this model, the value of $R^2 = 0.75$.

Considering again the state of Arkansas, with the previous model (18), the estimative for $\widehat{LVC}(AR)$ is 6.4511. With this approach, the information about the behavior of the predicted variable is obviously poorer.

5. CONCLUSION AND PERSPECTIVES

The main advantages of the proposed *DSD Model* are that: 1) it allows predicting the distributions taken by one histogram-valued variable from the distributions taken by explicative histogram-valued variables; 2) the parameters are easily estimated by solving a quadratic optimization problem or a constrained least squares problem, subject to nonnegative constraints on the unknowns; 3) from the model, the prediction of the distributions for the observations of the response histogram-valued variable is immediate; and 4) it is possible to deduce a goodness-of-fit measure from the model. This measure is deduced similarly as to classical statistics and appears to have a good behavior. When we compare the predicted and observed quantile functions for each unit, we have good estimates when the value of the goodness-of-fit measure is close to one, whereas the predicted and observed quantile functions are more discrepant when the value of the goodness-of-fit measure is lower.

An extension of the *DSD Model*, where instead of a real number a quantile function is used as the independent parameter , is under development. With this new approach, we expect to obtain a more flexible model. As interval-valued variables are a particular case of histogram-valued variables, it is possible to particularize the two approaches of the model to interval-valued variables.

Finally, and as a future research perspective, other models and methods in Symbolic Data Analysis based on linear relations between variables may now be developed using this approach.

6. ACKNOWLEDGMENTS

Project "NORTE-07-0124-FEDER-000059" is financed by the North Portugal Regional Operational Programme (ON.2 - O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

The authors would also like to thank Professor Antonio Irpino, of the Second University of Naples (Italy), for kindly providing the Matlab code of the Verde and Irpino Model, used in Section 4.

APPENDIX A: FIRST ORDER PARTIAL DERIVATIVES OF THE FUNCTION SEE

$$\begin{aligned}SEE = \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Bigg[&\left(c_{Y(j)_i} - \sum_{k=1}^{p}\left(a_k c_{X_k(j)_i} - \beta_k c_{X_k(j)_{n-i+1}}\right) - v\right)^2\\ &+ \frac{1}{3}\left(r_{Y(j)_i} - \sum_{k=1}^{p}\left(a_k r_{X_k(j)_i} + b_k r_{X_k(j)_{n-i+1}}\right)\right)^2\Bigg]\end{aligned}$$

In these partial derivatives the subintervals of the histograms are defined from the center and half-range of the intervals.

$$\begin{aligned}\frac{\partial SEE}{\partial a_{\mathbf{k}}} &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Bigg[2\left(c_{Y(j)_i} - v - \sum_{k=1}^{p}\left(a_k c_{X_k(j)_i} + b_k\left(-c_{X_k(j)_{n-i+1}}\right)\right)\right)\\ &\quad \times(-c_{X_{\mathbf{k}}(j)_i}) + \tfrac{2}{3}\left(r_{Y(j)_i} - \sum_{k=1}^{p}\left(a_k r_{X_k(j)_i} + b_k r_{X_k(j)_{n-i+1}}\right)\right)(-r_{X_{\mathbf{k}}(j)_i})\Bigg]\\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2\sum_{k=1}^{p} c_{X_k(j)_i} c_{X_{\mathbf{k}}(j)_i} + \frac{2}{3}\sum_{k=1}^{p} r_{X_k(j)_i} r_{X_{\mathbf{k}}(j)_i}\right) a_k\\ &\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2\sum_{k=1}^{p} c_{X_k(j)_{n-i+1}} c_{X_{\mathbf{k}}(j)_i} + \frac{2}{3}\sum_{k=1}^{p} r_{X_k(j)_{n-i+1}} r_{X_{\mathbf{k}}(j)_i}\right) b_k\\ &\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} 2 p_i c_{X_{\mathbf{k}}(j)_i} v + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2 c_{Y(j)_i} c_{X_{\mathbf{k}}(j)_i} - \frac{2}{3} r_{Y(j)_i} r_{X_{\mathbf{k}}(j)_i}\right)\end{aligned}$$

$$\begin{aligned}
\tfrac{\partial SEE}{\partial b_{\mathbf{k}}} &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[2\left(c_{Y(j)_i} - v - \sum_{k=1}^{p}\left(a_k c_{X_k(j)_i} + b_k\left(-c_{X_k(j)_{n-i+1}}\right)\right)\right)\right. \\
&\quad \times \left(c_{X_{\mathbf{k}}(j)_{n-i+1}}\right) + \tfrac{2}{3}\left(r_{Y(j)_i} - \sum_{k=1}^{p}\left(\alpha_k r_{X_k(j)_i} + b_k r_{X_k(j)_{n-i+1}}\right)\right) \\
&\quad \left.\times \left(-r_{X_{\mathbf{k}}(j)_{n-i+1}}\right)\right] \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2\sum_{k=1}^{p} c_{X_k(j)_i} c_{X_{\mathbf{k}}(j)_{n-i+1}} + \frac{2}{3}\sum_{k=1}^{p} r_{X_k(j)_i} r_{X_{\mathbf{k}}(j)_{n-i+1}}\right) a_k \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2\sum_{k=1}^{p} c_{X_k(j)_{n-i+1}} c_{X_{\mathbf{k}}(j)_{n-i+1}}\right. \\
&\quad \left. + \tfrac{2}{3}\sum_{k=1}^{p} r_{X_k(j)_{n-i+1}} r_{X_{\mathbf{k}}(j)_{n-i+1}}\right) b_k \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} -2 p_i c_{X_{\mathbf{k}}(j)_{n-i+1}} v \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 c_{Y(j)_i} c_{X_{\mathbf{k}}(j)_{n-i+1}} - \frac{2}{3} r_{Y(j)_i} r_{X_{\mathbf{k}}(j)_{n-i+1}}\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial SEE}{\partial v} &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[-2\left(c_{Y(j)_i} - \sum_{k=1}^{p}\left(a_k c_{X_k(j)_i} + b_k\left(-c_{X_k(j)_{n-i+1}}\right)\right) - v\right)\right] \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2\sum_{k=1}^{p} \alpha_k c_{X_k(j)_i}\right) + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2\sum_{k=1}^{p} \beta_k c_{X_k(j)_{n-i+1}}\right) \\
&\quad + 2mv - \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 c_{Y(j)_i}\right)
\end{aligned}$$

APPENDIX B: PROOF OF PROPOSITION 3.5.

Defining the quantile functions $\Psi^{-1}_{Y(j)}(t)$ and $\Psi^{-1}_{\widehat{Y}(j)}(t)$ from the centers and half-ranges of the subintervals, according expression (4), in *Section 2.1* we have,

$$\begin{aligned}
&\sum_{j=1}^{m}\int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right) dt \\
&\quad = \sum_{j=1}^{m}\sum_{i=1}^{n}\int_{w_{i-1}}^{w_i} \left[c_{Y(j)_i} + \left(2\frac{t-w_{i-1}}{w_i - w_{i-1}} - 1\right) r_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right. \\
&\qquad \left. - \left(2\tfrac{t-w_{i-1}}{w_i-w_{i-1}} - 1\right) r_{\widehat{Y}(j)_i}\right]\left[c_{\widehat{Y}(j)_i} + \left(2\tfrac{t-w_{i-1}}{w_i-w_{i-1}} - 1\right) r_{\widehat{Y}(j)_i} - \overline{Y}\right] dt \\
&\quad = \sum_{j=1}^{m}\sum_{i=1}^{n}\int_{w_{i-1}}^{w_i} \left[\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right) + \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right)\left(2\frac{t-w_{i-1}}{w_i - w_{i-1}} - 1\right)\right] \\
&\qquad \times \left[\left(c_{\widehat{Y}(j)_i} - \overline{Y}\right) + r_{\hat{Y}(j)_i}\left(2\tfrac{t-w_{i-1}}{w_i-w_{i-1}} - 1\right)\right] dt \\
&\quad = \sum_{j=1}^{m}\sum_{i=1}^{n}\int_{w_{i-1}}^{w_i} \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\left(c_{\widehat{Y}(j)_i} - \overline{Y}\right) + \left(\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right) r_{\widehat{Y}(j)_i}\right. \\
&\qquad \left. + \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right)\left(c_{\widehat{Y}(j)_i} - \overline{Y}\right)\right)\left(2\tfrac{t-w_{i-1}}{w_i-w_{i-1}} - 1\right) \\
&\qquad + \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) r_{\widehat{Y}(j)_i}\left(2\tfrac{t-w_{i-1}}{w_i-w_{i-1}} - 1\right)^2 dt
\end{aligned}$$

Solving the definite integral, after some algebra and considering $w_i - w_{i-1} = p_i$, we obtain,

$$\begin{aligned}
\sum_{j=1}^{m}\sum_{i=1}^{n} p_i &\left[\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right) c_{\widehat{Y}(j)_i} + \frac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) r_{\widehat{Y}(j)_i}\right. \\
&\quad \left. - \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\overline{Y}\right]
\end{aligned}$$

From the expression (13) in *Section 3.3*,

$$\begin{aligned}
c_{\widehat{Y}(j)_i} &= \sum_{k=1}^{p} \alpha^*_k c_{X_k(j)_i} - b^*_k c_{X_k(j)_{n-i+1}} + v^*; \\
r_{\widehat{Y}(j)_i} &= \sum_{k=1}^{p} \alpha^*_k r_{X_k(j)_i} + b^*_k r_{X_k(j)_{n-i+1}}
\end{aligned}$$

from Proposition 3 also in *Section 3.3* we have $\overline{Y} = \overline{\widehat{Y}}$, so

$$\begin{aligned}
&\sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big[(c_{Y(j)_i} - c_{\widehat{Y}(j)_i})\Big(\sum_{k=1}^{p} \alpha^*_k c_{X_k(j)_i} - b^*_k c_{X_k(j)_{n-i+1}} + v^*\Big) \\
&\quad + \tfrac{1}{3}(r_{Y(j)_i} - r_{\widehat{Y}(j)_i})\Big(\sum_{k=1}^{p} \alpha^*_k r_{X(j)_i} + b^*_k r_{X(j)_{n-i+1}}\Big) \\
&\quad - (c_{Y(j)_i} - c_{\widehat{Y}(j)_i})\overline{\widehat{Y}}\Big] \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big[\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\sum_{k=1}^{p} \alpha^*_k c_{X_k(j)_i} + \frac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) \\
&\quad \times \sum_{k=1}^{p} \alpha^*_k r_{X_k(j)_i}\Big] + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \Big[-\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\sum_{k=1}^{p} \beta^*_k c_{X_k(j)_{n-i+1}} \\
&\quad + \tfrac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right)\sum_{k=1}^{p} b^*_k r_{X_k(j)_{n-i+1}}\Big] \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\left(v^* - \overline{\widehat{Y}}\right) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} \alpha^*_1 \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right) c_{X_1(j)_i} + \frac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) r_{X_1(j)_i}\right] \\
&\quad + \ldots + a^*_p \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right) c_{X_p(j)_i} + \tfrac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) r_{X_p(j)_i}\right] \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} \beta^*_1 \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\left(-c_{X_1(j)_{n-i+1}}\right)\right. \\
&\quad \left. + \tfrac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) r_{X_1(j)_{n-i+1}}\right] + \ldots \\
&\quad + a^*_p \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\left(-c_{X_p(j)_{n-i+1}}\right)\right. \\
&\quad \left. + \tfrac{1}{3}\left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i}\right) r_{X_p(j)_{n-i+1}}\right] \\
&\quad - \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i}\right)\left(v^* - \overline{\widehat{Y}}\right)
\end{aligned}$$

Comparing this expression with the partial derivatives of the function SEE (see Appendix A) we may write

$$\begin{aligned}
&\sum_{j=1}^{m}\int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right) dt \\
&\quad = -\tfrac{1}{2}\sum_{k=1}^{p} a^*_k \frac{\partial SEE}{\partial a_k}(B^*) - \frac{1}{2}\sum_{k=1}^{p} b^*_k \frac{\partial SEE}{\partial b_k}(B^*) \\
&\qquad + \tfrac{1}{2}\sum_{k=1}^{p} \frac{\partial SEE}{\partial v}(B^*)\left(v^* - \overline{\widehat{Y}}\right)
\end{aligned}$$

From the Kuhn Tucker conditions presented in *Section 3.3*, we have $\frac{\partial SEE}{\partial v}(B^*) = 0$; $a^*_k \frac{\partial SEE}{\partial \alpha_k}(B^*) = 0$ and $b^*_k \frac{\partial SEE}{\partial b_k}(B^*) = 0$ for all $k \in \{1, \ldots, p\}$ and $B^* = \begin{bmatrix}\alpha^*_1 & b^*_1 & a^*_2 & b^*_2 & \cdots & a^*_n & b^*_n & v^*\end{bmatrix}^T$. So,

$$\sum_{j=1}^{m}\int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right) dt = 0. \qquad \square$$

APPENDIX C: SIMULATION RESULTS OF THE STUDY PRESENTED IN SECTION 4.1

Table C1. Results, in different conditions, of the *DSD Model* with $a = 2$, $b = 1$ and $v = -1$.

Distribution of microdata	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)
Uniform distribution	Level I	10	2.0015 (0.1368)	0.0187	0.9988 (0.1375)	0.0189	−1.0024 (0.4364)	0.1903	0.9921 (0.0027)	0.3098 (0.0553)	0.3093 (0.0547)	0.3114 (0.0558)
		30	2.0020 (0.0810)	0.0066	0.9979(0.0808)	0.0065	−1.0090 (0.2316)	0.0537	0.9907 (0.0016)	0.3497 (0.0312)	0.3490 (0.0309)	0.3512 (0.0315)
		100	1.9977 (0.0559)	0.0031	1.0023 (0.0561)	0.0032	−0.9933 (0.1627)	0.0265	0.9792 (0.0019)	0.5083 (0.0242)	0.5080 (0.0241)	0.5091 (0.0243)
		250	2.0008 (0.0403)	0.0016	0.9992 (0.0404)	0.0016	−1.0023 (0.1220)	0.0149	0.9786 (0.0012)	0.5182 (0.0153)	0.5179 (0.0153)	0.5191 (0.0154)
	Level II	10	1.8793 (1.0839)	1.1882	1.1714 (1.0565)	1.1444	−0.5298 (3.4948)	12.4228	0.5854 (0.0937)	3.1077 (0.5710)	3.1031 (0.5652)	3.1229 (0.5753)
		30	1.9441 (0.7508)	0.5663	1.0631 (0.7357)	0.5446	−0.8140 (2.1537)	4.6685	0.5276 (0.0485)	3.4968 (0.3139)	3.4895 (0.3101)	3.5135 (0.3172)
		100	2.0002 (0.5497)	0.3019	1.0015 (0.5466)	0.2985	−0.9903 (1.6097)	2.5886	0.3265 (0.0225)	5.0721 (0.2285)	5.0688 (0.2273)	5.0797 (0.2296)
		250	1.9959 (0.2677)	0.0716	1.0032 (0.2680)	0.0718	−0.9854 (0.8166)	0.6663	0.5042 (0.0156)	3.4809 (0.1034)	3.4749 (0.1023)	3.4937 (0.1044)
Normal distribution	Level I	10	1.9829 (0.4405)	0.1942	1.0170 (0.4409)	0.1945	−0.9563 (1.5537)	2.4135	0.9801 (0.0070)	0.6153 (0.1159)	0.6153 (0.1147)	0.6191 (0.1166)
		30	1.9977 (0.1944)	0.0378	1.0025 (0.1942)	0.0377	−0.9939 (0.6529)	0.4259	0.9763 (0.0042)	0.6460 (0.0587)	0.6458 (0.0578)	0.6505 (0.0594)
		100	1.9970 (0.1038)	0.0108	1.0031 (0.1038)	0.0108	−0.9923 (0.3124)	0.0975	0.9761 (0.0023)	0.6358 (0.0309)	0.6354 (0.0306)	0.6381 (0.0311)
		250	2.0009 (0.1037)	0.0107	0.9990 (0.1038)	0.0108	−0.9999 (0.3132)	0.0980	0.9759 (0.0022)	0.6374 (0.0299)	0.6370 (0.0297)	0.6397 (0.0301)
	Level II	10	1.7233 (1.4533)	2.1864	1.4121 (1.4455)	2.2572	0.2289 (5.6502)	33.4031	0.3549 (0.0809)	6.3839 (1.0909)	6.3804 (1.0800)	6.4185 (1.0989)
		30	1.7704 (1.2191)	1.5374	1.2834 (1.2086)	1.5396	−0.1550 (4.1590)	17.9939	0.3062 (0.0448)	6.4961 (0.6143)	6.4896 (0.6055)	6.5406 (0.6210)
		100	1.9387 (0.8810)	0.7792	1.0727 (0.8702)	0.7618	−0.7882 (2.6483)	7.0514	0.2944 (0.0214)	6.3760 (0.3008)	6.3721 (0.2979)	6.3986 (0.3032)
		250	1.9564 (0.6377)	0.4082	1.0450 (0.6349)	0.4047	−0.8522 (1.9215)	3.7102	0.2867 (0.0133)	6.5374 (0.1951)	6.5317 (0.1928)	6.5649 (0.1971)
LogNormal distribution	Level I	10	2.0020 (0.0350)	0.0010	0.9980 (0.0330)	0.0010	−1.0030 (0.4640)	0.2150	0.9760 (0.0080)	1.2860 (0.2170)	1.2860 (0.2170)	1.3010 (0.2170)
		30	2.0001 (0.0180)	3.3688E-4	0.9998 (0.0180)	3.2923E-4	−1.0080 (0.2510)	0.0630	0.9750 (0.0040)	1.3120 (0.1130)	1.3120 (0.1130)	1.3150 (0.1130)
		100	2.0003 (0.0100)	9.4048E-5	0.9997 (0.0100)	9.1453E-5	−1.0040 (0.1750)	0.0310	0.9750 (0.0020)	1.7250 (0.0750)	1.7240 (0.0740)	1.7280 (0.0750)
		250	1.9999 (0.0050)	2.2135E-5	1.0001 (0.0050)	2.1963E-5	−0.9900 (0.1120)	0.0130	0.9860 (0.0010)	1.7280 (0.0500)	1.7270 (0.0500)	1.7300 (0.0500)
	Level II	10	2.0110 (0.3500)	0.1220	0.9970 (0.3370)	0.1130	−0.9890 (4.7560)	22.5990	0.3150 (0.0790)	12.8310 (2.1230)	12.8320 (2.1190)	12.9780 (2.1300)
		30	2.0100 (0.1900)	0.0360	0.9890 (0.1870)	0.0350	−1.1940 (2.4730)	6.1460	0.2910 (0.0400)	13.1140 (1.1650)	13.1190(1.1660)	13.1460 (1.1670)
		100	1.9940 (0.0960)	0.0090	1.0060 (0.0960)	0.0090	−1.0090 (1.7580)	3.0890	0.2830 (0.0200)	17.2270 (0.7900)	17.2220 (0.7910)	17.2560 (0.7900)
		250	1.9997 (0.0470)	0.0020	0.9998 (0.0470)	0.0020	−0.9880 (1.1150)	1.2420	0.4130 (0.0140)	17.2620 (0.4760)	17.2570 (0.4760)	17.2830 (0.4760)
Mixture of different distributions	Level I	10	2.0014 (0.0588)	0.0035	0.9985 (0.0588)	0.0035	−1.0067 (0.3851)	0.1482	0.9764 (0.0076)	1.0213 (0.1753)	1.0285 (0.1752)	1.0289 (0.1755)
		30	2.0008 (0.0293)	0.0009	0.9993 (0.0294)	0.0009	−0.9991 (0.2217)	0.0491	0.9725 (0.0048)	1.1094 (0.1010)	1.1109 (0.1009)	1.1115 (0.1012)
		100	1.9998 (0.0199)	0.0004	1.0002 (0.0199)	0.0004	−0.9933 (0.1068)	0.0114	0.9740 (0.0023)	1.0089 (0.0463)	1.0093 (0.0461)	1.0101 (0.0463)
		250	2.0000 (0.0114)	0.0001	1.0000 (0.0114)	0.0001	−0.9983 (0.0723)	0.0052	0.9743 (0.0014)	1.0847 (0.0305)	1.0848 (0.0305)	1.0855 (0.0306)
	Level II	10	2.0092 (0.5748)	0.3302	0.9919 (0.5688)	0.3233	−1.1080 (3.8687)	14.9639	0.3259 (0.0855)	10.2512 (1.7637)	10.3166 (1.7653)	10.3241 (1.7699)
		30	1.9982 (0.2841)	0.0806	1.0016 (0.2838)	0.0805	−0.9473 (2.2149)	4.9035	0.2697 (0.0377)	11.0789 (0.9735)	11.0919 (0.9723)	11.0995 (0.9742)
		100	2.0024 (0.1945)	0.0378	0.9972 (0.1946)	0.0378	−1.0134 (1.0434)	1.0877	0.2780 (0.0247)	10.0615 (0.4612)	10.0654 (0.4595)	10.0734 (0.4619)
		250	1.9983 (0.1115)	0.0124	1.0017 (0.1119)	0.0125	−1.0047 (0.7308)	0.5336	0.2762 (0.0136)	10.8351 (0.3004)	10.8354 (0.2993)	10.8430 (0.3008)

Table C2. Results, in different conditions, of the *DSD Model* with $a = 2$, $b = 8$ and $v = 3$.

			Estimated parameters						Goodness of fit measures			
Distribution of microdata	Error level	n	$\overline{a^*}$ (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)
Uniform distribution	Level I	10	1.9915 (0.4520)	0.2042	8.0086 (0.4506)	0.2029	3.0229 (1.4417)	2.0769	0.9924 (0.0027)	1.0225 (0.1930)	1.0216 (0.1913)	1.0271 (0.1942)
		30	1.9847 (0.2701)	0.0731	8.0131 (0.2711)	0.0736	3.0384 (0.7709)	0.5952	0.9911 (0.0016)	1.1585 (0.1068)	1.1565 (0.1058)	1.1637 (0.1076)
		100	2.0008 (0.1323)	0.0175	7.9991 (0.1321)	0.0174	3.0001 (0.3828)	0.1464	0.9907 (0.0008)	1.1436 (0.0526)	1.1419 (0.0522)	1.1473 (0.0530)
		250	1.9983 (0.0942)	0.0089	8.0015 (0.0942)	0.0089	3.0030 (0.2837)	0.0804	0.9905 (0.0005)	1.1598 (0.0334)	1.1579 (0.0330)	1.1640 (0.0337)
	Level II	10	3.0342 (3.2418)	11.5686	7.1862 (3.4119)	12.2918	0.1597 (10.8013)	124.6187	0.5856 (0.0885)	10.4196 (1.8327)	10.4029 (1.8114)	10.4716 (1.8500)
		30	2.3117 (2.1745)	4.8209	7.7527 (2.2562)	5.1464	2.2254 (6.4742)	42.4731	0.5331 (0.0519)	11.6755 (1.1063)	11.6534 (1.0948)	11.7283 (1.1160)
		100	2.0129 (1.2074)	1.4564	7.9954 (1.2206)	1.4884	3.0452 (3.5845)	12.8382	0.5183 (0.0268)	11.4784 (0.5447)	11.4622 (0.5388)	11.5153 (0.5502)
		250	2.0223 (0.8549)	0.7306	7.9790 (0.8556)	0.7317	2.9444 (2.6689)	7.1191	0.5110 (0.0164)	11.6145 (0.3396)	11.5951 (0.3356)	11.6567 (0.3433)
Normal distribution	Level I	10	2.0967 (1.4226)	2.0311	7.9062 (1.4281)	2.0463	2.6628 (5.0032)	25.1205	0.9801 (0.0068)	2.0608 (0.3781)	2.0598 (0.3742)	2.0737 (0.3800)
		30	1.9842 (0.6694)	0.4479	8.0168 (0.6701)	0.4488	3.0522 (2.2418)	5.0233	0.9763 (0.0042)	2.1646 (0.1986)	2.1636 (0.1955)	2.1796 (0.2004)
		100	1.9902 (0.3540)	0.1253	8.0098 (0.3539)	0.1252	3.0282 (1.0801)	1.1663	0.9762 (0.0021)	2.1268 (0.0952)	2.1249 (0.0945)	2.1349 (0.0958)
		250	2.0000 (0.2199)	0.0483	7.9999 (0.2199)	0.0483	2.9990 (0.6745)	0.4545	0.9757 (0.0014)	2.1806 (0.0623)	2.1788 (0.0616)	2.1893 (0.0629)
	Level II	10	4.5615 (4.8027)	29.6046	5.9059 (4.9355)	28.7201	−5.2812 (18.7400)	419.4156	0.3563 (0.0844)	21.2601 (3.6607)	21.2454 (3.6329)	21.3778 (3.6803)
		30	3.5385 (3.9068)	17.6148	6.6568 (3.9765)	17.6010	−1.7030 (13.6088)	207.1313	0.3044 (0.0413)	21.7775 (1.9671)	21.7607 (1.9364)	21.9164 (1.9923)
		100	2.5780 (2.6539)	7.3699	7.4812 (2.7122)	7.6180	1.3196 (8.3472)	72.4307	0.2966 (0.0204)	21.2287 (0.9388)	21.2085 (0.9303)	21.3103 (0.9462)
		250	2.1668 (1.9088)	3.6676	7.8517 (1.9313)	3.7484	2.5374 (5.9072)	35.0738	0.2890 (0.0136)	21.8057 (0.6483)	21.7876 (0.6421)	21.8918 (0.6537)
LogNormal distribution	Level I	10	1.9960 (0.1160)	0.0130	8.0030 (0.1140)	0.0130	2.9920 (1.6020)	2.5630	0.9780 (0.0070)	4.4380 (0.7950)	4.4650 (0.7780)	4.4640 (0.8130)
		30	1.9960 (0.0590)	0.0040	8.0040 (0.0590)	0.0030	3.0430 (0.8550)	0.7320	0.9790 (0.0040)	4.4000 (0.3890)	4.4060 (0.3870)	4.4070 (0.3910)
		100	1.9998 (0.0310)	0.0010	8.0004 (0.0320)	0.0010	2.9760 (0.5950)	0.3540	0.9780 (0.0020)	5.7680 (0.2670)	5.7710 (0.2660)	5.7720 (0.2690)
		250	2.0010 (0.0160)	2.4392E-4	7.9990 (0.0160)	2.4331E-4	3.0090 (0.3840)	0.1470	0.9880 (0.0010)	5.7690 (0.1660)	5.7710 (0.1650)	5.7730 (0.1660)
	Level II	10	2.0340 (1.1180)	1.2490	7.9850 (1.1770)	1.3840	3.6840 (16.3180)	266.4640	0.3300 (0.0900)	44.2560 (7.8380)	44.5340 (7.6590)	44.5010 (8.0260)
		30	1.9990 (0.6390)	0.4080	8.0010 (0.6410)	0.4110	2.9310 (8.7020)	75.6620	0.3190 (0.0430)	44.0600 (3.8910)	44.1120 (3.8660)	44.1440 (3.9140)
		100	2.0040 (0.3280)	0.1070	7.9970 (0.3310)	0.1090	2.7960 (6.1880)	38.2960	0.3150 (0.0230)	57.5170 (2.6890)	57.5550 (2.6750)	57.5670 (2.7040)
		250	2.0130 (0.1530)	0.0240	7.9880 (0.1530)	0.0230	3.0730 (3.5690)	12.7280	0.4550 (0.0160)	57.7450 (1.6470)	57.7650 (1.6380)	57.7840 (1.6560)
Mixture of different distributions	Level I	10	2.0075 (0.1945)	0.0379	7.9930 (0.1941)	0.0377	2.9145 (1.2898)	1.6693	0.9800 (0.0066)	3.3860 (0.5900)	3.4096 (0.5896)	3.4111 (0.5903)
		30	1.9984 (0.0970)	0.0094	8.0016 (0.0969)	0.0094	2.9919 (0.7528)	0.5662	0.9762 (0.0041)	3.6657 (0.3234)	3.6707 (0.3229)	3.6723 (0.3238)
		100	1.9997 (0.0651)	0.0042	8.0002 (0.0650)	0.0042	3.0012 (0.3476)	0.1207	0.9788 (0.0019)	3.3561 (0.1501)	3.3576 (0.1496)	3.3597 (0.1500)
		250	1.9993 (0.0380)	0.0014	8.0006 (0.0379)	0.0014	3.0059 (0.2433)	0.0592	0.9784 (0.0012)	3.6120 (0.1047)	3.6123 (0.1045)	3.6142 (0.1050)
	Level II	10	2.0651 (1.7394)	3.0268	8.0708 (1.9194)	3.6855	2.5964 (12.9825)	168.5408	0.3568 (0.0958)	34.2123 (5.7510)	34.4588 (5.7669)	34.3920 (5.7614)
		30	2.0079 (0.9464)	0.8948	7.9963 (0.9474)	0.8966	2.8864 (7.2998)	53.2464	0.2940 (0.0423)	37.0044 (3.2426)	37.0508 (3.2410)	37.0681 (3.2482)
		100	1.9959 (0.6567)	0.4309	8.0050 (0.6577)	0.4322	3.0427 (3.5225)	12.3973	0.3191 (0.0319)	33.5675 (1.5549)	33.5832 (1.5526)	33.6035 (1.5552)
		250	1.9863 (0.3816)	0.1456	8.0128 (0.3819)	0.1458	3.1761 (2.5117)	6.3335	0.3131 (0.0175)	36.1132 (1.0192)	36.1160 (1.0177)	36.1365 (1.0198)

Table C3. Results, in different conditions, of the *DSD Model* with $a = 6$, $b = 0$ and $v = 2$.

Distribution of microdata	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)
Uniform distribution	Level I	10	5.9060 (0.1897)	0.0448	0.1174 (0.1703)	0.0427	2.3269 (0.5950)	0.4605	0.9924 (0.0025)	0.6374 (0.1106)	0.6364 (0.1097)	0.6404 (0.1113)
		30	5.9553 (0.1025)	0.0125	0.0589 (0.0903)	0.0116	2.1385 (0.2876)	0.1018	0.9914 (0.0015)	0.7057 (0.0654)	0.7044 (0.0646)	0.7087 (0.0660)
		100	5.9784 (0.0499)	0.0030	0.0294 (0.0433)	0.0027	2.0741 (0.1470)	0.0271	0.9914 (0.0008)	0.6886 (0.0326)	0.6875 (0.0323)	0.6908 (0.0328)
		250	5.9836 (0.0353)	0.0015	0.0210 (0.0316)	0.0014	2.0554 (0.1075)	0.0146	0.9912 (0.0005)	0.6962 (0.0213)	0.6951 (0.0210)	0.6987 (0.0215)
	Level II	10	5.1445 (1.7747)	3.8784	1.0960 (1.5941)	3.7397	4.9585 (5.5954)	40.0296	0.5889 (0.0893)	6.3568 (1.1442)	6.3474 (1.1346)	6.3873 (1.1516)
		30	5.4839 (1.0569)	1.3823	0.6388 (0.9542)	1.3176	3.5398 (3.0400)	11.6035	0.5446 (0.0487)	7.0279 (0.6605)	7.0141 (0.6542)	7.0601 (0.6658)
		100	5.7588 (0.5229)	0.3313	0.3181 (0.4604)	0.3130	2.8271 (1.5093)	2.9597	0.5380 (0.0265)	6.8855 (0.3232)	6.8754 (0.3205)	6.9081 (0.3256)
		250	5.8372 (0.3433)	0.1442	0.2115 (0.3058)	0.1382	2.5513 (1.0297)	1.3631	0.5299 (0.0170)	6.9544 (0.2078)	6.9432 (0.2061)	6.9793 (0.2093)
Normal distribution	Level I	10	5.6565 (0.5278)	0.3963	0.3642 (0.5128)	0.3953	3.2455 (1.8674)	5.0351	0.9794 (0.0070)	1.2709 (0.2312)	1.2707 (0.2285)	1.2778 (0.2332)
		30	5.8513 (0.2450)	0.0821	0.1623 (0.2353)	0.0816	2.5023 (0.8316)	0.9432	0.9768 (0.0039)	1.3022 (0.1125)	1.3006 (0.1109)	1.3110 (0.1138)
		100	5.9207 (0.1287)	0.0228	0.0870 (0.1231)	0.0227	2.2503 (0.4030)	0.2249	0.9769 (0.0022)	1.2760 (0.0619)	1.2755 (0.0613)	1.2801 (0.0624)
		250	5.9522 (0.0807)	0.0088	0.0528 (0.0768)	0.0087	2.1498 (0.2471)	0.0834	0.9765 (0.0013)	1.3057 (0.0375)	1.3045 (0.0371)	1.3106 (0.0378)
	Level II	10	3.9562 (2.8522)	12.3041	2.3298 (2.7704)	13.0955	9.7470 (10.9972)	180.8330	0.3557 (0.0842)	12.7916 (2.2538)	12.7806 (2.2253)	12.8635 (2.2764)
		30	4.6760 (2.1116)	6.2071	1.4787 (2.0209)	6.2663	6.7264 (7.2579)	74.9641	0.3081 (0.0412)	13.0324 (1.1632)	13.0133 (1.1503)	13.1227 (1.1724)
		100	5.2244 (1.2684)	2.2087	0.8513 (1.2124)	2.1932	4.3969 (3.9058)	20.9856	0.2994 (0.0222)	12.7847 (0.5737)	12.7784 (0.5688)	12.8257 (0.5780)
		250	5.5403 (0.8005)	0.8514	0.5077 (0.7634)	0.8399	3.4450 (2.4552)	8.1101	0.2948 (0.0139)	13.0683 (0.3792)	13.0559 (0.3763)	13.1177 (0.3817)
LogNormal distribution	Level I	10	5.9960 (0.0700)	0.0050	0.0270 (0.0390)	0.0020	2.0750 (0.8900)	0.7960	0.9850 (0.0050)	2.5890 (0.4200)	2.5730 (0.4250)	2.6370 (0.4150)
		30	5.9960 (0.0350)	0.0010	0.0160 (0.0230)	0.0010	2.0370 (0.4870)	0.2380	0.9840 (0.0030)	2.6220 (0.2280)	2.6210 (0.2290)	2.6320 (0.2280)
		100	5.9990 (0.0180)	3.2272E-4	0.0070 (0.0110)	1.6840E-4	2.0130 (0.3590)	0.1290	0.9840 (0.0010)	3.4480 (0.1500)	3.4440 (0.1500)	3.4580 (0.1500)
		250	5.9997 (0.0090)	8.7406E-5	0.0040 (0.0060)	4.5620E-5	2.0060 (0.2200)	0.0480	0.9920 (4.7992E-4)	3.4500 (0.1010)	3.4470 (0.1010)	3.4570 (0.1000)
	Level II	10	5.9400 (0.7110)	0.5080	0.2890 (0.4030)	0.2460	2.8770 (8.9940)	81.5750	0.4020 (0.0930)	26.0160 (4.0850)	25.8460 (4.1260)	26.5160 (4.0510)
		30	5.9710 (0.3400)	0.1170	0.1480 (0.2090)	0.0650	2.1640 (4.8970)	23.9830	0.3900 (0.0480)	26.2500 (2.2940)	26.2290 (2.2960)	26.3410 (2.2920)
		100	5.9930 (0.1810)	0.0330	0.0760 (0.1120)	0.0180	2.1140 (3.5020)	12.2680	0.3880 (0.0250)	34.4440 (1.5240)	34.4030 (1.5240)	34.5440 (1.5230)
		250	5.9996 (0.0900)	0.0080	0.0370 (0.0510)	0.0040	2.1190 (2.1670)	4.7060	0.5450 (0.0160)	34.4870 (0.9780)	34.4540 (0.9780)	34.5600 (0.9770)
Mixture of different distributions	Level I	10	5.9925 (0.1079)	0.0117	0.0494 (0.0658)	0.0068	2.0804 (0.7350)	0.5461	0.9852 (0.0047)	2.0509 (0.3404)	2.0540 (0.3402)	2.0673 (0.3426)
		30	5.9934 (0.0512)	0.0027	0.0250 (0.0334)	0.0017	2.0395 (0.4276)	0.1843	0.9814 (0.0032)	2.2197 (0.1950)	2.2204 (0.1949)	2.2235 (0.1950)
		100	5.9989 (0.0386)	0.0015	0.0160 (0.0224)	0.0008	2.0107 (0.2064)	0.0427	0.9855 (0.0013)	2.0172 (0.0912)	2.0175 (0.0912)	2.0190 (0.0911)
		250	5.9993 (0.0215)	0.0005	0.0088 (0.0134)	0.0003	2.0135 (0.1387)	0.0194	0.9847 (0.0008)	2.1642 (0.0598)	2.1642 (0.0597)	2.1653 (0.0598)
	Level II	10	5.9118 (1.0729)	1.1577	0.5004 (0.6533)	0.6767	2.9336 (7.1752)	52.3032	0.4163 (0.1022)	20.4358 (3.3003)	20.4698 (3.2934)	20.5977 (3.3195)
		30	5.9389 (0.4937)	0.2472	0.2382 (0.3215)	0.1600	2.4506 (4.2470)	18.2219	0.3490 (0.0492)	22.2670 (1.9401)	22.2732 (1.9403)	22.3021 (1.9417)
		100	6.0041 (0.3893)	0.1514	0.1557 (0.2203)	0.0727	2.1594 (2.0098)	4.0607	0.4069 (0.0374)	20.1143 (0.8733)	20.1180 (0.8731)	20.1319 (0.8726)
		250	5.9735 (0.2286)	0.0529	0.1054 (0.1419)	0.0312	2.1551 (1.4054)	1.9973	0.3904 (0.0224)	21.6308 (0.6286)	21.6308 (0.6283)	21.6415 (0.6284)

Table C4. Results, in different conditions, of the *DSD Model* with $a_1 = 2$, $b_1 = 1$, $a_2 = 0.5$, $b_2 = 3$, $a_3 = 1.5$, $b_3 = 1$ and $v = -1$.

Distribution of microdata	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)
Uniform distribution	Level I	10	2.0180 (1.0040)	1.0080	1.0290 (0.8540)	0.7300	0.6700 (0.7690)	0.6200	3.0440 (1.1870)	1.4090	1.3570 (0.8650)	0.7680	1.0160 (0.8580)	0.7360
		30	2.0300 (0.5480)	0.3010	0.9860 (0.5230)	0.2730	0.5770 (0.5320)	0.2890	3.0060 (0.6660)	0.4430	1.5120 (0.4670)	0.2180	0.9360(0.4450)	0.2020
		100	1.9951 (0.2630)	0.0691	1.0055 (0.2627)	0.0690	0.4964 (0.2665)	0.0710	3.0023 (0.2848)	0.0811	1.4974 (0.2459)	0.0604	1.0031 (0.2466)	0.0607
		250	1.9969 (0.1752)	0.0307	0.9981 (0.1779)	0.0316	0.5008 (0.1813)	0.0329	3.0005 (0.1812)	0.0328	1.5077 (0.1766)	0.0312	0.9929 (0.1779)	0.0317
	Level II	10	3.7313 (5.5866)	34.17618	3.2144 (5.3537)	33.5365	1.8612 (3.1480)	11.7530	2.1846 (3.3425)	11.8259	0.68015 (1.4447)	2.7573	0.6521 (1.4055)	2.0944
		30	2.5566 (3.5936)	13.2105	2.1629 (3.4342)	13.1344	1.2743 (2.2707)	5.7504	2.3932 (3.0064)	9.3976	1.0544 (1.5866)	2.7132	0.9354 (1.5036)	2.2627
		100	2.2417 (2.4698)	6.1521	1.3448 (1.9563)	3.9421	1.1146 (1.7354)	3.3862	2.6963 (2.5101)	6.3863	1.1442 (1.3671)	1.9938	1.0365 (1.3062)	1.7058
		250	1.9969 (0.1752)	0.0307	0.9981 (0.1779)	0.0316	0.5008 (0.1813)	0.0329	3.0005 (0.1812)	0.0328	1.5077 (0.1766)	0.0312	0.9929 (0.1779)	0.0317
Normal distribution	Level I	10	2.1101 (2.0446)	4.1884	1.7887 (1.9431)	4.3939	0.7246 (0.9897)	1.0289	2.2317 (1.4262)	2.6222	1.5038 (1.4331)	2.0517	1.3207 (1.3280)	1.8647
		30	2.0907 (1.5439)	2.3894	1.2460 (1.3024)	1.7552	0.6695 (0.7621)	0.6089	2.8822 (0.8897)	0.8047	1.3987 (0.8784)	0.7811	0.9657 (0.8482)	0.7200
		100	1.9522 (0.9401)	0.8851	1.0975 (0.8524)	0.7353	0.5731 (0.5403)	0.2970	2.9575 (0.6137)	0.3780	1.4726 (0.6094)	0.3718	0.9849 (0.5983)	0.3578
		250	1.9973 (0.6053)	0.3661	1.0128 (0.5843)	0.3412	0.5265 (0.3428)	0.1181	2.9793 (0.3613)	0.1308	1.4947 (0.4133)	0.1707	0.9967 (0.4161)	0.1730
	Level II	10	6.1020 (8.5801)	90.3710	5.8700 (8.7266)	99.7942	0.7455 (1.7663)	3.1770	0.9229 (2.0136)	8.3647	1.0657 (1.9449)	3.9676	1.1209 (1.9571)	3.8409
		30	5.1883 (7.4166)	65.1165	4.6084 (7.2786)	65.9450	0.9933 (1.7250)	3.2162	1.4892 (2.1051)	6.7099	0.9281 (1.7970)	3.5529	0.9547 (1.8582)	3.4514
		100	3.0643 (4.1770)	18.5624	3.1008 (4.3948)	23.7084	1.1631 (1.8017)	3.6825	1.6705 (2.0395)	5.9229	1.1911 (1.7758)	3.2458	1.2351 (1.7757)	3.2052
		250	2.6133 (3.2783)	11.1127	2.2798 (2.9082)	10.0869	1.0509 (1.5058)	2.5687	2.0568 (1.8915)	4.4638	1.3016 (1.5597)	2.4695	1.1420 (1.4702)	2.1794
LogNormal distribution	Level I	10	1.9363 (0.5625)	0.3201	0.9704 (0.6123)	0.3754	0.5443 (0.5019)	0.2536	3.0296 (0.5918)	0.3508	1.5045 (0.1457)	0.0212	1.0010 (0.1396)	0.0195
		30	2.0029 (0.1942)	0.0377	0.9965 (0.1915)	0.0367	0.5089 (0.2694)	0.0726	2.9934 (0.2722)	0.0741	1.4986 (0.0372)	0.0014	1.0010 (0.0372)	0.0014
		100	1.9966 (0.0878)	0.0077	1.0035 (0.0871)	0.0076	0.5040 (0.1320)	0.0174	2.9960 (0.1317)	0.0173	1.5005 (0.0236)	0.0006	0.9995 (0.0232)	0.0005
		250	1.9982 (0.0331)	0.0011	1.0017 (0.0325)	0.0011	0.5057 (0.0869)	0.0076	2.9943 (0.0864)	0.0075	1.4991 (0.0112)	0.0001	1.0009 (0.0111)	0.0001
	Level II	10	1.4470 (2.1128)	4.7653	1.5148 (2.0260)	4.3657	1.7741 (2.4277)	7.5113	2.9833 (3.1078)	9.6493	1.3239 (0.8571)	0.7650	0.8539 (0.7641)	0.6046
		30	2.0171 (1.6541)	2.7336	1.2535 (1.4276)	2.1003	1.3990 (1.8138)	4.0949	3.1881 (2.5762)	6.6655	1.3660 (0.3212)	0.1210	0.9130 (0.3405)	0.1234
		100	1.8891 (0.8364)	0.7111	1.0523 (0.7637)	0.5854	0.7938 (0.9088)	0.9115	3.0153 (1.2865)	1.6536	1.4755 (0.2206)	0.0492	0.9671 (0.2370)	0.0572
		250	1.9937 (0.3306)	0.1092	0.9730 (0.3306)	0.1099	0.6279 (0.6534)	0.4429	3.0213 (0.8386)	0.7031	1.4809 (0.1028)	0.0109	0.9974 (0.1168)	0.0136
Mixture of different distributions	Level I	10	1.9777 (0.2846)	0.0814	1.0031 (0.3050)	0.0929	0.5417 (0.3659)	0.1355	2.9683 (0.3812)	0.1462	1.5059 (0.1068)	0.0114	0.9964 (0.1078)	0.0116
		30	1.9929 (0.1564)	0.0245	1.0069 (0.1552)	0.0241	0.4998 (0.0778)	0.0060	3.0000 (0.0766)	0.0059	1.5020 (0.0306)	0.0009	0.9981 (0.0304)	0.0009
		100	1.9989 (0.1346)	0.0181	1.0002 (0.1347)	0.0181	0.4981 (0.0558)	0.0031	3.0019 (0.0555)	0.0031	1.4990 (0.0233)	0.0005	1.0012 (0.0233)	0.0005
		250	2.0001 (0.0506)	0.0026	1.0000 (0.0511)	0.0026	0.5003 (0.0367)	0.0013	2.9991 (0.0365)	0.0013	1.5017 (0.0188)	0.0004	0.9986 (0.0187)	0.0004
	Level II	10	2.5108 (2.8676)	8.4760	1.9512 (2.5185)	7.2411	1.1774 (1.4834)	2.6572	2.0922 (1.7827)	3.9989	1.2844 (1.0560)	1.1606	0.9198 (0.9291)	0.8688
		30	2.0814 (1.6958)	2.8794	1.1883 (1.3511)	1.8592	0.5467 (0.6034)	0.3659	3.0429 (1.0352)	1.0723	1.4411 (0.3326)	0.1140	0.9757 (0.2832)	0.0807
		100	2.0760 (1.3360)	1.7890	1.1120 (1.1021)	1.2260	0.4796 (0.4623)	0.2140	3.0397 (0.6322)	0.4008	1.4778 (0.2476)	0.0618	0.9715 (0.2320)	0.0546
		250	1.9814 (0.5260)	0.2767	0.9926 (0.5340)	0.2850	0.5079 (0.3562)	0.1268	3.0186 (0.4099)	0.1682	1.5000 (0.1890)	0.0357	0.9940 (0.1897)	0.0360

Table C5. Results, in different conditions, of the *DSD 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 C4).

Distribution of microdata	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)
Uniform distribution	Level I	10	−1.5390 (15.2030)	231.1920	0.9930 (0.0030)	1.7990 (0.3940)	1.7960 (0.3880)	1.8180 (0.3980)
		30	−1.8260 (10.4680)	110.1520	0.9900 (0.0020)	2.1080 (0.2070)	2.1020 (0.2040)	2.1240 (0.2100)
		100	−0.9182 (4.5996)	21.1414	0.9894 (0.0011)	2.2340 (0.1132)	2.2265 (0.1107)	2.2520 (0.1154)
		250	−1.0406 (3.0068)	9.0336	0.9892 (0.0007)	2.2584 (0.0689)	2.2505 (0.0676)	2.2759 (0.0700)
	Level II	10	−16.2481 (46.2304)	2367.6185	0.5700 (0.0923)	20.2636 (3.7344)	20.1923 (3.6756)	20.4681 (3.7797)
		30	−10.0067 (37.7315)	1503.3633	0.5128 (0.0486)	21.6993 (2.0073)	21.6227 (1.9667)	21.8825 (2.0436)
		100	−7.1346 (31.9188)	1055.4226	0.4871 (0.0272)	22.4286 (1.1417)	22.3489 (1.1177)	22.6125 (1.1628)
		250	−4.3557 (24.3173)	601.9989	0.4814 (0.0165)	22.5629 (0.6960)	22.4852 (0.6838)	22.7374 (0.7067)
Normal distribution	Level I	10	−1.7818 (12.8664)	165.9908	0.9781 (0.0079)	5.2894 (0.9987)	5.3373 (0.9797)	5.4002 (1.0084)
		30	−1.4290 (8.8043)	77.6226	0.9742 (0.0048)	5.6637 (0.5404)	5.6703 (0.5264)	5.7603 (0.5490)
		100	−1.0779 (5.4809)	30.0168	0.9726 (0.0027)	5.8658 (0.2956)	5.8589 (0.2884)	5.9399 (0.3007)
		250	−1.1130 (3.7736)	14.2387	0.9720 (0.0017)	5.9871 (0.1851)	5.9760 (0.1801)	6.0580 (0.1890)
	Level II	10	−4.9998 (34.4850)	1204.0237	0.3298 (0.0830)	55.0942 (9.8672)	55.0224 (9.6353)	55.9747 (10.0384)
		30	−4.8885 (26.6366)	723.9217	0.2824 (0.0412)	57.9609 (5.4172)	57.8505 (5.2743)	58.7480 (5.5337)
		100	−3.6457 (18.2035)	338.0369	0.2651 (0.0193)	59.1260 (2.6834)	59.0011 (2.6232)	59.8200 (2.7333)
		250	−3.4715 (15.4462)	244.4548	0.2597 (0.0124)	60.0610 (1.8103)	59.9207 (1.7640)	60.7524 (1.8495)
LogNormal distribution	Level I	10	−1.0437 (6.6573)	44.2771	0.9749 (0.0082)	5.5943 (0.9805)	5.6663 (0.9673)	5.6887 (0.9768)
		30	−1.1024 (4.4514)	19.8058	0.9684 (0.0056)	8.6758 (0.8009)	8.6971 (0.7964)	8.7223 (0.8031)
		100	−1.0438 (2.2897)	5.2393	0.9671 (0.0029)	9.1786 (0.4271)	9.1822 (0.4252)	9.2070 (0.4275)
		250	−1.0915 (1.4540)	2.1204	0.9687 (0.0017)	9.6863 (0.2736)	9.6835 (0.2731)	9.7115 (0.2741)
	Level II	10	−8.4105 (38.0447)	1500.8634	0.3281 (0.0841)	56.8560 (9.9320)	57.0863 (9.9399)	57.5179 (9.8937)
		30	−6.1572 (34.0822)	1187.0296	0.2565 (0.0400)	86.4383 (7.5017)	86.6241 (7.4774)	86.8222 (7.5162)
		100	−3.4937 (19.7820)	397.1544	0.2333 (0.0185)	91.7504 (4.1788)	1.7881 (4.1741)	92.0029 (4.1857)
		250	−1.7408 (12.8365)	165.1590	0.2393 (0.0117)	96.6891 (2.7337)	96.6724 (2.7232)	96.9214 (2.7415)
Mixture of different distributions	Level I	10	−1.5361 (5.6379)	32.0416	0.9784 (0.0070)	4.5320 (0.7769)	4.6074 (0.7754)	4.6093 (0.7776)
		30	−0.9821 (1.3223)	1.7471	0.9796 (0.0035)	5.7982 (0.5045)	5.8124 (0.5045)	5.8170 (0.5059)
		100	−1.0786 (0.9928)	0.9909	0.9774 (0.0020)	6.4328 (0.2873)	6.4384 (0.2867)	6.4435 (0.2875)
		250	−1.0355 (0.5313)	0.2833	0.9781 (0.0013)	4.9282 (0.1466)	4.9280 (0.1460)	4.9350 (0.1471)
	Level II	10	−13.0526 (26.5748)	850.7790	0.3640 (0.0834)	45.9825 (7.7313)	46.5534 (7.6858)	46.5829 (7.7703)
		30	−0.5398 (13.3243)	177.5710	0.3472 (0.0660)	57.8498 (5.2121)	57.9755 (5.2090)	58.0411 (5.2300)
		100	−0.7405 (7.2973)	53.2646	0.2976 (0.0316)	57.0197 (2.7292)	57.0658 (2.7256)	57.1239 (2.7347)
		250	−0.9364 (5.1880)	26.8924	0.3125 (0.0181)	49.2312 (1.3779)	49.2283 (1.3750)	49.3006 (1.3832)

Table C6. Results, in different conditions, of the *DSD Model* with $a_1 = 6$, $b_1 = 0$, $a_2 = 2$, $b_2 = 8$, $a_3 = 10$, $b_3 = 5$ and $v = 3$.

Distribution of microdata	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)
Uniform distribution	Level I	10	6.3715 (4.1743)	17.5452	1.5427 (2.2722)	7.5375	2.8373 (3.3580)	11.9658	7.9944 (4.7880)	22.9021	9.3881 (3.8229)	14.9742	4.6417 (3.6571)	13.4892
		30	6.2297 (2.5382)	6.4889	0.9146 (1.3949)	2.7804	2.2264 (2.0872)	4.4033	8.1975 (2.8075)	7.9129	9.6613 (1.9165)	3.7840	4.7633 (1.8954)	3.6449
		100	6.2787 (1.4522)	2.1845	0.4560 (0.6806)	0.6707	1.9767 (1.1170)	1.2471	7.9660 (1.1831)	1.3995	9.8788 (1.1363)	1.3045	4.9499 (1.1680)	1.3655
		250	6.1695 (0.9477)	0.9260	0.3016 (0.4384)	0.2830	1.9678 (0.7684)	0.5908	7.9975 (0.7851)	0.6158	9.9601 (0.7612)	0.5804	4.9273 (0.7345)	0.5442
	Level II	10	15.2212 (24.1391)	667.1439	12.4554 (21.1855)	603.5122	8.1637 (13.5112)	220.3613	9.2567 (14.4161)	209.1941	3.1279 (6.2838)	86.6726	2.8147 (6.1209)	42.2040
		30	10.9816 (15.8180)	274.7761	8.0165 (13.1785)	237.7624	5.8666 (10.0152)	115.1538	7.8857 (11.3599)	128.9303	5.1868 (6.9554)	71.4964	4.1814 (6.4541)	42.2833
		100	8.8226 (10.9566)	127.8936	5.0310 (8.0127)	89.4498	5.1229 (7.6110)	67.6221	8.0074 (9.4182)	88.6130	6.5383 (6.2197)	50.6291	4.5743 (5.6959)	32.5919
		250	8.0084 (7.9529)	67.2192	2.9439 (5.1385)	35.0445	3.8007 (5.3914)	32.2800	8.0732 (7.4319)	55.1833	8.0554 (5.7973)	37.3570	4.5135 (4.9062)	24.2834
Normal distribution	Level I	10	6.1899 (6.4202)	41.2144	4.2460 (5.6656)	50.0951	2.4327 (3.4550)	12.1122	5.8788 (4.8876)	28.3644	8.8327 (6.8724)	48.5451	6.7459 (6.3410)	43.2159
		30	6.7039 (5.4925)	30.6331	2.5740 (3.8666)	21.5608	2.6683 (2.7546)	8.0269	7.6296 (3.2566)	10.7319	9.0163 (3.7933)	15.3425	4.8627 (3.7216)	13.8556
		100	5.5686 (3.1611)	10.1686	1.5645 (2.1965)	7.2676	2.3975 (2.1930)	4.9625	7.6173 (2.4409)	6.0984	9.8201 (2.4105)	5.8371	4.8835 (2.3846)	5.6942
		250	5.7034 (2.0794)	4.4073	0.9581 (1.4121)	2.9100	2.0421 (1.4118)	1.9930	7.8796 (1.4782)	2.1973	9.8789 (1.5967)	2.5615	5.0331 (1.5963)	2.5467
	Level II	10	19.5384 (30.1263)	1089.9741	23.0826 (33.6988)	1667.2778	3.3160 (7.7868)	62.3047	3.4185 (7.7595)	81.1396	5.0435 (8.2166)	92.0114	4.4917 (8.0309)	64.6890
		30	20.0250 (30.6497)	1135.1646	18.3701 (29.0715)	1181.7682	4.0303 (7.2559)	56.7183	5.1479 (7.8904)	70.3302	4.3349 (7.6641)	90.7732	3.6631 (7.3326)	55.4998
		100	12.5434 (17.1361)	336.1673	10.7856 (16.5852)	391.1231	3.6611 (6.2046)	41.2185	4.7456 (6.9110)	58.3060	6.9008 (8.2102)	76.9455	5.5393 (7.5360)	57.0254
		250	8.0842 (10.4216)	112.8460	6.8109 (9.9712)	145.7133	3.7954 (5.4784)	33.2067	5.8792 (6.5608)	47.4994	7.6233 (7.7100)	65.0334	5.5481 (6.7991)	46.4825
LogNormal distribution	Level I	10	5.7541 (2.4136)	5.8802	1.0681 (1.5536)	3.5521	2.3909 (2.1968)	4.9739	7.3313 (2.1254)	4.9598	9.9844 (0.6076)	0.3690	4.8485 (0.4659)	0.2398
		30	5.9884 (0.9078)	0.8234	0.3502 (0.5150)	0.3876	2.0275 (1.3054)	1.7032	7.9665 (1.4072)	1.9795	9.9868 (0.1773)	0.0316	4.9659 (0.1620)	0.0274
		100	5.9827 (0.3918)	0.1537	0.1607 (0.2320)	0.0796	1.9998 (0.6186)	0.3822	7.9657 (0.6116)	0.3749	10.0043 (0.1119)	0.0125	4.9749 (0.0968)	0.0100
		250	6.0078 (0.1557)	0.0243	0.0593 (0.0846)	0.0107	2.0104 (0.4052)	0.1641	7.9665 (0.3962)	0.1580	9.9985 (0.0528)	0.0028	4.9951 (0.0503)	0.0026
	Level II	10	6.3221 (9.9316)	98.6420	3.7750 (6.6825)	58.8611	6.6268 (10.0238)	121.7837	8.9186 (10.7192)	115.6308	8.4787 (4.1420)	19.4531	3.6631 (3.0099)	10.8378
		30	6.6484 (6.8399)	47.1579	2.9793 (4.6593)	30.5637	6.2718 (8.4137)	88.9684	10.4564 (10.9321)	125.4248	9.2326 (1.5082)	2.8613	4.1608 (1.5184)	3.0075
		100	5.6426 (3.7238)	13.9806	1.5381 (2.3138)	7.7138	3.3822 (4.0155)	18.0189	8.1756 (5.5977)	31.3334	9.8158 (1.0184)	1.0701	4.6169 (0.9733)	1.0931
		250	5.8846 (1.5230)	2.3305	0.6164 (0.8660)	1.1292	2.8735 (2.9117)	9.2327	7.6140 (3.6918)	13.7646	9.8951 (0.4639)	0.2260	4.9371 (0.5056)	0.2594
Mixture of different distributions	Level I	10	6.3798 (1.7457)	3.1888	0.5378 (0.7632)	0.8710	2.0761 (1.5963)	2.5514	7.7189 (1.4624)	2.2156	9.9191 (0.4230)	0.1853	4.8631 (0.4800)	0.2489
		30	6.2112 (0.9328)	0.9138	0.2791 (0.3972)	0.2355	1.9003 (0.3939)	0.1649	7.8800 (0.3511)	0.1375	9.9782 (0.1498)	0.0229	4.9773 (0.1198)	0.0149
		100	6.2074 (0.7713)	0.6373	0.2093 (0.3071)	0.1380	1.9227 (0.2984)	0.0949	7.9114 (0.2143)	0.0538	9.9797 (0.1045)	0.0113	4.9828 (0.0934)	0.0090
		250	6.0513 (0.2608)	0.0706	0.0777 (0.1154)	0.0193	1.9801 (0.1590)	0.0256	7.9700 (0.1287)	0.0174	9.9931 (0.0754)	0.0057	4.9905 (0.0744)	0.0056
	Level II	10	11.3412 (12.2949)	179.5422	6.1341 (9.3815)	125.5513	3.6998 (5.0037)	27.9012	4.9059 (5.3881)	38.5762	8.6930 (4.8385)	25.0960	3.5144 (3.6269)	15.3484
		30	8.0850 (7.8615)	66.0883	3.1629 (5.1847)	36.8584	1.9189 (2.5902)	6.7091	7.4338 (4.7105)	22.4874	9.4406 (1.6407)	3.0022	4.6134 (1.2473)	1.7038
		100	7.0340 (5.9841)	36.8431	2.4709 (3.9780)	21.9146	1.8423 (1.8664)	3.5047	7.4722 (2.7343)	7.7474	9.7925 (1.1403)	1.3421	4.6403 (1.0608)	1.2536
		250	6.4910 (2.5077)	6.5234	0.7633 (1.1560)	1.9177	1.8327 (1.4102)	2.0148	7.8260 (1.4520)	2.1366	9.8569 (0.7649)	0.6049	4.9013 (0.7283)	0.5396

Table C7. Results, in different conditions, of the *DSD Model* with $a_1 = 6$, $b_1 = 0$, $a_2 = 2$, $b_2 = 8$, $a_3 = 10$, $b_3 = 5$ and $v = 3$ (continuation of the Table C6).

Distribution of microdata	Degree of linearity	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)
Uniform distribution	Level I	10	−1.2255 (62.8274)	3961.1896	0.9929 (0.0030)	7.6060 (1.6619)	7.6050 (1.6332)	7.6754 (1.6788)
		30	4.0446 (42.8217)	1832.9585	0.9904 (0.0018)	8.8441 (0.8493)	8.8182 (0.8333)	8.9179 (0.8628)
		100	3.3419 (19.8125)	392.2586	0.9893 (0.0011)	9.3806 (0.4800)	9.3452 (0.4707)	9.4612 (0.4880)
		250	3.3371 (13.0226)	169.5329	0.9892 (0.0007)	9.4723 (0.2946)	9.4386 (0.2895)	9.5480 (0.2991)
	Level II	10	−16.5829 (198.6805)	39817.9727	0.5718 (0.0939)	85.3536 (15.8259)	85.1213 (15.5336)	86.1528 (16.0521)
		30	−13.5004 (151.7463)	23276.1861	0.5129 (0.0493)	90.3790 (8.5273)	90.0812 (8.3531)	91.1389 (8.6806)
		100	−10.6807 (127.3979)	16401.1455	0.4864 (0.0258)	94.1281 (4.5080)	93.7650 (4.4251)	94.9425 (4.5789)
		250	−5.0230 (99.8573)	10025.8686	0.4792 (0.0169)	94.9141 (3.0191)	94.5741 (2.9494)	95.6737 (3.0814)
Normal distribution	Level I	10	11.1307 (50.4357)	2607.3210	0.9781 (0.0079)	20.5721 (3.9492)	20.7540 (3.8554)	20.9982 (4.0038)
		30	5.3845 (32.4760)	1059.3199	0.9740 (0.0050)	22.3414 (2.2300)	22.3750 (2.1718)	22.7271 (2.2632)
		100	3.6874 (20.7385)	430.1263	0.9724 (0.0027)	23.1175 (1.1568)	23.0805 (1.1248)	23.4257 (1.1814)
		250	4.8888 (14.0520)	200.8291	0.9720 (0.0016)	23.4521 (0.6825)	23.4066 (0.6645)	23.7289 (0.6969)
	Level II	10	17.5905 (132.8516)	17844.7831	0.3207 (0.0809)	220.0315 (38.6167)	219.8830 (37.7432)	223.1061 (39.3400)
		30	8.6759 (106.5402)	11371.6744	0.2804 (0.0407)	229.0872 (20.9494)	228.5872 (20.4949)	232.2221 (21.3402)
		100	6.1545 (74.9246)	5618.0313	0.2654 (0.0201)	231.8536 (11.1344)	231.2495 (10.8614)	234.7588 (11.3699)
		250	7.8800 (60.6098)	3693.6885	0.2599 (0.0121)	234.9127 (6.9491)	234.3231 (6.7804)	237.6361 (7.0973)
LogNormal distribution	Level I	10	−2.9196 (26.5045)	736.8273	0.9747 (0.0081)	24.0000 (4.1075)	24.2282 (4.1001)	24.4583 (4.0913)
		30	3.0677 (21.2563)	451.3850	0.9704 (0.0050)	41.8798 (3.6441)	41.9553 (3.6286)	42.1200 (3.6478)
		100	3.0176 (10.6110)	112.4817	0.9688 (0.0027)	41.9305 (1.8795)	41.9239 (1.8738)	42.0868 (1.8811)
		250	2.6938 (6.8659)	47.1876	0.9710 (0.0016)	44.7719 (1.3077)	44.7508 (1.3052)	44.8917 (1.3085)
	Level II	10	−14.9940 (146.8749)	21874.4360	0.3218 (0.0847)	245.4011 (42.9161)	245.7950 (42.7959)	249.2247 (43.0330)
		30	−8.9582 (151.7535)	23149.0842	0.2626 (0.0392)	422.3805 (36.5826)	422.8704 (36.6250)	424.4867 (36.5567)
		100	−5.6890 (85.9342)	7452.8067	0.2442 (0.0192)	418.3204 (19.8923)	418.2955 (19.8783)	419.7091 (19.8771)
		250	−3.9669 (57.4979)	3351.2388	0.2526 (0.0116)	447.6666 (12.6391)	447.4331 (12.6195)	448.8317 (12.6537)
Mixture of different distributions	Level I	10	0.5223 (22.9902)	534.1620	0.9790 (0.0069)	19.2216 (3.3664)	19.4548 (3.3612)	19.6036 (3.3939)
		30	2.7570 (5.6917)	32.4220	0.9778 (0.0039)	25.6419 (2.3166)	25.6867 (2.3155)	25.7347 (2.3272)
		100	2.9845 (3.3438)	11.1699	0.9745 (0.0022)	24.6052 (1.1158)	24.6234 (1.1137)	24.6502 (1.1164)
		250	2.8976 (1.9650)	3.8679	0.9775 (0.0013)	19.6451 (0.5761)	19.6421 (0.5748)	19.6735 (0.5772)
	Level II	10	−33.0838 (93.8292)	10097.1450	0.3706 (0.0906)	195.4006 (34.3356)	197.9795 (34.4766)	197.6777 (34.2227)
		30	2.2830 (57.2783)	3278.0417	0.3271 (0.0525)	257.4014 (23.6065)	257.9902 (23.6029)	258.2033 (23.6690)
		100	1.0807 (30.5694)	937.2343	0.2853 (0.0224)	246.3766 (11.3368)	246.5810 (11.3199)	246.7886 (11.3522)
		250	3.1898 (20.4726)	418.7453	0.3056 (0.0148)	196.5597 (5.6251)	196.5302 (5.6124)	196.8411 (5.6372)

APPENDIX D: OBSERVED AND PREDICTED HISTOGRAMS OF THE EXPERIMENTS PRESENTED IN SECTION 4.2.

Example of Section 4.2.1

In Tables D1 and D2, we present the observed histograms of each observation of the histogram-valued variable Y, the histograms $H_{\widehat{Y}_{DSD}(j)}$ predicted using the *DSD Model*, the histograms $H_{\widehat{Y}_{CM}(j)}$ predicted using the *Center Model*, the histograms $H_{\widehat{Y}_{BD}(j)}$ predicted using the *Billard and Diday Model*, both proposed by Billard and Diday [3], and the histograms $H_{\widehat{Y}_{VI}(j)}$ predicted with the Verde and Irpino model [15].

Table D1. Observed and predicted histograms (using different methods) of the hematocrit values shown in *Table 4* (part1: patients 1 to 5).

Patient	Distributions of the hematocrit values
$H_{Y(1)}$	{[33.29; 35.41) , 0.3; [35.41; 36.11) , 0.1; [36.11; 36.82) , 0.1; [36.82; 37.52) , 0.1; [37.52; 38.04) , 0.1; [38.04; 39.61] , 0.3}
$H_{\widehat{Y}_{DSD}(1)}$	{[33.84; 35.70) , 0.3; [35.70; 36.32) , 0.1; [36.32; 36.73) , 0.1; [36.73; 37.13) , 0.1; [37.13; 37.56) , 0.1; [37.56; 38.85] , 0.3}
$H_{\widehat{Y}_{CM}(1)}$	{[34.33; 35.87) , 0.3; [35.87; 36.38) , 0.1; [36.38; 36.70) , 0.1; [36.70; 37.02) , 0.1; [37.02; 37.35) , 0.1; [37.35; 38.31] , 0.3}
$H_{\widehat{Y}_{BD}(1)}$	{[35.12; 36.51) , 0.3; [36.51; 36.97) , 0.1; [36.97; 37.23) , 0.1; [37.23; 37.55) , 0.1; [37.55; 37.84) , 0.1; [37.84; 38.71] , 0.3}
$H_{\widehat{Y}_{VI}(1)}$	{[33.79; 35.70) , 0.3; [35.70; 36.34) , 0.1; [36.34; 36.73) , 0.1; [36.73; 37.13) , 0.1; [37.13; 37.53) , 0.1; [37.53; 38.73] , 0.3}
$H_{Y(2)}$	{[36.69; 39.11) , 0.3; [39.11; 39.97) , 0.1; [39.97; 40.83) , 0.1; [40.83; 41.69) , 0.1; [41.69; 42.54) , 0.1; [42.54; 45.12] , 0.3}
$H_{\widehat{Y}_{DSD}(2)}$	{[35.16; 38.04) , 0.3; [38.04; 39.00) , 0.1; [39.00; 39.96) , 0.1; [39.96; 40.67) , 0.1; [40.67; 41.38) , 0.1; [41.38; 43.51] , 0.3}
$H_{\widehat{Y}_{CM}(2)}$	{[36.00; 38.37) , 0.3; [38.37; 39.16) , 0.1; [39.16; 39.95) , 0.1; [39.95; 40.49) , 0.1; [40.49; 41.03) , 0.1; [41.03; 42.64] , 0.3}
$H_{\widehat{Y}_{BD}(2)}$	{[36.63; 38.76) , 0.3; [38.76; 39.47) , 0.1; [39.47; 40.18) , 0.1; [40.18; 40.67) , 0.1; [40.67; 41.15) , 0.1; [41.15; 42.60] , 0.3}
$H_{\widehat{Y}_{VI}(2)}$	{[35.13; 38.06) , 0.3; [38.06; 39.04) , 0.1; [39.04; 40.02) , 0.1; [40.02; 40.69) , 0.1; [40.69; 41.36) , 0.1; [41.36; 43.35] , 0.3}
$H_{Y(3)}$	{[36.69; 40.26) , 0.3; [40.26; 41.45) , 0.1; [41.45; 42.64) , 0.1; [42.64; 43.85) , 0.1; [43.85; 45.06) , 0.1; [45.06; 48.68] , 0.3}
$H_{\widehat{Y}_{DSD}(3)}$	{[35.45; 42.27) , 0.3; [42.27; 43.38) , 0.1; [43.38; 44.50) , 0.1; [44.50; 45.61) , 0.1; [45.61; 46.72) , 0.1; [46.72; 50.46] , 0.3}
$H_{\widehat{Y}_{CM}(3)}$	{[36.98; 42.74) , 0.3; [42.74; 43.62) , 0.1; [43.62; 44.51) , 0.1; [44.51; 45.39) , 0.1; [45.39; 46.28) , 0.1; [46.28; 48.93] , 0.3}
$H_{\widehat{Y}_{BD}(3)}$	{[37.51; 42.69) , 0.3; [42.69; 43.49) , 0.1; [43.49; 44.28) , 0.1; [44.28; 45.08) , 0.1; [45.08; 45.88) , 0.1; [45.88; 48.27] , 0.3}
$H_{\widehat{Y}_{VI}(3)}$	{[35.29; 42.42) , 0.3; [42.42; 43.51) , 0.1; [43.51; 44.61) , 0.1; [44.61; 45.71) , 0.1; [45.71; 46.80) , 0.1; [46.80; 50.1] , 0.3}
$H_{Y(4)}$	{[36.38; 39.75) , 0.3; [39.75; 40.87) , 0.1; [40.87; 41.96) , 0.1; [41.96; 43.05) , 0.1; [43.05; 44.14) , 0.1; [44.14; 47.41] , 0.3}
$H_{\widehat{Y}_{DSD}(4)}$	{[35.80; 40.08) , 0.3; [40.08; 41.50) , 0.1; [41.50; 42.92) , 0.1; [42.92; 43.81) , 0.1; [43.81; 44.70) , 0.1; [44.70; 47.37] , 0.3}
$H_{\widehat{Y}_{CM}(4)}$	{[36.98; 40.55) , 0.3; [40.55; 41.74) , 0.1; [41.74; 42.93) , 0.1; [42.93; 43.58) , 0.1; [43.58; 44.23) , 0.1; [44.23; 46.18] , 0.3}
$H_{\widehat{Y}_{BD}(4)}$	{[37.51; 40.72) , 0.3; [40.72; 41.97) , 0.1; [41.97; 42.86) , 0.1; [42.86; 43.45) , 0.1; [43.45; 44.03) , 0.1; [44.03; 45.79] , 0.3}
$H_{\widehat{Y}_{VI}(4)}$	{[35.71; 40.13) , 0.3; [40.13; 41.61) , 0.1; [41.61; 43.08) , 0.1; [43.08; 43.89) , 0.1; [43.89; 44.69) , 0.1; [44.69; 47.12] , 0.3}
$H_{Y(5)}$	{[39.19; 42.69) , 0.3; [42.69; 43.86) , 0.1; [43.86; 45.03) , 0.1; [45.03; 46.19) , 0.1; [46.19; 47.36) , 0.1; [47.36; 50.86] , 0.3}
$H_{\widehat{Y}_{DSD}(5)}$	{[39.68; 42.52) , 0.3; [42.52; 43.64) , 0.1; [43.64; 44.75) , 0.1; [44.75; 45.86) , 0.1; [45.86; 46.97) , 0.1; [46.97; 50.25] , 0.3}
$H_{\widehat{Y}_{CM}(5)}$	{[40.78; 42.99) , 0.3; [42.99; 43.87) , 0.1; [43.87; 44.76) , 0.1; [44.76; 45.64) , 0.1; [45.64; 46.53) , 0.1; [46.53; 49.19] , 0.3}
$H_{\widehat{Y}_{BD}(5)}$	{[40.92; 42.92) , 0.3; [42.92; 43.71) , 0.1; [43.71; 44.51) , 0.1; [44.51; 45.31) , 0.1; [45.31; 46.10) , 0.1; [46.10; 48.49] , 0.3}
$H_{\widehat{Y}_{VI}(5)}$	{[39.80; 42.54) , 0.3; [42.54; 43.64) , 0.1; [43.64; 44.74) , 0.1; [44.74; 45.83) , 0.1; [45.83; 46.93) , 0.1; [46.93; 50.22] , 0.3}

Table D2. Observed and predicted histograms (using different methods) of the hematocrit values shown in *Table 4* (part2: patients 6 to 10).

Patient	Distributions of the hematocrit values
$H_{Y(6)}$	{[39.70; 43.17) , 0.3; [43.17; 44.32) , 0.1; [44.32; 44.81) , 0.1; [44.81; 45.29) , 0.1; [45.29; 45.78) , 0.1; [45.78; 47.24] , 0.3}
$H_{\widehat{Y}_{DSD}(6)}$	{[40.93; 42.92) , 0.3; [42.92; 43.58) , 0.1; [43.58; 44.04) , 0.1; [44.04; 44.51) , 0.1; [44.51; 44.99) , 0.1; [44.99; 46.45] , 0.3}
$H_{\widehat{Y}_{CM}(6)}$	{[41.50; 43.14) , 0.3; [43.14; 43.68) , 0.1; [43.68; 44.05) , 0.1; [44.05; 44.42) , 0.1; [44.42; 44.79) , 0.1; [44.79; 45.90] , 0.3}
$H_{\widehat{Y}_{BD}(6)}$	{[41.58; 43.05) , 0.3; [43.05; 43.54) , 0.1; [43.54; 43.87) , 0.1; [43.87; 44.21) , 0.1; [44.21; 44.54) , 0.1; [44.54; 45.53] , 0.3}
$H_{\widehat{Y}_{VI}(6)}$	{[40.92; 42.95) , 0.3; [42.95; 43.62) , 0.1; [43.62; 44.08) , 0.1; [44.08; 44.54) , 0.1; [44.54; 44.99) , 0.1; [44.99; 46.47] , 0.3}
$H_{Y(7)}$	{[41.56; 44.11) , 0.3; [44.11; 44.95) , 0.1; [44.95; 45.80) , 0.1; [45.80; 46.65) , 0.1; [46.65; 47.19) , 0.1; [47.19; 48.81] , 0.3}
$H_{\widehat{Y}_{DSD}(7)}$	{[42.67; 43.86) , 0.3; [43.86; 44.26) , 0.1; [44.26; 44.65) , 0.1; [44.65; 45.22) , 0.1; [45.22; 45.78) , 0.1; [45.78; 47.48] , 0.3}
$H_{\widehat{Y}_{CM}(7)}$	{[43.18; 44.07) , 0.3; [44.07; 44.37) , 0.1; [44.37; 44.66) , 0.1; [44.66; 45.13) , 0.1; [45.13; 45.60) , 0.1; [45.60; 47.00] , 0.3}
$H_{\widehat{Y}_{BD}(7)}$	{[43.09; 43.89) , 0.3; [43.89; 44.16) , 0.1; [44.16; 44.42) , 0.1; [44.42; 44.85) , 0.1; [44.85; 45.27) , 0.1; [45.27; 46.53] , 0.3}
$H_{\widehat{Y}_{VI}(7)}$	{[42.76; 43.87) , 0.3; [43.87; 44.24) , 0.1; [44.24; 44.61) , 0.1; [44.61; 45.19) , 0.1; [45.19; 45.77) , 0.1; [45.77; 47.51] , 0.3}
$H_{Y(8)}$	{[38.4; 40.34) , 0.3; [40.34; 40.99) , 0.1; [40.99; 41.64) , 0.1; [41.64; 42.28) , 0.1; [42.28; 42.93) , 0.1; [42.93; 45.22] , 0.3}
$H_{\widehat{Y}_{DSD}(8)}$	{[39.26; 40.74) , 0.3; [40.74; 41.24) , 0.1; [41.24; 41.72) , 0.1; [41.72; 42.20) , 0.1; [42.20; 42.79) , 0.1; [42.79; 44.54] , 0.3}
$H_{\widehat{Y}_{DSDII}(8)}$	{[38.78; 40.36) , 0.3; [40.36; 40.98) , 0.1; [40.98; 41.63) , 0.1; [41.63; 42.34) , 0.1; [42.34; 43.08) , 0.1; [43.08; 45.33] , 0.3}
$H_{\widehat{Y}_{CM}(8)}$	{[39.80; 40.95) , 0.3; [40.95; 41.33) , 0.1; [41.33; 41.72) , 0.1; [41.72; 42.10) , 0.1; [42.10; 42.58) , 0.1; [42.58; 44.00] , 0.3}
$H_{\widehat{Y}_{BD}(8)}$	{[40.04; 41.08) , 0.3; [41.08; 41.43) , 0.1; [41.43; 41.77) , 0.1; [41.77; 42.12) , 0.1; [42.12; 42.55) , 0.1; [42.55; 43.83] , 0.3}
$H_{\widehat{Y}_{VI}(8)}$	{[39.31; 40.74) , 0.3; [40.74; 41.22) , 0.1; [41.22; 41.70) , 0.1; [41.70; 42.17) , 0.1; [42.17; 42.76) , 0.1; [42.76; 44.52] , 0.3}
$H_{Y(9)}$	{[28.83; 32.86) , 0.3; [32.86; 34.21) , 0.1; [34.21; 35.55) , 0.1; [35.55; 36.84) , 0.1; [36.84; 38.12) , 0.1; [38.12; 41.98] , 0.3}
$H_{\widehat{Y}_{DSD}(9)}$	{[27.66; 33.54) , 0.3; [33.54; 35.50) , 0.1; [35.50; 36.70) , 0.1; [36.70; 37.91) , 0.1; [37.91; 39.20) , 0.1; [39.20; 43.08] , 0.3}
$H_{\widehat{Y}_{CM}(9)}$	{[29.20; 34.09) , 0.3; [34.09; 35.72) , 0.1; [35.72; 36.68) , 0.1; [36.68; 37.63) , 0.1; [37.63; 38.59) , 0.1; [38.59; 41.47] , 0.3}
$H_{\widehat{Y}_{BD}(9)}$	{[30.51; 34.91) , 0.3; [34.91; 36.37) , 0.1; [36.37; 37.23) , 0.1; [37.23; 38.10) , 0.1; [38.10; 38.96) , 0.1; [38.96; 41.55] , 0.3}
$H_{\widehat{Y}_{VI}(9)}$	{[27.54; 33.59) , 0.3; [33.59; 35.61) , 0.1; [35.61; 36.80) , 0.1; [36.80; 37.90) , 0.1; [37.90; 39.18) , 0.1; [38.18; 42.74] , 0.3}
$H_{Y(10)}$	{[44.48; 46.90) , 0.3; [46.90; 47.70) , 0.1; [47.70; 48.51) , 0.1; [48.51; 49.31) , 0.1; [49.31; 50.12) , 0.1; [50.12; 52.53] , 0.3}
$H_{\widehat{Y}_{DSD}(10)}$	{[45.85; 47.48) , 0.3; [47.48; 48.03) , 0.1; [48.03; 48.58) , 0.1; [48.58; 49.13) , 0.1; [49.13; 49.68) , 0.1; [49.68; 51.33] , 0.3}
$H_{\widehat{Y}_{CM}(10)}$	{[46.43; 47.73) , 0.3; [47.73; 48.17) , 0.1; [48.17; 48.61) , 0.1; [48.61; 49.05) , 0.1; [49.05; 49.48) , 0.1; [49.48; 50.80] , 0.3}
$H_{\widehat{Y}_{BD}(10)}$	{[46.02; 47.18) , 0.3; [47.18; 47.58) , 0.1; [47.58; 47.97) , 0.1; [47.97; 48.37) , 0.1; [48.37; 48.76) , 0.1; [48.76; 49.94] , 0.3}
$H_{\widehat{Y}_{VI}(10)}$	{[45.91; 47.51) , 0.3; [47.51; 48.06) , 0.1; [48.06; 48.60) , 0.1; [48.60; 49.14) , 0.1; [49.14; 49.68) , 0.1; [49.68; 51.31] , 0.3}

Example of Section 4.2.2

In Figs. D1 and D2, we compare the observed and predicted distributions of the logarithm of the number of violent crimes using the *DSD* linear regression model and the models proposed by Billard and Diday [3] and Irpino and Verde [15,19].

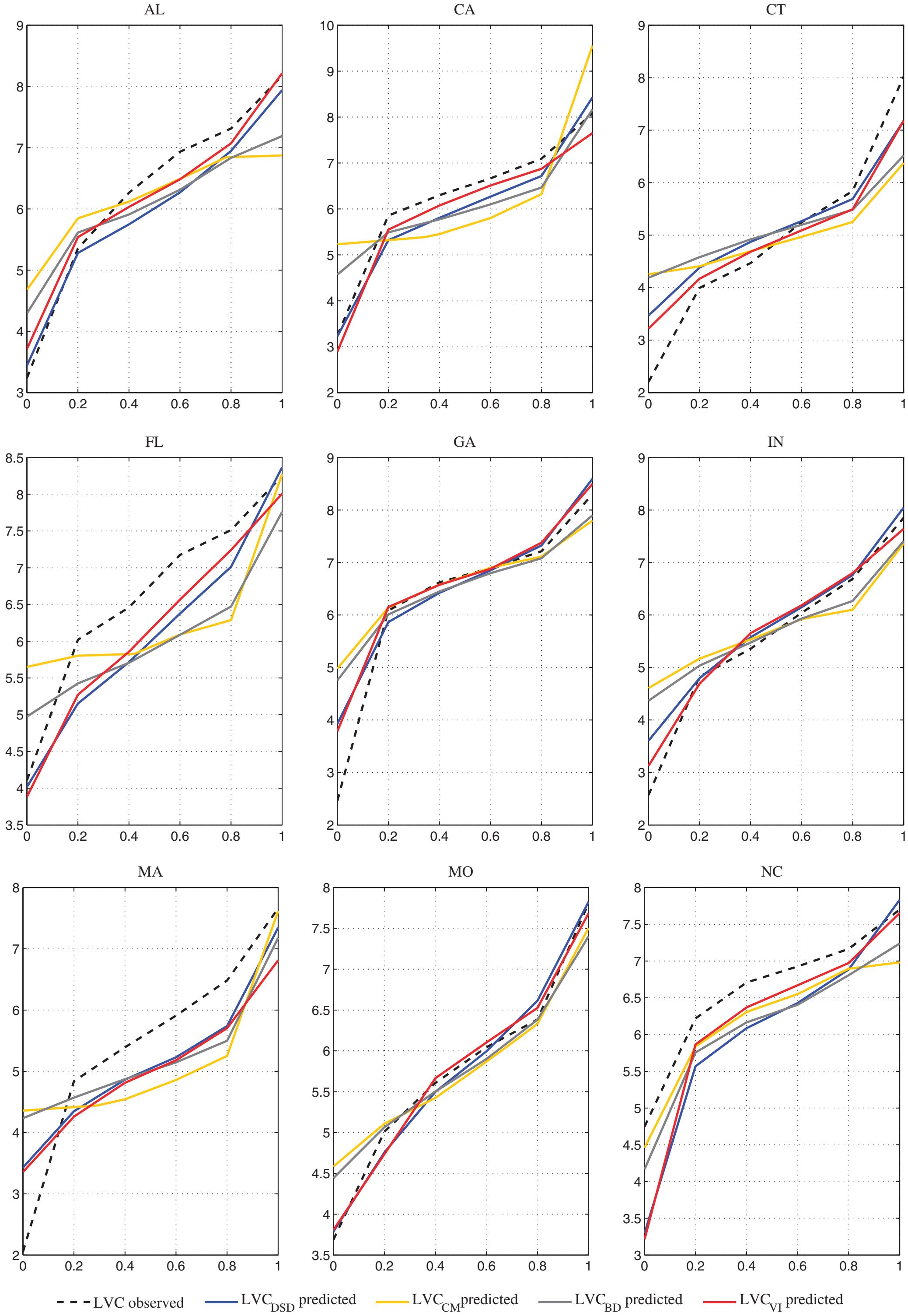

Fig. D1 Observed and predicted quantile functions of LVC considering the models: *DSD, CM, BD, VI* (part1).

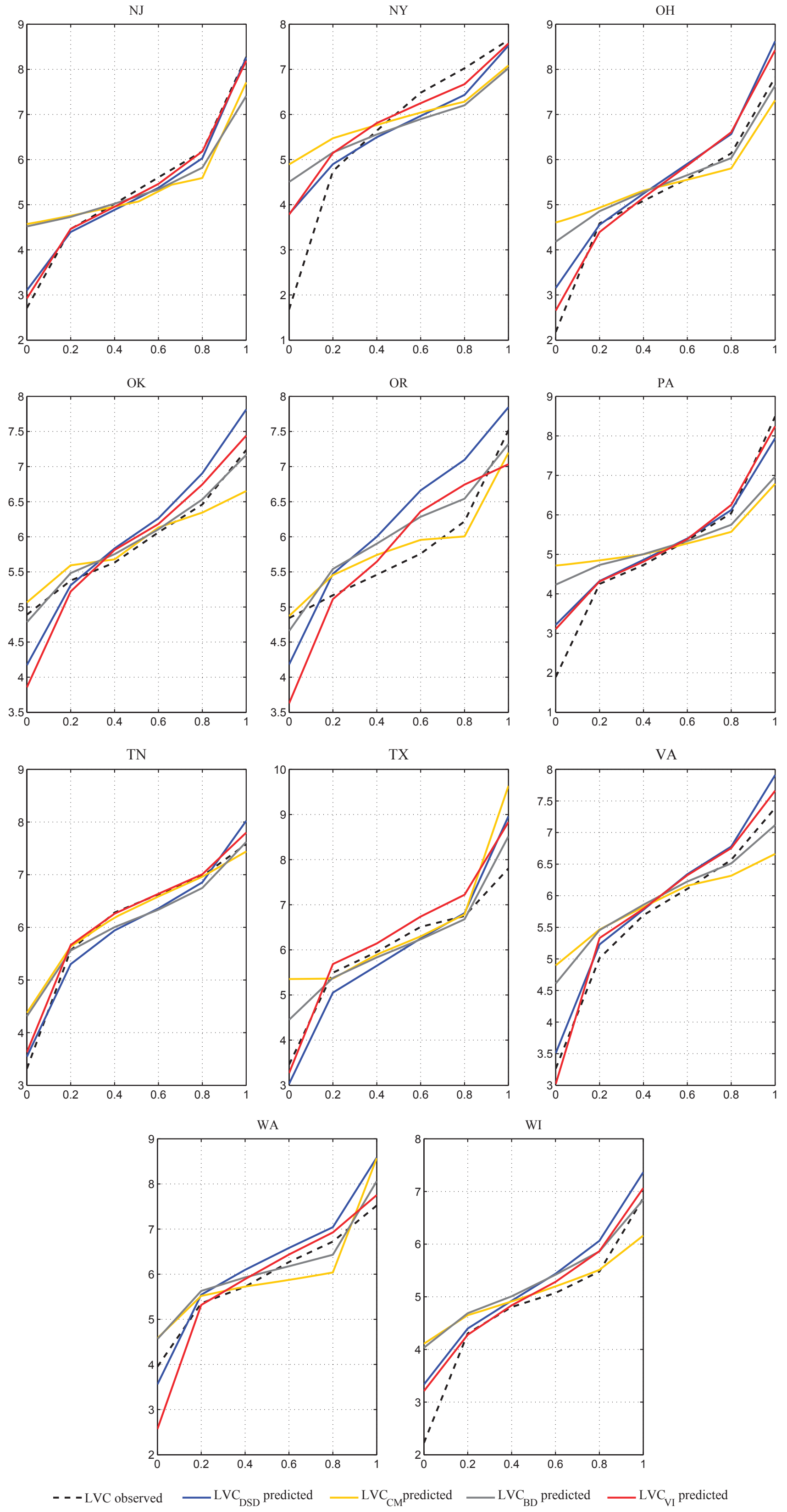

Fig. D2 Observed and predicted quantile functions of LVC considering the models: *DSD, CM, BD, VI* (part 2).

REFERENCES

[1] E. Diday, The symbolic approach in clustering and related methods of data analysis: the basic choices, In Classification and Related Methods of Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS'87), H.-H. Bock, ed. Amsterdam, Holland, 1988, 673–684.

[2] J. Arroyo, Métodos de Predicción para series temporales de Intervalos e Histogramas, Tesis para la obtención del título de Doctor; Universidad Pontificia Comillas, Madrid, 2008.

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

[4] L. Billard, and E. Diday, From the statistics of data to the statistics of knowledge: symbolic data analysis, J Am Stat Assoc 98(462) (2003), 470–487.

[5] H-H Bock, E. Diday eds., Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Berlin-Heidelberg, Springer-Verlag, 2000.

[6] E. Diday and M. Noirhomme-Fraiture, eds., Symbolic Data Analysis and the SODAS Software, Chichester, John Wiley & Sons, Ltd., 2008.

[7] M. Noirhomme-Fraiture, and P. Brito, Far beyond the classical data models: symbolic data analysis, Stat Anal Data Min 4(2) (2011), 157–170.

[8] L. Billard, and E. Diday Symbolic Regression Analysis, In Classification, Clustering and Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS'02), K. Jajuga, A. Sokolowski, and H.-H. Bock, eds. Heidelberg, Springer, 2002, 281–288.

[9] A. Irpino, and R. Verde, A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data, In Classification and Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS'06), V. Batagelj, H.-H. Bock, and A. Ferligoj, eds. Heidelberg, Springer, 2006, 185–192.

[10] O. Rodriguez, E. Diday, and S. Winsberg, Generalization of the principal components analysis to histogram data, Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases (PKDD), Workshop on Symbolic Data Analysis; Lyon, France, 2000.

[11] O. Rodriguez, and A. Pacheco, Applications of histogram principal components analysis, The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Data Bases (PKDD), Pisa, Italy, 2004.

[12] P. Brito, and G. Polaillon, Classification conceptuelle avec généralisation par intervalles [Conceptual clustering with generalization by intervals], Revue des Nouvelles Technologies de l'Information E.23 (2012), 35–40.

[13] P. Brito, and M. Chavent, Divisive monothetic clustering for interval and histogram-valued data, Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012), Vilamoura, Portugal, 2012.

[14] J. Arroyo, and C. Maté, Forecasting histogram time series with K-nearest neighbours methods, Int J Forecasting 25 (2009), 192–207.

[15] R. Verde, and A. Irpino, Ordinary least squares for histogram data based on Wasserstein distance, Proceedings of COMPSTAT'2010, Y. Lechevallier, and G. Saporta, eds. Heidelberg, Physica Verlag, 2010, 581–589.

[16] L. Billard, and E. Diday Regression analysis for interval-valued data, In Data Analysis, Classification and Related Methods. Proceedings of the Conference of the International Federation of Classification Societies (IFCS'00), H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen, and M. Schader, eds. Heidelberg, Springer, 2000, 369–374.

[17] E. A. L. Neto, and F. A. T. De Carvalho, Constrained linear regression models for symbolic interval-valued variables, Comput Stat Data Anal 54 (2010), 333–347.

[18] E. A. L. Neto, and F. A. T. De Carvalho, Centre and range method for fitting a linear regression model to symbolic intervalar data, Comput Stat Data Anal 52 (2008), 1500–1515.

[19] A. Irpino, and R. Verde. Linear regression for numeric symbolic variables: an ordinary least squares approach based on Wasserstein Distance. Adv Data Anal Classif 9(1) (2015), 81–106.

[20] P. Brito, and A. P. Duarte Silva, Modelling interval data with Normal and Skew-Normal distributions, J Appl Stat 39(1) (2011), 3–20.

[21] E. A. L. Neto, G. M. Cordeiro, and F. A. T. De Carvalho, Bivariate symbolic regression models for interval-valued variables, J Stat Comput Simulation 81 (2011), 1727–1744.

[22] R. Williamson, Probabilistic Arithmetic. Thesis for the obtencion the degree of Doctor, Department of Electrical Engineering, University of Queensland, Australia, 1989.

[23] R. Verde, and A. Irpino, Comparing histogram data using a Mahalanobis-Wasserstein distance, In Proceedings of COMPSTAT'2008 P. Brito, ed. Heidelberg, Physica Verlag, 2008, 77–89.

[24] R. Verde, and A. Irpino, Dynamic clustering of histogram data: using the right metric, In Selected Contributions in Data Analysis and Classification, P. Brito, P. Bertrand, G. Cucumel, and F. De Carvalho, eds. Heidelberg, Springer, 2007, 123–134.

[25] A. Colombo, and R. Jaarsma, A powerful numerical method to combine random variables, IEEE Trans Rel 29(2) (1980), 126–129.

[26] J. Case, Interval arithmetic and analysis the college mathematics journal 30(2) (1999), 106–111.

[27] A. Irpino, and E. Romano, Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. In Actes des cinquièmes journées Extraction et Gestion des Connaissances. M. Noirhomme-Fraiture and G. Venturini, eds. Cèpadués- Éditions (2007), 99–110.

[28] C. L. Mallows, A note on asymptotic joint normality, Ann Math Stat 43(2) (1972), 508–515.

[29] W. Winston, Operations Research. Applications and Algorithms, (3rd ed.), California, Duxbury Press, 1994.

[30] M. Redmond, UCI machine learning repository, communities and crime data set, Irvine, CA, University of California, School of Information and Computer Science, 2011, http://www.ics.uci.edu/mlearn/MLRepository.html.