Parametric and Nonparametric Mixture Models Based on Interval Regression

Roberta A. de A. Fagundes[1], Bruno A. Pimentel[2], Renata M.C.R. de Souza[2] and Francisco José A. Cysneiros[3]
[1] **Universidade de Pernambuco, Campus Gov. Miguel Arraes de Alencar, Polo Comercial, BR 104, Km 62 Caruaru (PE)- Brazil**
[2] **Universidade Federal de Pernambuco, Centro de Informática, Av. Jornalista Anibal Fernandes s/n, Cidade Universitária, CEP 50.740-560, Recife (PE)-Brazil**
[3] **Universidade Federal de Pernambuco - Depto. de Estatística, Av. Prof. Luiz Freire, s/n, Cidade Universitária,CEP 50740-540-Recife (PE) - Brazil**

Abstract— *It is increasingly common to use tools of Symbolic Data Analysis to reduce data sets and create new data ones, called symbolic data sets, without losing much information. These new data sets can be obtained for preserving the privacy of individuals when their information are present in the original data sets. In this work, we propose prediction models based on regression mixtures for interval symbolic data. The advantage of these mixtures is that they allow to assume a nonparametric function for center (midpoint) and a parametric function for range of the intervals or a nonparametric function for range and a parametric function for center. The proposed models are applied to a scientific production interval data set of institutions from Brazil. Here, this interval data set was built in order to reduce data and preserve the the information privacy. The quality of the interval prediction obtained by the models is assessed by a mean magnitude of relative error.*

Keywords: symbolic data analysis; regression mixture; interval data; interval regression

1. Introduction

The statistical treatment of interval data has been considered in the context of *Symbolic Data Analysis* (SDA) [1] which is a knowledge discovery and data management field related to multivariate analysis, pattern recognition and artificial intelligence. An extensive coverage of earlier symbolic data analysis methods can be found in [2]. *SDA* focuses on the analysis of data sets where individuals are described by variables that can represent internal variation and/or structure. Symbolic data values can be intervals, histograms, distributions, lists of values, taxonomies, etc. The term symbolic is used to stress the fact that the values are of a different nature.

Some data sets naturally consist of symbolic interval data as for example, the data set of minimum and maximum temperatures naturally represented by intervals, while many other interval symbolic data sets result from the aggregation of large classical data sets. For example, regarding scientific production data, the interest is in describing the behavior of some group of researchers rather than each scientific production by itself. By aggregating the scientific production data through institution and area of knowledge categorical variables it is obtained the information of interest; here the observed variability within each group is of utmost importance.

Regression analysis is one of the most widely used techniques in engineering, management and many other fields. In the framework of regression models for symbolic interval data, several models have been introduced. Most of these models consider parametric functions. The purpose of this work is to investigate the use of regression mixture models for interval-valued data. Four interval models are adopted and each one uses parametric and nonparametric functions. For each model a function fits midpoint data and another function fits range data of the intervals. Here, linear and robust regressions are considered as options for parametric functions and kernel regression as option for nonparametric regression.

In previous work [3], we proposed two mixture models for intervals data based on kernel and linear regressions. However, it is well known in the literature that linear regression fits the parameters using least squares approach that is sensitive to outliers. Thus, this paper generalizes the nonparametric and parametric mixture model for interval data based on the use of kernel and robust regressions for midpoint and range of the intervals. The linear regression is a particular case of the robust regression when the weights for items of the data set are identical to 1.0.

In Fagundes *et.al* [4] interval outliers are defined in the context of linear regression based on midpoint and range. The advantages of mixture models based on robust and kernel regressions are: kernel regression provides a versatile method of exploring a general relationship between variables and gives good predictions of observations yet to be made without reference to a fixed parametric model; and robust

regression model is not sensitive in the presence of outliers.

The proposed models are applied to a scientific production interval data set of institutions from Brazil. An educational data analysis is a domain of application that has not yet been explored in the *SDA* framework. Concerning social services domain, Neto and De Carvalho (2002) [5] showed an application about administrative management of Brazilian cities (in Pernambuco state) using interval-valued variables. Da Silva, Lechevallier, de Carvalho, and Trousse (2006) [6] made experiments using information of web users whose aim is to cluster users with the same web usage behavior together, for this, a dynamic clustering method for interval-valued variable was used. Zuccolotto (2006) [7] presented the use of Symbolic Data Analysis in a database about job satisfaction of Italian workers through the principal component analysis method.

Here, the scientific production interval data set was built taking into account the following advantages:

1) Summarize data: Initially, the data set contains more than 140000 individuals described by points in the R^{33}. These data can be aggregated using one or more categorical variables and a new data set smaller than the old one without losing much information can be obtained;
2) Ensure the privacy of individuals: The original data contains information that explicitly identify the individuals. The generalization process allows to ensure confidentiality of original data;
3) Use higher-level category: The original data set represents scientific production of researchers whereas the aggregated data set is able to represent profiles of scientific production taking into account variability intrinsic to each profile. The aggregated of this paper is to study Brazilian production in the particular scientific area of Computer Science.

The rest of this paper is organized as follows, section 2 describes regression mixture models for interval data proposed in this paper. Section 3 describes the scientific production data considered in this paper and highlights the aggregation process adopted to obtain symbolic data. Section 4 presents a performance analysis of these models using on the scientific production data. Finally, Section 5 gives the concluding remarks.

2. Regression mixture models for interval data

Let $\Omega = 1, \ldots, n$ be a data set of n objects described by the response interval-valued variable Y and p predictor interval-valued variables $\mathbf{X} = (X_1, \ldots, X_p)$. Each object i of Ω is represented as an interval feature vector $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ where $x_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$ $(j = 1, \ldots, p)$ and $y_i = [\alpha_i, \lambda_i] \in \Im$.

This method aims to find a smooth and nonlinear relationship between the interval response variable Y and the vector de interval predictor variables $\mathbf{X} = (X_1, \ldots, X_p)^T$ using the information of center (midpoint) and range of the intervals as:

2.1 Representing intervals

The interval response $[\alpha_i, \lambda_i]$ can be rewritten by:

$$y_i = [\alpha_i, \lambda_i] = [y^c - y^r/2, y^c + y^r/2]$$

Assuming the result above, the interval response variable is represented by a pair of quantitative variables (Y_c, Y_r) that describes the center and range of the intervals, respectively. Consider also that each predictor interval variable X_j is represented by a pair of quantitative variables (X_j^c, X_j^r) that describes the center and range of this interval variable.

Let $\mathbf{x}_i^c = (x_{i1}^c, \ldots, x_{ip}^c)^T$ where $x_{ij}^c = [a_{ij} + b_{ij}]/2$ and $\mathbf{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)^T$ where $x_{ij}^r = b_{ij} - a_{ij}$. Consider $y_i^c = [\alpha_i + \lambda_i]/2$ and $y_i^r = \lambda_i - \alpha_i$. Thus, X is represented by (X^c, X^r).In this regression method, to explore Y by X is equivalent to explore Y^c by $\mathbf{X}^c$ and Y^r by $\mathbf{X}^r$, separately.

2.2 Modeling the relationship between intervals

A relationship between Y and X is give as:

$$E(Y/\mathbf{X}) = [E(Y^c/\mathbf{X}^c) - E(Y^r/\mathbf{X}^r), E(Y^c/\mathbf{X}^c) + E(Y^r/\mathbf{X}^r)],$$

$$E(Y/\mathbf{X}) = \left[m^c(\mathbf{X}^c) - \frac{1}{2}m^r(\mathbf{X}^r), m^c(\mathbf{X}^c) + \frac{1}{2}m^r(\mathbf{X}^r)\right].$$

where m^c and m^r are parametric and nonparametric functions. Examples of m^c and m^r are described in Table 1.

Table 1: Mixture Models

Models	$m^c(\mathbf{X}^c)$	$m^r(\mathbf{X}^r)$
1	$\sum_{i=1}^n \omega_i^c y_i^c$ with $\omega_i^c = \frac{K(d(\mathbf{x}^c, \mathbf{x}_i^c))}{\sum_{i=1}^n K(d(\mathbf{x}^c, \mathbf{x}_i^c))}$	$(\mathbf{x}^r)^T \widehat{\beta}^r$ with $\widehat{\beta}^r = ({\mathbf{X}^r}^T \mathbf{X}^r)^{-1} {\mathbf{X}^r}^T \mathbf{y}^r$
2	$(\mathbf{x}^c)^T \widehat{\beta}^c$ with $\widehat{\beta}^c = ({\mathbf{X}^c}^T \mathbf{X}^c)^{-1} {\mathbf{X}^c}^T \mathbf{y}^c$	$\sum_{i=1}^n \omega_i^r y_i^r$ with $\omega_i^r = \frac{K(d(\mathbf{x}^r, \mathbf{x}_i^r))}{\sum_{i=1}^n K(d(\mathbf{x}^r, \mathbf{x}_i^r))}$
3	$\sum_{i=1}^n \omega_i^c y_i^c$ with $\omega_i^c = \frac{K(d(\mathbf{x}^c, \mathbf{x}_i^c))}{\sum_{i=1}^n K(d(\mathbf{x}^r, \mathbf{x}_i^r))}$	$(\mathbf{x}^r)^T \widehat{\beta}^r$ with $\widehat{\beta}^r = ({\mathbf{X}^r}^T \mathbf{W}^r \mathbf{X}^r)^{-1} {\mathbf{X}^r}^T \mathbf{W}^r \mathbf{y}^r$
4	$(\mathbf{x}^c)^T \widehat{\beta}^c$ with $\widehat{\beta}^c = ({\mathbf{X}^c}^T \mathbf{W}^c \mathbf{X}^c)^{-1} {\mathbf{X}^c}^T \mathbf{W}^c \mathbf{y}^c$	$\sum_{i=1}^n \omega_i^r y_i^r$ with $\omega_i^r = \frac{K(d(\mathbf{x}^r, \mathbf{x}_i^r))}{\sum_{i=1}^n K(d(\mathbf{x}^r, \mathbf{x}_i^r))}$

where $\mathbf{W}^c$ and $\mathbf{W}^r$ are weight matrices for center and range data, respectively;$\mathbf{X}^c$ and $\mathbf{X}^r$ input data matrices for center and range data, respectively; $\mathbf{y}^c$ and $\mathbf{y}^r$ are response data vectors for center and range data, respectively.

In this paper four parametric and nonparametric regression mixture models are investigated:

1) The model 1 (here called $MM : CK + RL$) combines kernel regression for center data and linear (multiple) regression for range data.
2) The model 2 (here called $MM : CL + RK$) combines linear (multiple) regression for center data and kernel regression for range data.
3) The model 3 (her called $MM : CK + RR$) combines kernel regression for center data and robust regression for range data.
4) The model 4 (here called $MM : CR + RK$) combines robust regression for center data and kernel regression for range data.

The regression mixture models 1 and 2 assume that the data set does not include interval outliers. The parameter are estimated from data using the least squares approach. The regression mixture models 3 and 4 consider interval outliers. Here, these outliers can be identified by investigating if there are point outliers on midpoint or range data of the intervals. The Fisher scoring method [8] can be easily applied to get $\hat{\boldsymbol{\beta}}^c$ and $\hat{\boldsymbol{\beta}}^r$ where the process for $\hat{\boldsymbol{\beta}}^c$ and $\hat{\boldsymbol{\beta}}^r$ can be interpreted as a modified least square.

There is a number of popular robust criterion function ρ. The least square is a particular case when the weight given to each residual is equal to 1.0. So, the robust regression method can be classified by the their ψ function that controls the weight given to each residual (Montegomery *et. al* [8]).

These regression mixture models use Guassian kernel functions and squared Euclidean distance applied to center and range of the intervals data. In these kernels functions, the bandwidth h is the standard deviation for a normal distributions centered on $\mathbf{x}_i^c$ or $\mathbf{x}_i^r$.

3. Scientific Production Data

The data were extracted from the National Council for Scientific and Technological Development (`http://www.cnpq.br`) that is an agency of the Ministry of Science, Technology and Innovation in order to promote scientific and technological research and the training of human resources for research in the country. Other important Brazilian agency is the Coordination for the Improvement of Higher Level Personnel (`http://www.capes.gov.br`) whose main activity is to evaluate the Brazilian research institutes. This agency evaluates the Brazilian post-graduate courses based on the scientific production of the researchers.

The scientific production of each researcher is described by a set of 33 continuous numerical and 3 categorical variables. The continuous variables are averages of production values computed in three years (2006, 2007 and 2008) for each researcher. They are: 1. National journal, 2. International journal, 3. Presentation of papers, 4. Books , 5. Chapter of book, 6. Other publications, 7. Summary of journal, 8. Summary of annals, 9. Publication, 10. PhD guidelines finished, 11. Master guidelines finished, 12. Specialization guidelines finished, 13. Graduate guidelines finished, 14. UR guidelines finished, 15. PhD guidelines unfinished, 16. Master guidelines unfinished, 17. Specialization guidelines unfinished, 18. Graduate guidelines unfinished, 19. UR guidelines unfinished, 20. Guidelines finished, 21. Guidelines unfinished, 22. Other intellectual productions, 23. Other types of production, 24. Registered software, 25. Unregistered software, 26. Unregistered product, 27. Registered techniques, 28. Unregistered techniques, 29. Technique works, 30. Technique presentations, 31. Other production-related techniques, 32. Techniques and 33. Artworks. The categorical variables are: institute, area of knowledge and sub-area of knowledge.

The data base considers 141260 researchers from 410 institutions such as federal, state, municipal and private universities, colleges integrated, colleges, institutes, schools, technical education centers that have at least one course of masters or doctorate degree recognized by Coordination for the Improvement of Higher Level Personnel, public institutes of scientific research, public technological institutes and federal centers of technological education or research laboratories and development of state enterprisers. Each institution is organized into several areas of knowledge such as Biological Sciences, Exact Science, Engineering, Agricultural Science, Health Sciences, Applied Social Sciences, Humanities and Linguistics-Literature-Arts. Each area of knowledge is divided in 76 sub-areas of knowledge. Each researcher is related to only one sub-area of knowledge.

Let Ω be a data set of researches indexed by $i (i = 1, ..., 141260)$. Each researcher is described by a vector of 33 continuous numerical and 3 categorial values $v_i = (v_i^1, \ldots, v_i^3 3, c_i^1, c_i^2, c_i^3)$ where $v_i^j \in \Re$ $(j = 1, \ldots, 33)$ and c_i^1, c_i^2, c_i^3 are the institute, area and sub-area of knowledge of the researcher i. Tools of SDA are applied on this researcher data base in order to build new units. These new units are modeled by interval symbolic data. Here, three reasons are considered by using SDA: 1) to reduce the size of the original base since SDA starts as mining data process applied on a large data base; 2) to ensure the privacy of researches and 3) to use higher-level objects described by the variables that allow to take into account variability and/or uncertainty.

This work analyzes the interval scientific production data of Brazilian institutes. In this context, the interval data proposed for [9] and [10] available at `http://www.cin.ufpe.br/~bap/ScientificProduction` are aggregated the sub-area of knowledge categorical variables Computer Science generating a new data base of size 166. These data represent new concepts of scientific production are describe in Table 2. In order to apply regression models to this interval data set, predict variables are choice using an *a priori* knowledge of software estimation experts and they are: $NPhd$(PhD guidelines finished), $NMaster$ (Master guidelines finished) and $NScientific$ (UR (scientific) guidelines finished). These predictor variables explain the number of publications of the researchers from science

computer response variable ($NPub$ - Publications) as it is showed interval graph in Figure 1, 2 and 3. In these Figures illustrates the interval-valued data set containing very large rectangles in a data set can mean the presence of atypical intervals. Note in these figures that, there are rectangle *outliers* that are remote in the X and Y coordinate.

Table 2 presents part of the interval scientific production data set from proposed in this work. Each row of this table corresponds to a number of publications of the researchers from science computer described by the variables $NPhd$, $NMaster$, $NScientific$ and $NPub$ with an interval description the minimum and maximum values of these variables . The rows of this table describe concepts of scientific production. These concepts model the information taking into account variability. There is remote intervals on the response variable can be easily observed in Figure 1, 2 and 3 in both coordinates(X and Y) and confirmed by the values described in Table 2.

Table 2: Concepts of scientific production data described by interval data.

	$NPub$	PhD guidelines ($NPhd$)	Master guidelines ($NMaster$)	UR guidelines ($NScientific$)
1	[0,14.25]	[0,6.5]	[0,1]	[0,2]
2	[0.5,16.25]	[0,0]	[0,0]	[0,5.75]
3	[0,28.5]	[0,817]	[0,3.75]	[0,10]
...	...	...	...	...
162	[0,39.75]	[0,0.25]	[0,4.75]	[0,26.6]
163	[0,9.25]	[0,0.25]	[0,0]	[0,2.75]
164	[0,8.75]	[0,1.75]	[0,1.25]	[0,0.75]

4. Performance analysis

The four regression mixture models ($MM : CK + RL$, $MM : CL+RK$, $MM : CK+RR$ and $MM : CR+RK$) proposed in this paper are applied to scientific production interval data set and a performance analysis is carried out. Moreover, in order to make a comparative study with other distribution-free regression methods of the *SDA* literature, the interval robust regression model based on center and range information [4] (here called IRR), the interval linear regression model based on center and range information [11] (here called ILR) and the interval kernel regression model based on center and range information[3] (here called IKR) are also to applied to this data set.

The prediction accuracy of the models are measured by the mean magnitude of relative error ($MMRE$) that is estimated by the hold-out method in the framework of a Monte Carlo simulation with 1000 replications. The test and learning sets are randomly selected from each input data set. The learning set corresponds to 75% of the data and the test data set corresponds to 25%. The experiments are performed using the Language R {http://www.r-project.org/}. The $MMRE$ is given as

$$MMRE = \sum_{i=1}^{n} \frac{1}{2n} \left\{ \left| \frac{\alpha_i - \hat{\alpha}_i}{\alpha_i} \right| + \left| \frac{\lambda_i - \hat{\lambda}_i}{\lambda_i} \right| \right\}. \quad (1)$$

The problem of an automatic choice of the bandwidth (h) is important in kernel regression. An appropriate bandwidth can be defined by studying the $MMRE$ behavior regarding different bandwidths [12]. Here, the bandwidth is chosen based on the lowest value of the $MMRE$ that in this simulation study is 0.01.

The comparison between the regression methods of the SDA literature and regression mixture methods are achieved by applying the statistical *Wilcoxon* test for not paired samples at a significance level of 5%. Let μ_1 and μ_2 be the average of the $MMRE$ for quantitative data and interval-valued data, respectively. The null(H_0) and alternative (H_1) hypotheses are:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

Tables 4, 6 and 8 present the comparison between regression methods based on the *p-value* of the statistics tests.

4.1 Results considering $NPhd$ and $NMaster$ explanatory variables

Table 3 presents the average and standard deviation of the $MMRE$ for IRR, IKR, ILR, $MM : CK + RL$, $MM : CL+RK$, $MM : CK+RR$ and $MM : CR+RK$ models. The predict variables $NPhd$(PhD guidelines) and $NMaster$ (Master guidelines) explain the response variable, that is, number the publications of the researches of computer science ($NPub$). The comparison between each two methods is achieved based on the wilcox-test for a difference in mean of the $MMRE$ with independent samples at the significance 5%. Table 4 shows the wilcox-test statistics computed in this study. From the values in Tables 3 and 4 some remarks are extracted.

- The IRR and $MM : CR+RK$ regression models have the best prediction performances due to the presence of *outliers* in the center of intervals as it is showed in Figure 1. Thus, the use robust regression in the center of the intervals is indicated. However, it can be observed that the $MM : CR+RK$ should be preferred because the mathematical coherence for intervals;
- The regression mixture models based on nonparametric functions to model center data $MM : CK + RR$ and $MM : CK + RL$ have the worst performances. The ILR and IKR methods have similar performances;
- The p-values in Table 4 support the previous remarks.

Table 3: $MMRE$ for interval data set.

Models	$Average \pm St.Deviation$
ILR	4.47943 ± 1.675311
IRR	**2.749475 ± 1.163158**
IKR	4.282684 ± 4.660366
MM:CK+RL	7.460052 ± 7.410414
MM:CL+RK	3.262939 ± 1.251656
MM:CK+RR	7.530917 ± 7.719749
MM:CR+RK	**2.998639 ± 0.874532**

Table 4: Comparison between regression methods.

Comparison	*p-value*
$\mu(MM:CRRK) \times \mu(IRR)$	6.9508×10^{-08}
$\mu(MM:CRRK) \times \mu(MM:CKRL)$	6.0250×10^{-69}
$\mu(MM:CRRK) \times \mu(MM:CLRK)$	2.5151×10^{-08}
$\mu(MM:CRRK) \times \mu(MM:CKRR)$	3.8637×10^{-66}
$\mu(MM:CRRK) \times \mu(ILR)$	2.8836×10^{-114}
$\mu(MM:CRRK) \times \mu(IKR)$	1.8853×10^{-17}

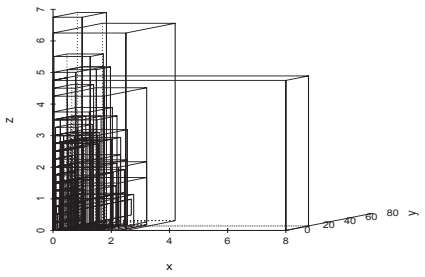

Fig. 1: Interval plot: $NPhd$ (X), $NMaster$ (Z) and $NPub$ (Y).

4.2 Results considering $NPhd$ and $NScientific$ explanatory variables

Table 5 shows the average and the standard deviation of the $MMRE$ for IRR, IKR, ILR, $MM:CK+RL$, $MM:CL+RK$, $MM:CK+RR$ and $MM:CR+RK$ models. The predict variables $NPhd$ (Phd guidelines finished) and $NScientific$(UR guidelines finished) explain the response variable, that is, number the publications of the researches of science computer ($NPub$). The comparison between each two methods is achieved based on the wilcox-test for a difference in mean of the $MMRE$ with independent samples at the significance 5%. Table 6 shows the wilcox-test statistics computed in this study. From the values in Tables 5 and 6 some remarks are extracted.

- The $MM:CR+RK$ regression mixture model proposed in this paper exhibited the best values of performance. As expected, this model consider a parametric function for the center that is sensitive to outliers and a nonparametric function for the range guaranteed the mathematical coherence for intervals;
- The ILR, IKR and IRR regression models and $MM:CL+RK$ have similar performances in terms of $MMRE$. However, it can be observed that the $MM:CL+RK$ regression mixture model should be preferred because the mathematical coherence for intervals.
- The $MM:CK+RL$ and $MM:CK+RR$ regression mixture models have the worst performance among all the regression models because there are remote rectangles as it is displayed in Figure 2;
- The p-values in Table 6 support the previous remarks.

Table 5: $MMRE$ for interval data set.

Models	$Average \pm St.Deviation$
ILR	5.327373 ± 1.919245
IKR	4.901354 ± 5.154982
IRR	3.739869 ± 1.425767
MM:CK+RL	8.225326 ± 6.470068
MM:CL+RK	5.019499 ± 1.996001
MM:CK+RR	8.368377 ± 7.055955
MM:CR+RK	**2.619647 ± 1.513572**

Table 6: Comparison between regression methods.

Comparison	*p-value*
$\mu(MM:CRRK) \times \mu(IRR)$	6.9508×10^{-08}
$\mu(MM:CRRK) \times \mu(MM:CKRL)$	6.0250×10^{-69}
$\mu(MM:CRRK) \times \mu(MM:CLRK)$	2.5151×10^{-08}
$\mu(MM:CRRK) \times \mu(MM:CKRR)$	3.8637×10^{-66}
$\mu(MM:CRRK) \times \mu(ILR)$	2.8836×10^{-114}
$\mu(MM:CRRK) \times \mu(IKR)$	1.8853×10^{-17}

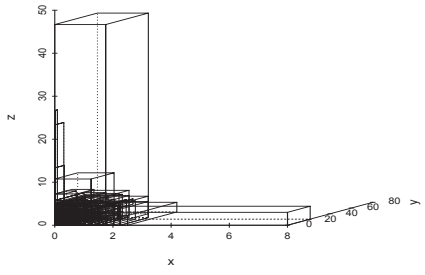

Fig. 2: Interval plot: $NPhd$ (X), $NScientific$ (Z) and $NPub$ (Y).

4.3 Results considering $NMaster$ and $NScientific$ explanatory variables

Table 7 presents the average and the standard deviation of the $MMRE$ for IRR, IKR, ILR, $MM:CK+RL$,$MM:CL+RK$, $MM:CK+RR$ and $MM:CR+RK$ models. The predict variables $NMaster$ (Master guidelines)

and $NScientific$(UR guidelines) explain the response variable, that is, number the publications of the researches of computer science ($NPub$). The comparison between each two methods is achieved based on the wilcox-test for a difference in mean of the $MMRE$ with independent samples at the significance 5%. Table 8 shows the wilcox-test statistics computed in this study. From the values in Tables 7 and 8 some remarks are extracted.

- The IKR, $MM : CL + RK$ and $MM : CR + RK$ regression mixture models are similar in terms of $MMRE$. However, it can be observed that the $MM : CR + RK$ and $MM : CL + RK$ regression mixture models should be preferred because the mathematical coherence for intervals. As expect, $MM : CL + RK$ is the best option in terms $MMRE$;
- These results show that the IRR and ILR models in the SDA literature have similar performance;
- The $MM : CK + RR$ and $MM : CK + RL$ regression mixture models have worst performance in terms $MMRE$ because the parametric form exists in the center of the intervals illustrated in the Figure 3 highlights the presence of a parametric form between the explanatory and response variables.

Table 7: $MMRE$ for interval data set.

Models	$Average \pm St.Deviation$
ILR	3.995003 ± 1.395748
IKR	4.878269 ± 6.794949
IRR	3.38593 ± 1.434064
MM:CK+RL	8.967796 ± 9.944816
MM:CL+RK	**2.643926 ± 1.068585**
MM:CK+RR	8.645091 ± 10.34362
MM:CR+RK	3.219962 ± 1.279143

Table 8: Comparison between regression methods.

Comparison	*p-value*
$\mu(MM : CLRK) \times \mu(IRR)$	5.7689×10^{-38}
$\mu(MM : CLRK) \times \mu(MM : CRRK)$	2.5336×10^{-27}
$\mu(MM : CLRK) \times \mu(MM : CKRL)$	1.2787×10^{-75}
$\mu(MM : CLRK) \times \mu(MM : CKRR)$	6.4910×10^{-65}
$\mu(MM : CLRK) \times \mu(ILR)$	6.2081×10^{-114}
$\mu(MM : CLRK) \times \mu(IKR)$	5.9507×10^{-24}

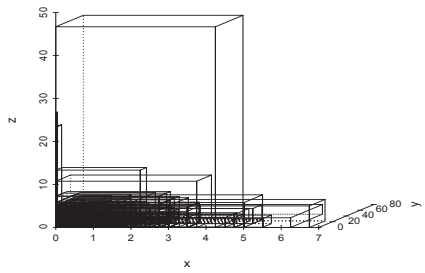

Fig. 3: Interval plot: $NMaster$ (X), $NScientificr$ (Z) and $NPub$ (Y).

4.4 Evaluating predicted interval

Figure 4 illustrates the predicted intervals versus real intervals. The predicted intervals are obtained by the $MM : CR+RK$ method from a test data set (scientific production) based on $NPhd$ and $NMaster$ explanatory variables. Figure 5 presents the predicted intervals based on $NPhd$ and $NScientific$ explanatory variables and $MM : CR + RK$ method. Figure 6 exhibits the predicted intervals based on $NMaster$ and $NScientific$ explanatory variables and $MM : CL + RK$ method.

As expect, the regression mixture methods have good linear fittings between predicted and real intervals. This means that the $MM : CR + RK$ method with $NPhd$ and $NMaster$ explanatory variables (Figure 4) and $MM : CL+ RK$ method with $NMaster$ and $NScientific$ explanatory variables (Figure 6) are adequacy to estimating scientific production data set.

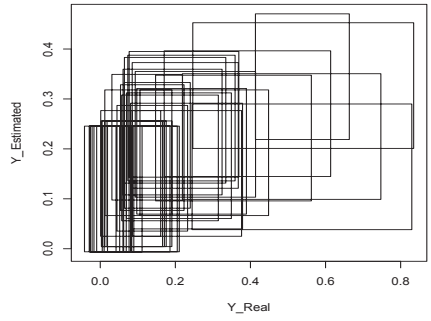

Fig. 4: Interval plot: Estimated Y versus real Y based on the $MM : CR + RK$.

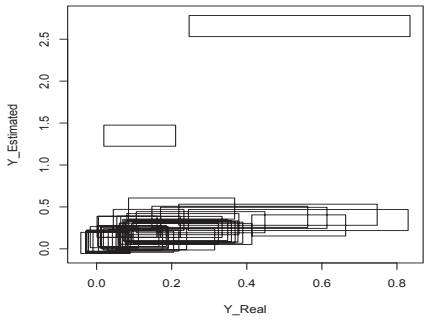

Fig. 5: Interval plot: Estimated Y versus real Y based on the $MM : CR + RK$.

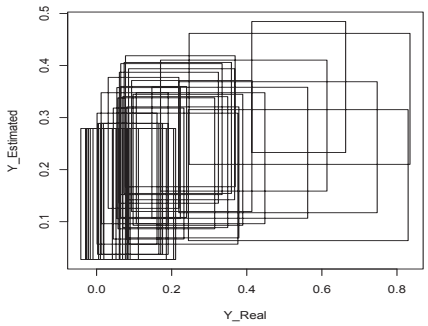

Fig. 6: Interval plot: Estimated Y versus real Y based on the $MM : CL + RK$.

5. Conclusion

This work presented a study of Brazilian scientific production based on tools of the Symbolic Data Analysis (SDA) and regression mixture models. The data set is originally formed by researchers from different centers of research. Tools of Symbolic Data Analysis are applied in order to model the information regarding variables that take into account variability. So, new units are obtained and they are described by interval data. Each unit represents aggregated data under the same institute and subject of research. The aggregation process provided the following advantages: reduction of the size of the base, assurance of the privacy of individuals.

The regression mixture models use kernel functions in center or range providing a versatile method of exploring a general relationship between interval variables and gives good predictions of interval observations yet to be made without reference to a fixed parametric model. Furthermore, the regression mixture models utilize robust regression in center or range as it is an alternative to least squares estimation in the presence of outliers.

References

[1] Billard L. and Diday E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, West Sussex, England, 2006.

[2] Diday E. and Noirhomme-Fraiture M. *Symbolic Data Analysis and the SODAS Software*. John Wesley & Sons, ltd, 2008.

[3] Fagundes, R.A.A., Souza, R.M.C.R. and Cysneiros, F.J.A. Interval Kernel Regression, vol. 128,371-388, 2014.

[4] Fagundes, R.A.A., Souza, R.M.C.R. and Cysneiros, F.J.A. Robust Regression with Application to Symbolic Interval Data, vol. 26,564-573, 2013.

[5] Neto, E.A.L., De Carvalho, F.A.T.: Symbolic Approach to Analyzing Administrative Management. The Electronic Journal of Symbolic Data Analysis 1(1), 1âĂŞ13, 2002.

[6] Silva, A.D., Lechevallier, Y., De Carvalho, F.A.T., Trousse, B.: Mining web usage data for discovering navigation clusters. In: IEEE Symposium on Computers and Communications,910âĂŞ915, 2006.

[7] Zuccolotto, P.: Principal components of sample estimates: an approach through symbolic data analysis. Applied and Metallurgical Statistics 16, 173âĂŞ192, 2006.

[8] Montgomery D.C., Peck E.A. and Vining G. G. *Introduction to Linear Regression Analysis*, Wisley-Interscince. 2006.

[9] Pimentel,B.A., Nóbrega, J.P. & Souza,R. M.C.R. Using Weighted Clustering and Symbolic Data to Evaluate InstitutesâĂŹs Scientific Production, 435-442, 2012.

[10] Pimentel,B.A. & Souza,R. M.C.R. A weighted multivariate Fuzzy C-Means method in interval-valued scientific production data, Expert Systems with Applications, 41, 3223-3236, 2014.

[11] Lima Neto E.A. and De Carvalho F.A.T. Centre and Range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis*, 52:1500-1515, 2008.

[12] Sheather, S.J. and Jones, M.C., A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society*, Series B, 53:683-690, 1991.