Symbolic Data Analysis: Taking Variability in Data into Account

Regression for Symbolic Data

Sónia Dias
I.P. Viana do Castelo & LIAAD-INESC TEC, Univ. Porto

Paula Brito
Fac. Economia & LIAAD-INESC TEC, Univ. Porto

ECI 2015 - Buenos Aires

Symbolic Data Analysis: Taking Variability in Data into Account

Methods for the Analysis of Symbolic Data

Regression

- *Linear Regression for interval-valued variables*
- *Linear Regression for histogram-valued variables*

Outline

- ***Linear regression for interval-valued variables***
  - State-of-the-art
  - *Introduction to intervals*
  - *Exploration of some proposed models*
  - *Examples*

- ***Linear regression for histogram-valued variables***
  - State-of-the-art
  - *Introduction to distributions*
  - *Exploration of proposed models*
  - *Examples*

Linear Regression for interval-valued variables
State-of-the-art

- **METHODS BASED IN SYMBOLIC COVARIANCE DEFINITIONS** (Billard and Diday, 2000;2006; Xu, 2010)

- **MINMAX METHOD** (Billard and Diday, 2002)

- **CENTER AND RANGE METHOD** (Lima Neto and De Carvalho,2008)

- **CENTER AND RANGE LEAST ABSOLUTE DEVIATION REGRESSION METHOD** (Maia and Carvalho, 2008)

- **CONSTRAINED CENTER AND RANGE METHOD** (Lima Neto and De Carvalho, 2010)

- **LASSO IR METHOD** (Giordani, 2014)

- **BIVARIANTE SYMBOLIC REGRESSION MODELS** (Lima Neto *et al*,2011)

-**LINEAR REGRESSION MODELS FOR SYMBOLIC INTERVAL DATA USING PSO ALGORITHM** (Yang *et al*, 2011)

- **MONTE CARLO METHOD** (Ahn *et al*,2012)

- **RADIAL BASIS FUNCTION NETWORKS** (Su *et al*, 2012)

- **COPULA INTERVAL REGRESSION METHOD** (Neto *et al*, 2012)

- **INTERVAL DISTRIBUTIONAL MODEL** (Dias and Brito, in study)

Linear Regression for interval-valued variables
Introduction

For each observation j the interval observations Y_j may be represented in different ways:

INTERVALS

- defined from the **bounds** $Y_j = \left[\underline{I}_{Y_j}, \overline{I}_{Y_j} \right]$ with $\underline{I}_{Y(j)} \leq \overline{I}_{Y(j)}$
- defined from the **centers and half-ranges** $I_{Y(j)} = \left[c_{Y(j)} - r_{Y(j)}, c_{Y(j)} + r_{Y(j)} \right]$ where

$c_{Y(j)} = \dfrac{\overline{I}_{Y(j)} + \underline{I}_{Y(j)}}{2}$ center of the interval; $r_{Y(j)} = \dfrac{\overline{I}_{Y(j)} - \underline{I}_{Y(j)}}{2}, \quad r_{Y(j)} \geq 0$ half range of the interval

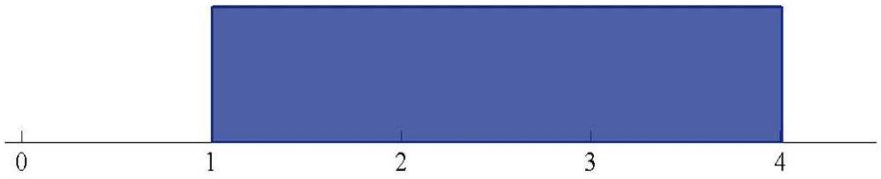
PARAMETERIZATION OF THE INTERVALS/ QUANTILE FUNCTONS

Assuming the Uniform distribution within the intervals:

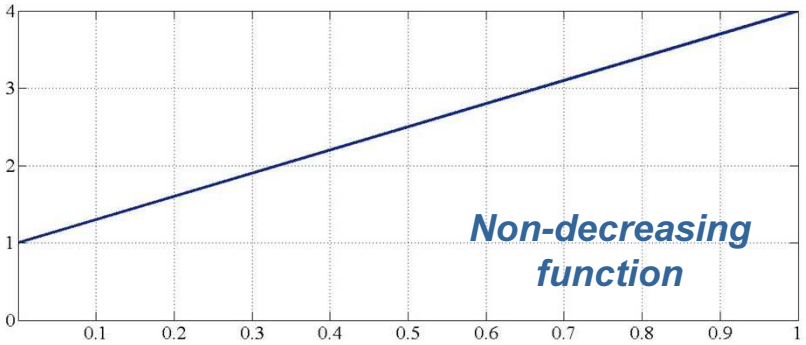
- defined from the **bounds** $\Psi^{-1}_{Y(j)}(t) = \underline{I}_{Y(j)} + \left(\overline{I}_{Y(j)} - \underline{I}_{Y(j)} \right) t, \quad 0 \leq t \leq 1$
- defined from the **centers and half-ranges** $\Psi^{-1}_{Y(j)}(t) = c_{Y(j)} + r_{Y(j)} (2t - 1), \quad 0 \leq t \leq 1$

Linear Regression for interval-valued variables
Introduction

Representations of the intervals



$$I_X = [1,4]$$



$$\Psi_X^{-1}(t) = 1 + 3t, \quad t \in [0,1]$$

In an interval $\left[\underline{I}_Y, \overline{I}_Y\right]$ we have always $\underline{I}_Y \leq \overline{I}_Y$. Consequently, the quantile function that represent the interval is a non decreasing function with domain [0,1].

Linear Regression for interval-valued variables
The Center Method (CM)

Billard and Diday, 2000. *Regression analysis for interval-valued data*. Proceedings of IFCS'00, pp.369-374. Springer.

Linear regression relation: $c_{Y(j)} = b_0 + b_1 c_{X_1(j)} + \ldots + b_p c_{X_p(j)} + e^c(j)$

Prediction of the intervals: $I_{\hat{Y}(j)} = \left[\underline{I}_{\hat{Y}(j)}, \overline{I}_{\hat{Y}(j)}\right]$ with $\underline{I}_{\hat{Y}(j)} = min\left\{b_0 + \sum_{k=1}^{p} b_k \underline{I}_{X_k(j)}, b_0 + \sum_{k=1}^{p} b_k \overline{I}_{X_k(j)}\right\}$

$$\overline{I}_{\hat{Y}(j)} = max\left\{b_0 + \sum_{k=1}^{p} b_k \underline{I}_{X_k(j)}, b_0 + \sum_{k=1}^{p} b_k \overline{I}_{X_k(j)}\right\}$$

- The coefficients of the model are estimated by applying the classical model to the mid-point of the intervals;
- Estimates separately the bounds of the interval;
- To write the estimated interval we have to consider the lower value for the lower bound and higher for the upper bound of the interval;
- The estimation of the parameters may be obtained by an adaptation of the solution obtained by the Least Square estimation method for the classical linear model, where symbolic definitions of variance and covariance are used;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables
Min Max Method (MinMax)

Billard and Diday, 2002. *Symbolic regression analysis*. Proceedings of IFCS'02, pp.281-288. Springer.

Linear regression relation :

$$\begin{cases} \underline{I}_{Y(j)} = b_0^L + b_1^L \underline{I}_{X_1(j)} + \ldots + b_p^L \underline{I}_{X_p(j)} + \underline{e}(j) \\ \overline{I}_{Y(j)} = b_0^U + b_1^U \overline{I}_{X_1(j)} + \ldots + b_p^U \overline{I}_{X_p(j)} + \overline{e}(j) \end{cases}$$

Prediction of the intervals: $I_{\hat{Y}(j)} = \left[\underline{I}_{\hat{Y}(j)}, \overline{I}_{\hat{Y}(j)}\right]$ with

$$\underline{I}_{\hat{Y}(j)} = min\left\{b_0^L + \sum_{k=1}^{p} b_k^L \underline{I}_{X_k(j)}, b_0^U + \sum_{k=1}^{p} b_k^U \overline{I}_{X_k(j)}\right\}$$

$$\overline{I}_{\hat{Y}(j)} = max\left\{b_0^L + \sum_{k=1}^{p} b_k^L \underline{I}_{X_k(j)}, b_0^U + \sum_{k=1}^{p} b_k^U \overline{I}_{X_k(j)}\right\}$$

- Requires the adjustment of two linear regression models, for the lower and upper bounds of the interval;
- The coefficients of the model are estimated by applying the classical model to the lower and upper bounds of the interval;
- The estimated value for the upper bound of the interval may be smaller than the lower. This can happen if there are negative coefficients in the model;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables
The Center and Range Method (CRM)

Lima Neto and de Carvalho, 2008. *Center and Range method for fitting a linear regression model to symbolic interval data*. Computational Statistics & Data Analysis 52 (3), 1500-1515.

Linear regression relation :

$$\begin{cases} c_{Y(j)} = b_0^c + b_1^c c_{X_1(j)} + \ldots + b_p^c c_{X_p(j)} + e^c(j) \\ r_{Y(j)} = b_0^r + b_1^r r_{X_1(j)} + \ldots + b_p^r r_{X_p(j)} + e^r(j) \end{cases},$$

Prediction of the intervals:

$$I_{\hat{Y}(j)} = \left[\underline{I}_{\hat{Y}(j)}, \overline{I}_{\hat{Y}(j)}\right] \text{ with } \underline{I}_{\hat{Y}(j)} = min\left\{c_{\hat{Y}(j)} - r_{\hat{Y}(j)}, c_{\hat{Y}(j)} + r_{\hat{Y}(j)}\right\}$$

$$\overline{I}_{\hat{Y}(j)} = max\left\{c_{\hat{Y}(j)} - r_{\hat{Y}(j)}, c_{\hat{Y}(j)} + r_{\hat{Y}(j)}\right\}$$

- Requires the adjustment of two linear regression models, for the mid-point and half range of the interval;
- The coefficients of the model are estimated by applying the classical model to the mid-point and half range of the interval;
- The estimated value for the range of the interval may be negative. This can happen if there are negative coefficients in the model that estimates the half range;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables
The Constrained Centre and Range Method (CCRM)

Lima Neto and de Carvalho, 2010. *Constrained linear regression models for symbolic interval-valued variables*. Computational Statistics & Data Analysis 54 (2), 333-347.

Linear regression relation :

$$\begin{cases} c_{Y(j)} = b_0^c + b_1^c c_{X_1(j)} + \ldots + b_p^c c_{X_p(j)} + e^c(j) \\ r_{Y(j)} = b_0^r + b_1^r r_{X_1(j)} + \ldots + b_p^r r_{X_p(j)} + e^r(j) \end{cases},$$

$$with \quad b_k^r \geq 0$$

Prediction of the intervals: $I_{\hat{Y}(j)} = \left[c_{\hat{Y}(j)} - r_{\hat{Y}(j)}, c_{\hat{Y}(j)} + r_{\hat{Y}(j)} \right]$

- The mid-points and half ranges of the intervals are estimated independently;
- The coefficients of the centers model are estimated by applying the classical model to the mid-points of the intervals;
- The coefficients of the half ranges model are estimated using the Lawson and Hanson's algorithm (Lawson and Hanson, 1995).
- Because of the restriction imposed, the linear relation between the half range of the intervals has to be always positive;
- Descriptive linear regression model;
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables
Bivariate Symbolic Method

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

Interval-valued variables: $Y = (Y_1, Y_2)$ and $X_k = (X_{1k}, X_{2k})$

- defined as bivariate quantitative vectors composed by two one-dimensional variables;
- the components are quantitative classical variables that may be represented by the bounds (lower and upper) or the center and half range;
- the response interval-valued variable Y has a random nature.

Linear regression relation

Adaptation to interval-valued variables of the Bivariate Generalized Linear Model.

- It is assumed that the components of the bivariate vectors follow a bivariate exponential family;
- It is possible to guarantee the non negativity for the predicted values of the half ranges;
- Probabilistic linear regression model. Inference techniques can be considered.
- Information about the variability within the intervals is not taken into account.
- Available in R Package: iRegression 1.2.

Linear Regression for interval-valued variables
Interval Distributional Model (ID)

Dias, 2014. *Linear regression with empirical distributions*. Ph.D thesis, Chapter 5. Universidade do Porto, Portugal.

$\psi^{-1}_{X_1(j)}(t), \psi^{-1}_{X_2(j)}(t), \ldots, \psi^{-1}_{X_p(j)}(t), \psi^{-1}_{Y(j)}(t)$ quantile functions representing the intervals of the explicative and response interval-valued variables for observation *j*.

Linear regression relation:

$$\psi^{-1}{}_{Y(j)}(t) = v + a_1\, \psi^{-1}{}_{X_1(j)}(t) - \beta_1\, \psi^{-1}{}_{X_1(j)}(1-t) + \ldots + a_p\, \psi^{-1}{}_{X_p(j)}(t) - b_p \psi^{-1}{}_{X_p(j)}(1-t) + e(j) \Leftrightarrow$$

$$\psi^{-1}{}_{Y(j)}(t) = v + \sum_{k=1}^{p}(a_k - b_k)\, c_{X_k(j)} + \sum_{k=1}^{p}(a_k + b_k)\, r_{X_k(j)}\ (2t-1) + e(j)$$

$$\text{with} \quad a_k, b_k \geq 0, \quad k = 1, \ldots, p \quad v \in IR \quad \text{e} \quad 0 \leq t \leq 1$$

The error $e(j)$, for each unit *j*, is a linear function, but not necessarily a quantile function.

Linear Regression for interval-valued variables
Interval Distributional Model (ID)

Dias, 2014. *Linear regression with empirical distributions*. Ph.D thesis, Chapter 5. Universidade do Porto, Portugal.

- As the model uses quantile functions to represent the intervals, the distributions within them are considered;
- In the present model, the Uniform distribution is assumed in each observed interval, however other distributions may be considered;
- The linear relations between the centers and half ranges induced by the model are different, although related;
- The non-negative parameters of the model are determined by solving a quadratic optimization problem, subject to non-negativity constraints. The distance used to quantify the dissimilarity between predicted and observed quantile functions is the Mallows Distance (Uniform distribution is assumed)

$$D_M^2\left(\psi_{Y(j)}^{-1}(t),\psi_{\hat{Y}(j)}^{-1}(t)\right)=\left(c_{Y(j)}-c_{\hat{Y}(j)}\right)^2+\frac{1}{3}\left(r_{Y(j)}-r_{\hat{Y}(j)}\right)^2$$

- It is possible to estimate the quantile functions for the response variable directly from the model;
- A goodness-of-fit measure, Ω is derived from the model;
- Descriptive linear regression model;
- Information about the variability within the intervals is taken into account.

Linear Regression for interval-valued variables

Measure the dissimilarity between intervals

Root Mean Square Errors (Lima Neto and De Carvalho, 2008, 2010)

$$RMSE_U = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\bar{I}_{\hat{Y}}(j) - \bar{I}_Y(j)\right)^2}$$

$$RMSE_L = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\underline{I}_{\hat{Y}}(j) - \underline{I}_Y(j)\right)^2}$$

Error measure defined with the Mallows Distance, assuming the Uniform distribution within the intervals (Irpino and Verde, 2015)

$$RMSE_M = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left[\left(c_{Y(j)} - c_{\hat{Y}(j)}\right)^2 + \frac{1}{3}\left(r_{Y(j)} - r_{\hat{Y}(j)}\right)^2\right]}$$

Linear Regression for interval-valued variables

Example: Hematocrit and hemoglobin study

Interval data:

Ranges of the values of hematocrit and hemoglobin of 16 patients in a hospital.

Variables:

Response variable: Y - hematocrit

Explicative variables: X - hemoglobin

Higher level units: pacients

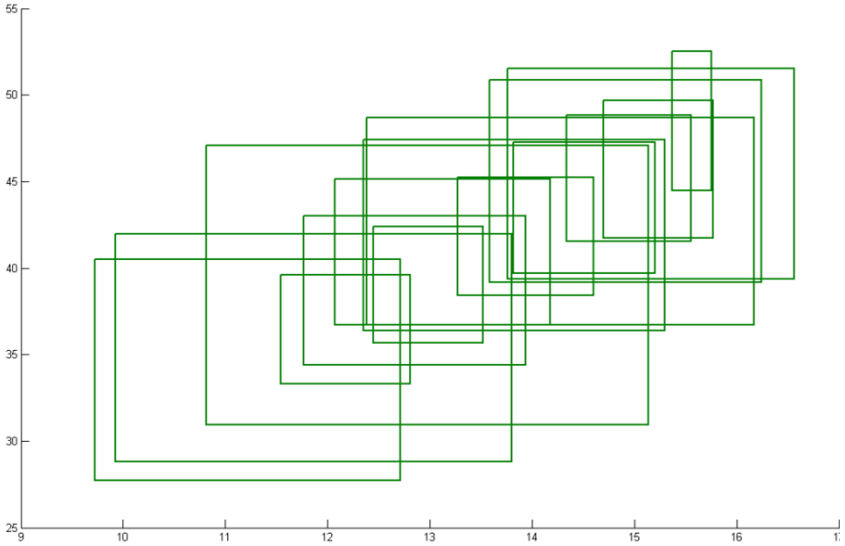
Goal: Study the ranges of hematocrit from the ranges of hemoglobin.

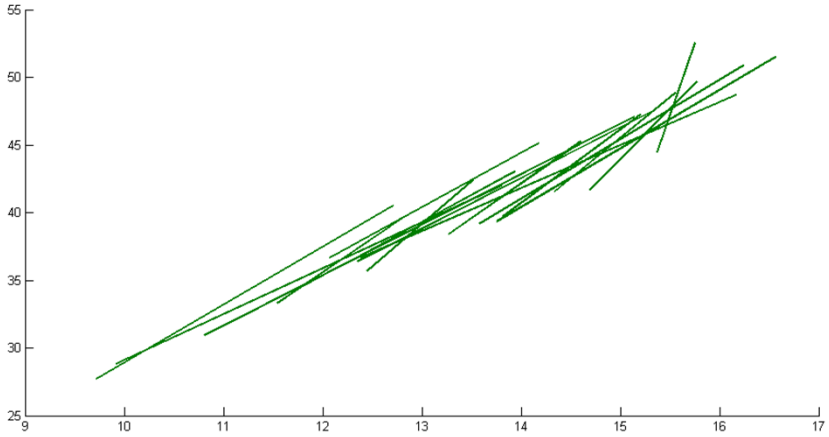
j	X	Y
1	[11.54;12.806]	[33.296; 39.601]
2	[12.075;14.177]	[36.694; 45.123]
3	[12.384;16.169]	[36.699; 48.685]
4	[12.354;15.298]	[36.386; 47.412]
5	[13.581;16.242]	[39.190; 50.866]
6	[13.819;15.203]	[39.701; 47.246]
7	[14.341;15.554]	[41.560; 48.814]
8	[13.274;14.601]	[38.404; 45.228]
9	[9.9220;13.801]	[28.831; 41.980]
10	[15.374;16.755]	[44.481; 52.536]
11	[9.722;12.712]	[27.713; 40.499]
12	[11.767;13.936]	[34.405; 43.027]
13	[10.812;15.142]	[30.919; 47.091]
14	[13.760;16.562]	[39.351; 51.510]
15	[14.698;15.769]	[41.710; 49.678]
16	[12.448;13.519]	[35.674; 42.382]

Linear Regression for interval-valued variables

Example: Hematocrit and hemoglobin study

Scatter plots of the data





RGui - [R Console]

File Edit View Misc Packages Windows Help

Package 'iRegression'

July 26, 2012

Type Package

Title Regression methods for interval-valued variables

Version 1.2

Date 2011-06-01

Depends mgcv

Author Eufrasio de A. Lima Neto with contribution from Claudio A. Vasconcelos

Maintainer Eufrasio de A. Lima Neto <eufrasio@de.ufpb.br>

Description This package contains some important regression methods for interval-valued variables. For each method, it is available the fitted values, residuals and some goodness-of-fit measures.

License GPL (>= 2)

LazyLoad yes

Repository CRAN

Date/Publication 2012-07-26 19:51:15

R topics documented:

iRegression-package 2
bivar 3
Cardiological.CR 5
Cardiological.MinMax 6
ccrm 7
cm 8
coef.bivar 10
coef.ccrm 11
coef.crm 11
coef.MinMax 12
crm 12

RGui - [R Console]

File Edit View Misc Packages Windows Help

```
> cm <- cm(MinY~MinX,MaxY~MaxX,data=BillardSangueMinMax)
> cm
Call:
cm.formula(formula1 = MinY ~ MinX, formula2 = MaxY ~ MaxX, data = BillardSangueMinMax)

$coefficients
(Intercept)           1
 -0.8146172   3.0805942

$sigma
[1] 0.5855333

$df
[1] 14

$fitted.values.l
 [1] 34.75084 36.38356 37.33546 37.24304 41.02293 41.75611 43.36418 40.07719 29.75104 46.54644

$fitted.values.u
 [1] 38.63547 42.85897 48.99551 46.31231 49.22039 46.01966 47.10094 44.16514 41.70066 47.72014

$residuals.l
 [1] -1.4548424  0.3104427 -0.6364609 -0.8570431 -1.8329321 -2.0551135 -1.8041837 -1.6731897 -0
[16] -1.8586189

$residuals.u
 [1]  0.9655284  2.2640338 -0.3105098  1.0996877  1.6456068  1.2263441  1.7130556  1.0628618  0
[16]  1.5500647
```

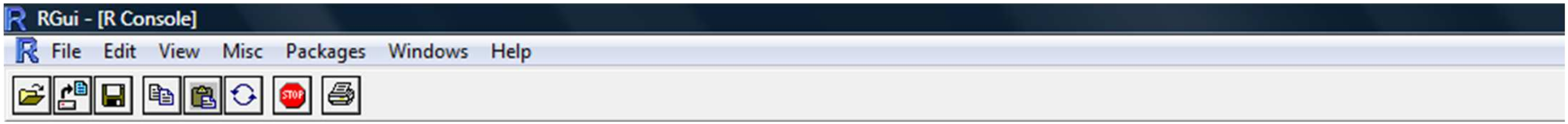
RGui - [R Console]

File Edit View Misc Packages Windows Help

```
> summary(cm)
Call:
cm.formula(formula1 = MinY ~ MinX, formula2 = MaxY ~ MaxX, data = BillardSangueMinMax)

              Estimate    StdErr
(Intercept) -0.8146172 1.5826708
1            3.0805942 0.1148388

RMSE.L:
[1] 1.651846
RMSE.U:
[1] 1.832516
```




```
> ccrm <- ccrm(CY~CX,RY~RX,data=BillardSangueCR)
> summary(ccrm)
Call:
ccrm.formula(formula1 = CY ~ CX, formula2 = RY ~ RX, data = BillardSangueCR)

            Estimate.C  StdErr.C
(Intercept) -0.8146172 1.5826708
CX           3.0805942 0.1148388

            Estimate.R  StdErr.R
(Intercept)   4.751821 0.6524150
RX            2.279818 0.2623395

RMSE.L:
[1] 0.656672

RMSE.U:
[1] 0.891152
```

Linear Regression for interval-valued variables

Example: Forest Fires

Original data:

- Collected from January 2000 to December 2003;
- The first database was collected by the responsible for the Montesinho fire occurrences. Several features were registered: date, spatial location, the total burned area,…
- The second database was collected by the Bragança Polytechnic Institute, containing several weather observations that were recorded by a meteorological station located in Montesinho park.

Original variables selected:

Response variable :

area - burned area of the forest *(in ha);*
Transformed in *LNarea=ln(area+1)*

Three explicative variables:

temp - temperature in Celsius degrees;
wind - wind speed in *km/h*;
rh - relative humidity in percentage;

Linear Regression for interval-valued variables

Example: Forest Fires

Temporal aggregation: by month

Higher level units: months

Goal: Study the burned area of the forest of the Montesinho natural park.

Some considerations:

For this study we considered only the months and the records in which forest fires occurred. For this reason January and November were eliminated.

Months	area	temp	wind	rh
Feb	[0.74;3.97]	[4.6;12.4]	[0.9; 9.4]	[35;82]
Mar	[0.67;3.63]	[5.3;17]	[0.9; 9.4]	[26;70]
Apr	[1.47;4.13]	[5.8;13.7]	[3.1; 9.4]	[33;64]
May	[3.68;3.68]	[18;18]	[4; 4]	[40;40]
June	[0.64;4.27]	[14.3;28]	[1.8; 9.4]	[34;79]
July	[0.31;5.63]	[13.4;30.2]	[0.9; 7.2]	[25;82]
Aug	[0.09;6.62]	[11.2;33.3]	[0.4; 8.9]	[22;88]
Sep	[0.29;7.0]	[10.1;29.6]	[0.9; 7.6]	[15;78]
Oct	[1.9;3.9]	[16.1;20.2]	[2.7; 4.5]	[25;45]
Dec	[1.9;3.2]	[2.2;5.1]	[4.9; 8.5]	[21;61]

Linear Regression for interval-valued variables

Example: Forest Fires

Models	Expressions that allows predicting the intervals
CM	$\hat{c}_{LNarea}(j) = 1.92 + 0.002c_{temp}(j) + 0.003c_{wind}(j) - 0.02c_{rh}(j)$
MinMax	$\hat{\underline{I}}_{LNarea}(j) = -0.39 + 0.01\underline{I}_{temp}(j) + 0.24\underline{I}_{wind}(j) + 0.02\underline{I}_{rh}(j)$ $\hat{\overline{I}}_{LNarea}(j) = 1.16 + 0.01\overline{I}_{temp}(j) - 0.04\overline{I}_{wind}(j) + 0.01\overline{I}_{rh}(j)$
CRM	$\hat{c}_{LNarea}(j) = 1.92 + 0.002c_{temp}(j) + 0.003c_{wind}(j) - 0.02c_{rh}(j)$ $\hat{r}_{LNarea}(j) = 0.01 + 0.07r_{temp}(j) - 0.01r_{wind}(j) + 0.01r_{rh}(j)$
CCRM	$\hat{c}_{LNarea}(j) = 1.92 + 0.002c_{temp}(j) + 0.003c_{wind}(j) - 0.02c_{rh}(j)$ $\hat{r}_{LNarea}(j) = 0.004 + 0.07r_{temp}(j) + 0.01r_{rh}(j)$
ID	$\psi^{-1}_{\hat{LNarea}(j)}(t) = 1.86 + 0.02\,\psi^{-1}_{temp(j)}(t) - 0.02\psi^{-1}_{temp(j)}(1-t) - 0.01\,\psi^{-1}_{rh(j)}(t), \quad t \in [0,1]$ $\hat{c}_{LNarea}(j) = 1.86 + 0.001c_{temp}(j) - 0.01c_{rh}(j); \quad \hat{r}_{LNarea}(j) = 0.04r_{temp}(j) + 0.01r_{rh}(j)$

Linear Regression for interval-valued variables

Example: Forest Fires

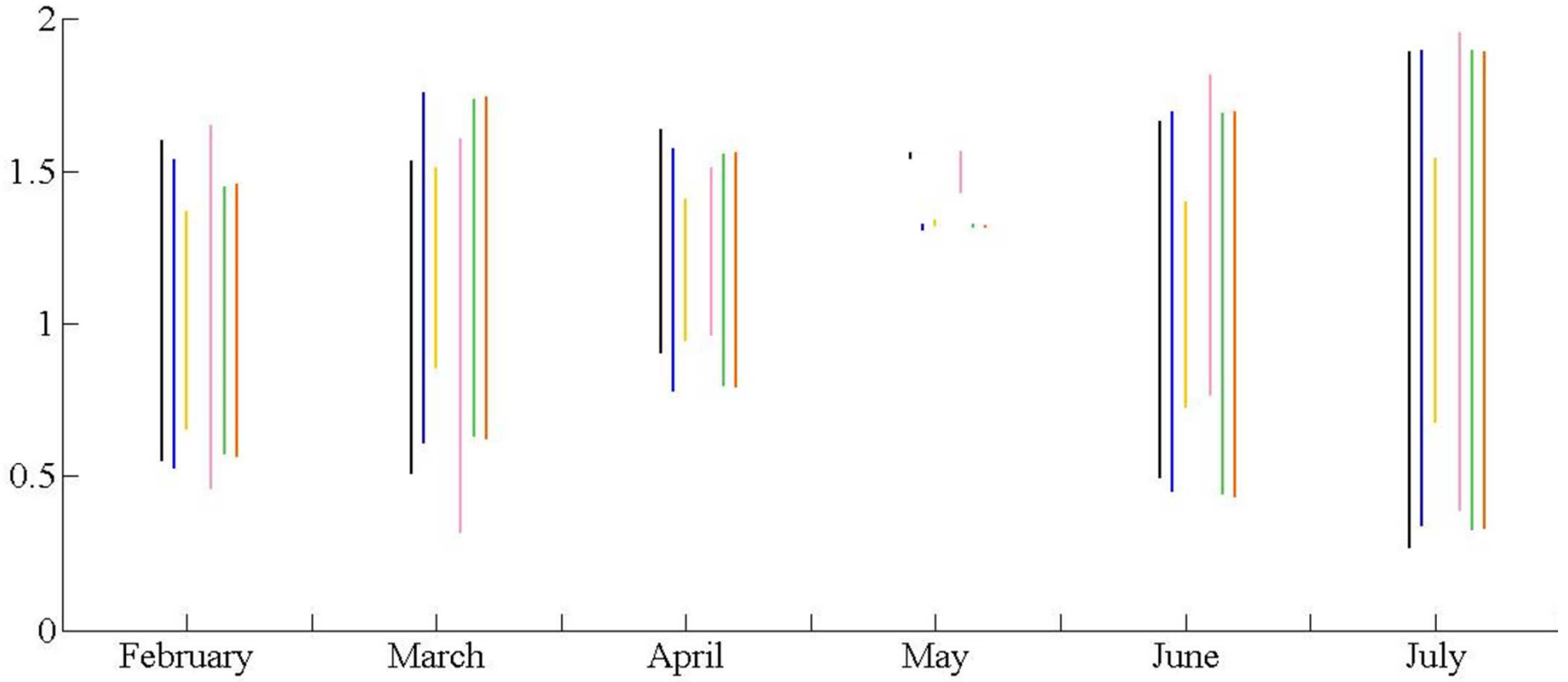
Evaluate the performance of the models

Models	$RMSE_L$	$RMSE_U$	$RMSE_M$	Mean Area
CM	0.3076	0.2676	0.1856	0.1504
MinMax	0.1481	0.0940	0.1044	0.0828
CRM	0.1030	0.1161	0.1038	0.0805
CCRM	0.1034	0.1156	0.1038	0.0804
ID	0.1106	0.1222	0.1066	0.0818

Linear Regression for interval-valued variables

Example: Forest Fires

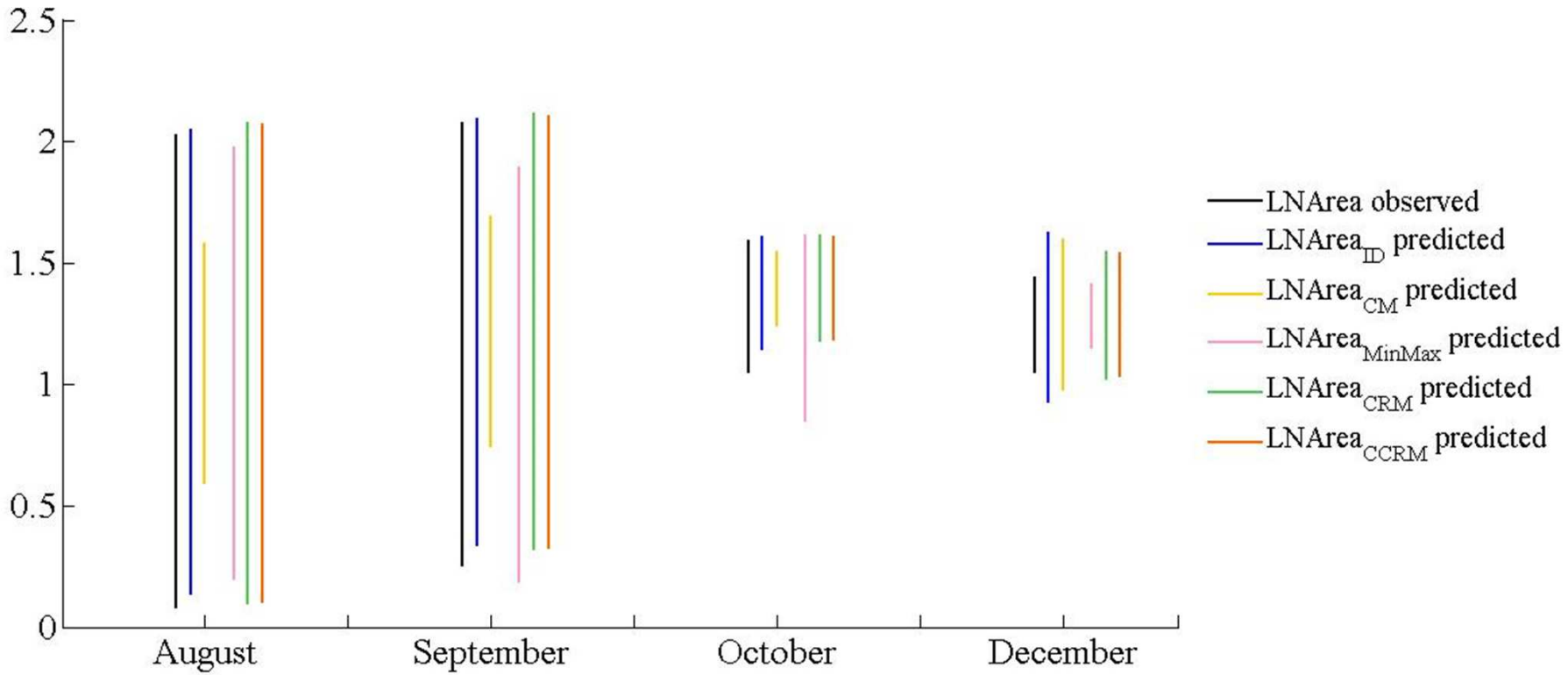
Observed and predicted intervals considering different methods



Linear Regression for interval-valued variables

Example: Forest Fires

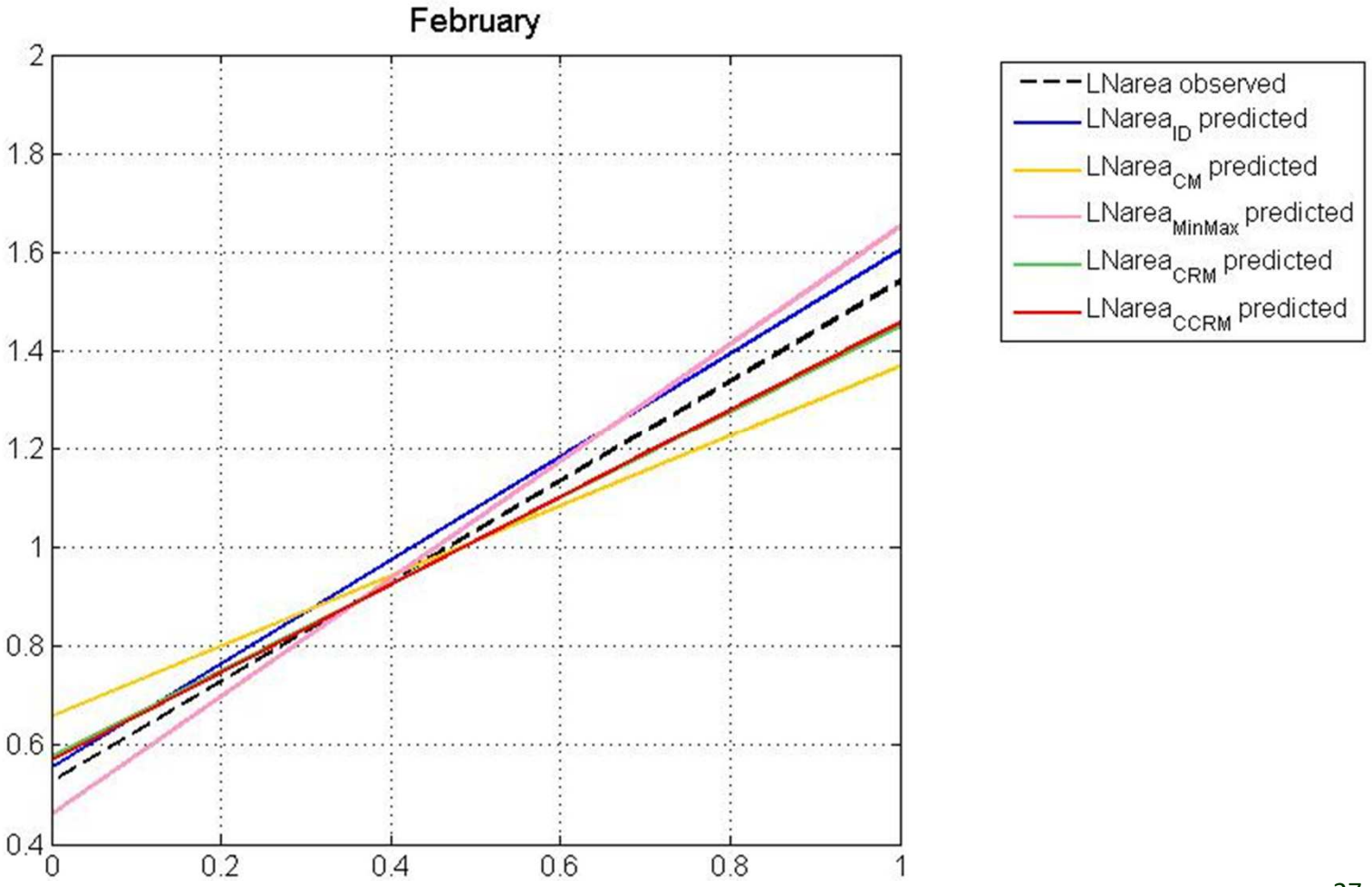
Observed and predicted intervals considering different methods



Linear Regression for interval-valued variables

Example: Forest Fires

Observed and predicted quantile functions considering different methods



Linear Regression for histogram-valued variables
State-of-the-art

- **METHODS BASED IN SYMBOLIC COVARIANCE DEFINITIONS** (Billard and Diday, 2006)

- **LINEAR REGRESSION FOR NUMERIC SYMBOLIC VARIABLES: AN ORDINARY LEAST SQUARES APPROACH BASEAD ON WASSERSTEIN DISTANCE** (Verde and Irpino, 2015)

- **LINEAR REGRESSION MODEL WITH HISTOGRAM-VALUED VARIABLES** (Dias and Brito, 2015)

Linear Regression for histogram-valued variables
Introduction

For each observation j the distribution observations Y_j may be represented in different ways:

HISTOGRAM

$$H_{Y(j)} = \left\{ \left[\underline{I}_{Y(j)_1}, \overline{I}_{Y(j)_1} \right[, p_{j1}; \left[\underline{I}_{Y(j)_2}, \overline{I}_{Y(j)_2} \right[, p_{j2}; \ldots; \left[\underline{I}_{Y(j)_n}, \overline{I}_{Y(j)_n} \right[, p_{jn_j} \right\},$$

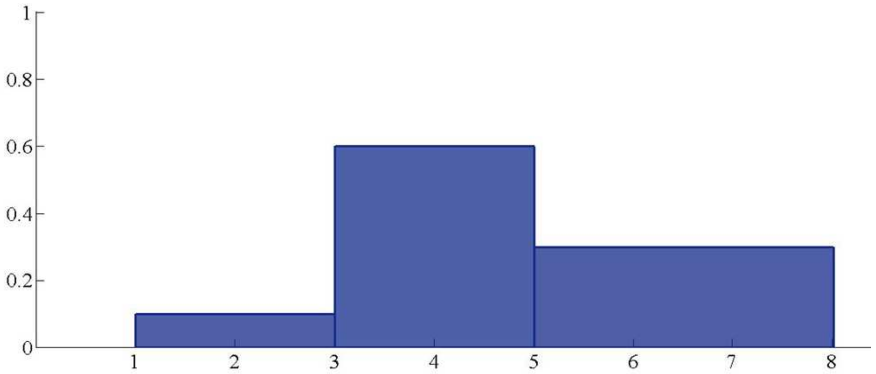
QUANTILE FUNCTION

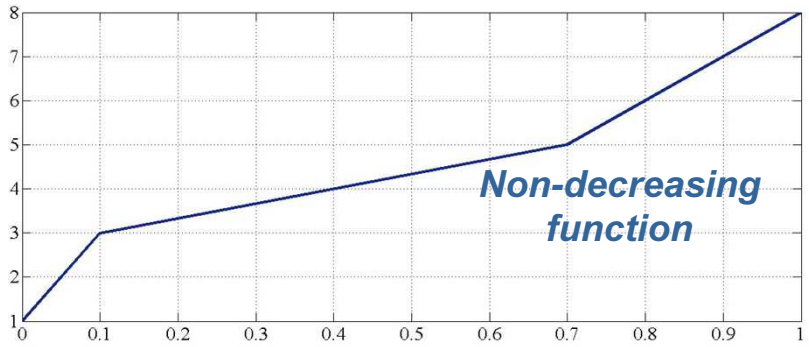
$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} \underline{I}_{Y(j)_1} + \dfrac{t}{w_{j1}}\left(\overline{I}_{Y(j)_1} - \underline{I}_{Y(j)_1}\right), & 0 \le t \le w_{j1} \\ \underline{I}_{Y(j)_2} + \dfrac{t - w_{j1}}{w_{j2} - w_{j1}}\left(\overline{I}_{Y(j)_2} - \underline{I}_{Y(j)_2}\right) & w_{j1} \le t \le w_{j2} \\ \vdots & \vdots \\ \underline{I}_{Y(j)_{n_j}} + \dfrac{t - w_{jn_j-1}}{1 - w_{jn_j-1}}\left(\overline{I}_{Y(j)_{n_j}} - \underline{I}_{Y(j)_{n_j}}\right) & w_{jn_j-1} \le t \le 1 \end{cases}$$

- n_j number of subintervals in the histogram for the j^{th} observation;
- $\underline{I}_{Y(j)_i} \le \overline{I}_{Y(j)_i}$
- $\sum_{i=1}^{n_j} p_{ji} = 1$
- $w_{jl} = \begin{cases} 0, & l = 0 \\ \sum_{h=1}^{l} p_{jh}, & l = 1, \ldots, n_j \end{cases}$
- It is assumed that within each sub-interval the values $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i} \right[$ are uniformly distributed.

Linear Regression for histogram-valued variables
Introduction

Representations of the distributions





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

$$\Psi_X^{-1}(t) = \begin{cases} 1+\frac{t}{0.1}\times 2 & 0 \le t < 0.1 \\ 3+\frac{t-0.1}{0.6}\times 2 & 0.1 \le t < 0.7 \\ 5+\frac{t-0.7}{0.3}\times 3 & 0.7 \le t \le 1 \end{cases}$$

In a histogram the lower bound of each subinterval is always less than or equal to the upper bound , $\underline{I}_{Y(j)i} \le \overline{I}_{Y(j)i}$ The lower bound of a subinterval is always greater or equal to the upper bound of the previous sub-interval , $\overline{I}_{Y(j)i} \le \underline{I}_{Y(j)i+1}$. **Consequently,** the quantile function that represents the distribution is a non-decreasing function in the domain [0,1].

Linear Regression for histogram-valued variables

Introduction

The observed empirical distributions are represented by quantile functions

Consequently

Work in the space where the elements are quantile functions

All involved functions are rewritten with an equal number of pieces and the domain of each piece has to be the same for all functions.
However, the corresponding histograms are not necessarily equiprobable histograms.

To rewrite all observations in the above conditions the Irpino and Verde (2006) process is used.

Linear Regression for histogram-valued variables
Introduction

Considerations

- p explicative histogram valued-variables X_k with $k \in \{1,\ldots,p\}$ and one response histogram valued-variable Y;

- m observations of each variable $X_k(j)$ and $Y(j)$, with $j \in \{1,\ldots,m\}$

- Each histogram $X_k(j)$ and $Y(j)$, for all $j \in \{1,\ldots,m\}$, is defined with n subintervals, $I_{X_k(j)_i}$ and $I_{Y(j)_i}$ with $i \in \{1,\ldots,n\}$, i.e.

$$Y(j) = \left\{ \left[\underline{I}_{Y(j)1}, \overline{I}_{Y(j)1} \right[, p_1; \ldots; \left[\underline{I}_{Y(j)n}, \overline{I}_{Y(j)n} \right], p_n \right\}$$

$$X_k(j) = \left\{ \left[\underline{I}_{X_k(j)1}, \overline{I}_{X_k(j)1} \right[, p_1; \ldots; \left[\underline{I}_{X_k(j)n}, \overline{I}_{X_k(j)n} \right], p_n \right\}$$

Linear Regression for histogram-valued variables
Introduction

Distance between distributions – Mallows distance

$\psi_X^{-1}(t), \psi_Y^{-1}(t)$: quantile functions that represent two distributions. Within each subinterval the Uniform distribution is assumed.

$$D_M^2\left(\psi_X^{-1}(t),\psi_Y^{-1}(t)\right)=\int_0^1\left(\psi_X^{-1}(t)-\psi_Y^{-1}(t)\right)^2dt=\sum_{j=1}^{m}\sum_{i=1}^{n}p_{ji}\left[\left(c_{X_i(j)}-c_{Y_i(j)}\right)^2+\frac{1}{3}\left(r_{X_i(j)}-r_{Y_i(j)}\right)^2\right]$$

$c_{X_i(j)}=\dfrac{\overline{I}_{X_i(j)}+\underline{I}_{X_i(j)}}{2}$; $c_{Y_i(j)}=\dfrac{\overline{I}_{Y_i(j)}+\underline{I}_{Y_i(j)}}{2}$ centers of the intervals;

$r_{X_i(j)}=\dfrac{\overline{I}_{X_i(j)}-\underline{I}_{X_i(j)}}{2}$; $r_{Y_i(j)}=\dfrac{\overline{I}_{Y_i(j)}-\underline{I}_{Y_i(j)}}{2}$ half-ranges of the intervals

Linear Regression for histogram-valued variables

Center method generalized to histogram-valued variables (CM)

Billard and Diday, 2006. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Chapter 6. John Wiley & Sons.
Dias, 2014. *Linear regression with empirical distributions*. Ph.D thesis, Chapter 3. Universidade do Porto, Portugal.

Linear regression relation: $\overline{Y}(j) = b_0 + b_1 \overline{X}_1(j) + \ldots + b_p \overline{X}_p(j) + e(j)$

where $\overline{Y}(j) = \sum_{i=1}^{n} c_{Y(j)} p_{ji}$ and $\overline{X_k}(j) = \sum_{i=1}^{n} c_{X_k(j)} p_{ji}$

Prediction of the histograms:

$$\hat{Y}(j) = \left\{ \left[\underline{I}_{\hat{Y}(j)1}, \overline{I}_{\hat{Y}(j)1} \right[, p_1; \ldots; \left[\underline{I}_{\hat{Y}(j)n}, \overline{I}_{\hat{Y}(j)n} \right], p_n \right\}$$

where to each subinterval $i \in \{1, \ldots, n\}$

$$\underline{I}_{\hat{Y}(j)i} = min \left\{ b_0 + \sum_{k=1}^{p} b_k \underline{I}_{X_k(j)i}, b_0 + \sum_{k=1}^{p} b_k \overline{I}_{X_k(j)i} \right\};$$

$$\overline{I}_{\hat{Y}(j)i} = max \left\{ b_0 + \sum_{k=1}^{p} b_k \underline{I}_{X_k(j)i}, b_0 + \sum_{k=1}^{p} b_k \overline{I}_{X_k(j)i} \right\}$$

Linear Regression for histogram-valued variables
Center method generalized to histogram-valued variables (CM)

Billard and Diday, 2006. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Chapter 6. John Wiley & Sons.
Dias, 2014. *Linear regression with empirical distributions*. Ph.D thesis, Chapter 3. Universidade do Porto, Portugal.

- The coefficients of the model are estimated by applying the classical model to the mean values of the observations of the histogram-valued variables $\overline{Y}(j), \overline{X}_k(j), k \in \{1,\ldots,p\}$
- The authors Billard and Diday don't present the process that may be used to predict the distributions for the response variables from the explicative variables;
- A process to predict the histograms is suggested in the *PhD thesis* of Dias, 2014. The presented process is a generalization of the process already used with the interval-valued variables and requires that all histograms are defined with the same number of subintervals;
- The method don’t predict distributions from other distributions;
- Descriptive linear regression model.

Linear Regression for histogram-valued variables
Irpino and Verde Model (IV)

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

$\psi^{-1}_{X_1(j)}(t), \psi^{-1}_{X_2(j)}(t), \ldots, \psi^{-1}_{X_p(j)}(t)$: quantile functions representing the distributions of the explicative histogram-valued variables for observation *j*. These distributions are decomposed:

- a part depending on the averages of the distributions $\overline{X_k}(j) = \sum_{i=1}^{n_j} c_{X_k(j)} p_{ji}$
- the centered quantile distributions $\psi^{c-1}{}_{X_k(j)}(t) = \psi^{-1}{}_{X_k(j)}(t) - \overline{X}_k(j)$

Linear regression relation:

$$\psi^{-1}{}_{Y(j)}(t) = b_0 + \sum_{k=1}^{p} b_k \, \overline{X}_k(j) + \sum_{k=1}^{p} a_k \, \psi^{c-1}{}_{X_k(j)}(t) + e_j(t)$$

$$with \quad a_k \geq 0, \quad b_k \in IR, \quad k \in \{1, \ldots, p\}.$$

- The method relies on the exploitation of the properties of a decomposition of the Mallows (Wasserstein) distance;
- The parameters are obtained by minimizing the Mallows's distance between the observed and the derived quantile functions of the dependent variable;

Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Models (DSD)

Dias and Brito, 2015. *Linear Regression Model with histogram-valued variables*. Stat Anal Data Min 8(2), 75-113.
Dias, 2014. *Linear regression with empirical distributions*. Ph.D thesis. Universidade do Porto, Portugal.

$\psi^{-1}_{X_1(j)}(t), \psi^{-1}_{X_2(j)}(t), \ldots, \psi^{-1}_{X_p(j)}(t), \psi^{-1}_{Y(j)}(t)$: quantile functions representing the distributions of the explicative and response histogram-valued variables for observation *j*.

Linear regression relation:

DSD I

$$\psi^{-1}{}_{Y(j)}(t) = v + a_1\, \psi^{-1}{}_{X_1(j)}(t) - b_1\, \psi^{-1}{}_{X_1(j)}(1-t) \ldots + a_p\, \psi^{-1}{}_{X_p(j)}(t) - b_p \psi^{-1}{}_{X_p(j)}(1-t) + e_j(t)$$

$$with \quad a_k, b_k \geq 0, \quad k \in \{1, \ldots, p\}, \quad v \in IR.$$

DSD II

$$\psi^{-1}{}_{\hat{Y}(j)}(t) = \psi^{-1}_{Constant}(t) + a_1\, \psi^{-1}{}_{X_1(j)}(t) - b_1\, \psi^{-1}{}_{X_1(j)}(1-t) \ldots + a_p\, \psi^{-1}{}_{X_p(j)}(t) - b_p \psi^{-1}{}_{X_p(j)}(1-t) + e_j(t)$$

$$with \quad a_k, b_k \geq 0, \quad k \in \{1, \ldots, p\}$$

The **error function** is a piecewise linear function (but not necessarily a quantile function)

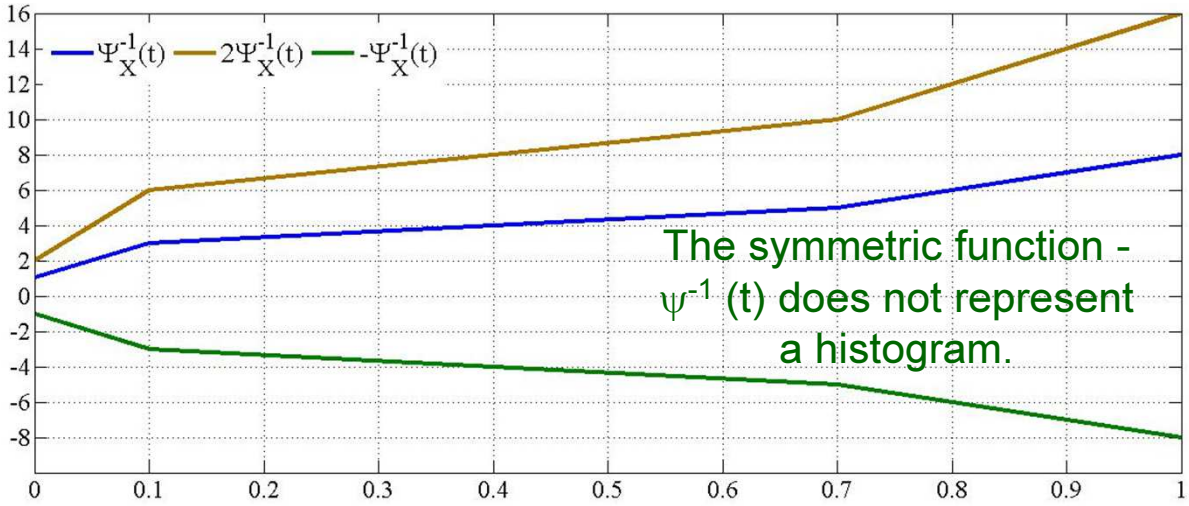
Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Models (DSD)

Why?

- Is it necessary to impose constraints to the parameters of the model?
- Is it necessary to include in the model the quantile functions $-\Psi^{-1}_{X_k(j)}(1-t)$?

What does it represent?

- If one β_k is negative, $\beta_k \psi^{-1}_{X_k(j)}(t)$ is not a **non-decreasing function**;
- Consequently, **it is not possible** to obtain the symmetric distribution multiplying the quantile function by -1.



The product of a quantile function by a real number λ:

- is a **non-decreasing function** if $\lambda \geq 0$.
- is **NOT A NON-DECREASING FUNCTION** if $\lambda < 0$.

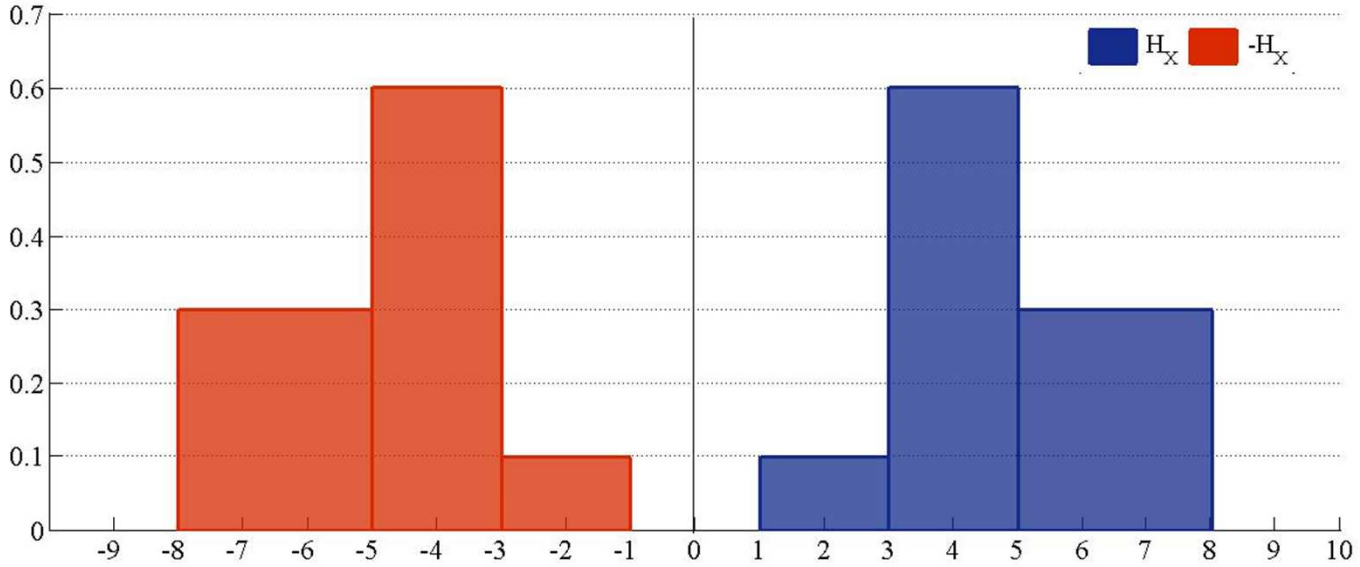
Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Models (DSD)

Product of an interval by a real number – interval arithmetic (Moore, 1966)

$$\alpha\left[\underline{I}_X, \overline{I}_X\right] = \begin{cases} \left[\alpha\underline{I}_X, \alpha\overline{I}_X\right] & if \quad \alpha \geq 0 \\ \left[\alpha\overline{I}_X, \alpha\underline{I}_X\right] & if \quad \alpha < 0 \end{cases}$$

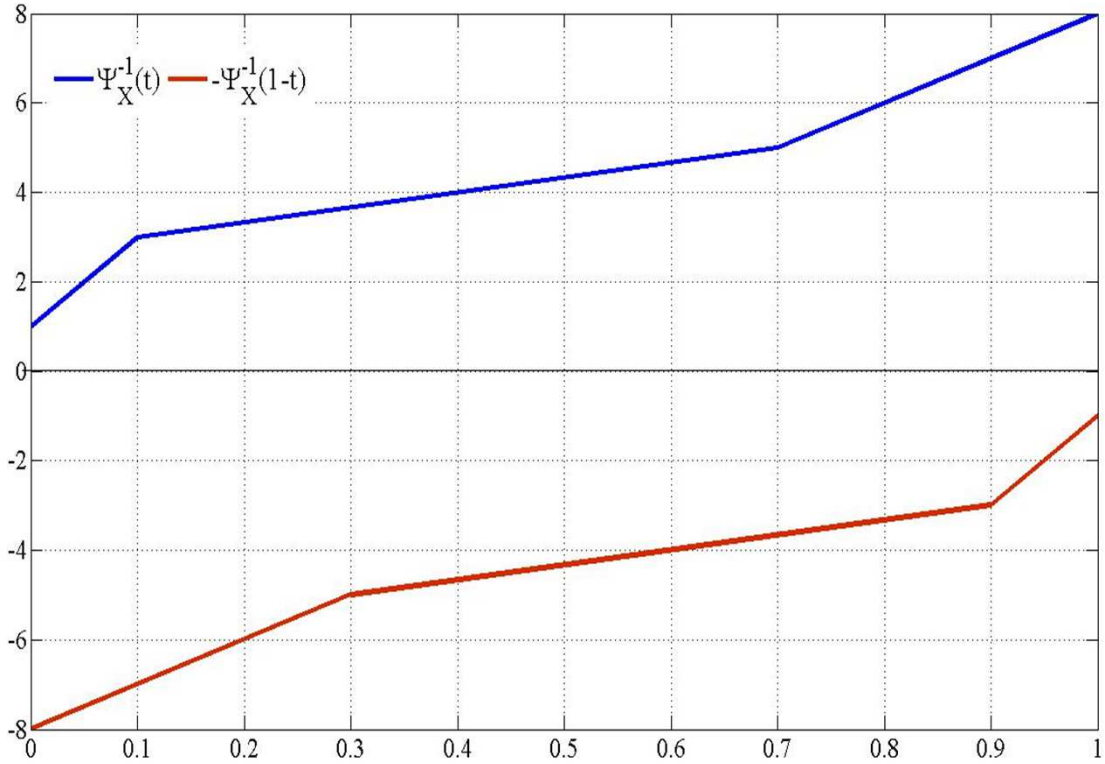
$$-H_X = \{[-8,-5[, 0.3; [-5,-3[, 0.6; [-3,-1], 0.1\} \qquad H_X = \{[1,3[, 0.1; [3,5[, 0.6; [5,8], 0.3\}$$

SYMMETRIC HISTOGRAM



HISTOGRAM

Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Models (DSD)



$$\Psi_X^{-1}(t) = \begin{cases} 1 + \dfrac{t}{0.1} \times 2 & 0 \le t < 0.1 \\ 3 + \dfrac{t - 0.1}{0.6} \times 2 & 0.1 \le t < 0.7 \\ 5 + \dfrac{t - 0.7}{0.3} \times 3 & 0.7 \le t \le 1 \end{cases}$$

$$\Psi_X^{-1}(1-t) = \begin{cases} -8 + \dfrac{t}{0.3} \times 3 & 0 \le t < 0.3 \\ -5 + \dfrac{t - 0.3}{0.4} \times 2 & 0.3 \le t < 0.7 \\ -3 + \dfrac{t - 0.7}{0.3} \times 2 & 0.7 \le t \le 1 \end{cases}$$

Quantile function representing the symmetric histogram $-\Psi^{-1}(1-t)$: obtained from the quantile function $\Psi^{-1}(t)$ with $t \in [0,1]$.

Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Models (DSD)

- The **quantile function that represents the symmetric histogram** is obtained from the quantile function $\Psi^{-1}(t)$ and is the function $-\Psi^{-1}(1-t)$ with $t \in [0,1]$;
- $-\Psi^{-1}(1-t)$ is a non-decreasing function;
- $\Psi^{-1}(t)+\left(-\Psi^{-1}(1-t)\right)$ is not a null function but is a quantile function with mean zero;
- $-\Psi^{-1}(1-t)=\Psi^{-1}(t)$ only when the histograms are symmetric;
- Applying the Irpino and Verde process to the functions $\Psi^{-1}(t)$ and $-\Psi^{-1}(1-t)$ the distributions are defined with and equal number of subintervals each of which have associated weights p_i that verify the condition $p_i = p_{n-i+1}$, with $i \in \{1,2,\ldots,n\}$.

$$\Psi_X^{-1}(t)=\begin{cases} 1+\dfrac{t}{0.1}\times 2 & 0\le t<0.1 \\ 3+\dfrac{t-0.1}{0.2}\times\dfrac{2}{3} & 0.1\le t<0.3 \\ \dfrac{11}{3}+\dfrac{t-0.3}{0.4}\times\dfrac{4}{3} & 0.3\le t<0.7 \\ 5+\dfrac{t-0.7}{0.2}\times 2 & 0.7\le t<0.9 \\ 7+\dfrac{t-0.9}{0.1} & 0.9\le t\le 1 \end{cases} \qquad -\Psi_X^{-1}(1-t)=\begin{cases} -8+\dfrac{t}{0.1} & 0\le t<0.1 \\ -7+\dfrac{t-0.1}{0.2}\times 2 & 0.1\le t<0.3 \\ -5+\dfrac{t-0.3}{0.4}\times\dfrac{4}{3} & 0.3\le t<0.7 \\ -\dfrac{11}{3}+\dfrac{t-0.7}{0.2}\times\dfrac{2}{3} & 0.7\le t<0.9 \\ -3+\dfrac{t-0.9}{0.1}\times 2 & 0.9\le t\le 1 \end{cases}$$

Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Model (DSD)

For classical variables

$$\hat{y}_j = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_p x_{pj}$$

$$Minimize \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \left(y_j - \hat{y}_j \right)^2$$

Least squares method

For histogram-valued variables

PREDICTION

$$\psi^{-1}_{\hat{Y}(j)}(t) = \psi^{-1}_{Constant}(t) + a_1 \psi^{-1}_{X_1(j)}(t) - b_1 \psi^{-1}_{X_1(j)}(1-t) + \ldots\ldots + a_p \psi^{-1}_{X_p(j)}(t) - b_p \psi^{-1}_{\tilde{X}_{p(j)}}(1-t)$$

$$with \quad a_k, b_k \geq 0, \quad k \in \{1, \ldots, p\}$$

ERROR MEASURE

DSD Model I

$$Minimize \quad SE = \sum_{j=1}^{n} \int_0^1 \left(\psi^{-1}_{Y(j)}(t) - \psi^{-1}_{\hat{Y}(j)}(t) \right)^2 dt$$

$$with \quad a_k, b_k \geq 0, \quad k \in \{1, \ldots, p\}$$

DSD Model II

$$Minimize \quad SE = \sum_{j=1}^{n} \int_0^1 \left(\psi^{-1}_{Y(j)}(t) - \psi^{-1}_{\hat{Y}(j)}(t) \right)^2 dt$$

$$with \quad a_k, b_k \geq 0, k \in \{1, \ldots, p\};$$
$$r_{v_i} \geq 0, i \in \{1, \ldots, n\}$$

Mallows distance is applied

Linear Regression for histogram-valued variables

Distribution and Symmetric Distribution Models (DSD)

For classical variables

$$\sum_{j=1}^{n}\left(y_j-\bar{y}\right)^2=\sum_{j=1}^{n}\left(\hat{y}_j-\bar{y}\right)^2+\sum_{j=1}^{n}\left(y_j-\hat{y}_j\right)^2$$

$$R^2=\frac{\sum_{j=1}^{m}\left(\hat{y}_j-\bar{y}\right)^2}{\sum_{j=1}^{m}\left(y_j-\bar{y}\right)^2}$$

COEFFICIENT OF DETERMINATION

$\bar{y}$ is the classical mean of the variable y

For histogram-valued variables

From the decomposition

$$\sum_{j=1}^{m}D^2{}_M\left(\psi^{-1}{}_{Y(j)}(t),\bar{Y}\right)=$$

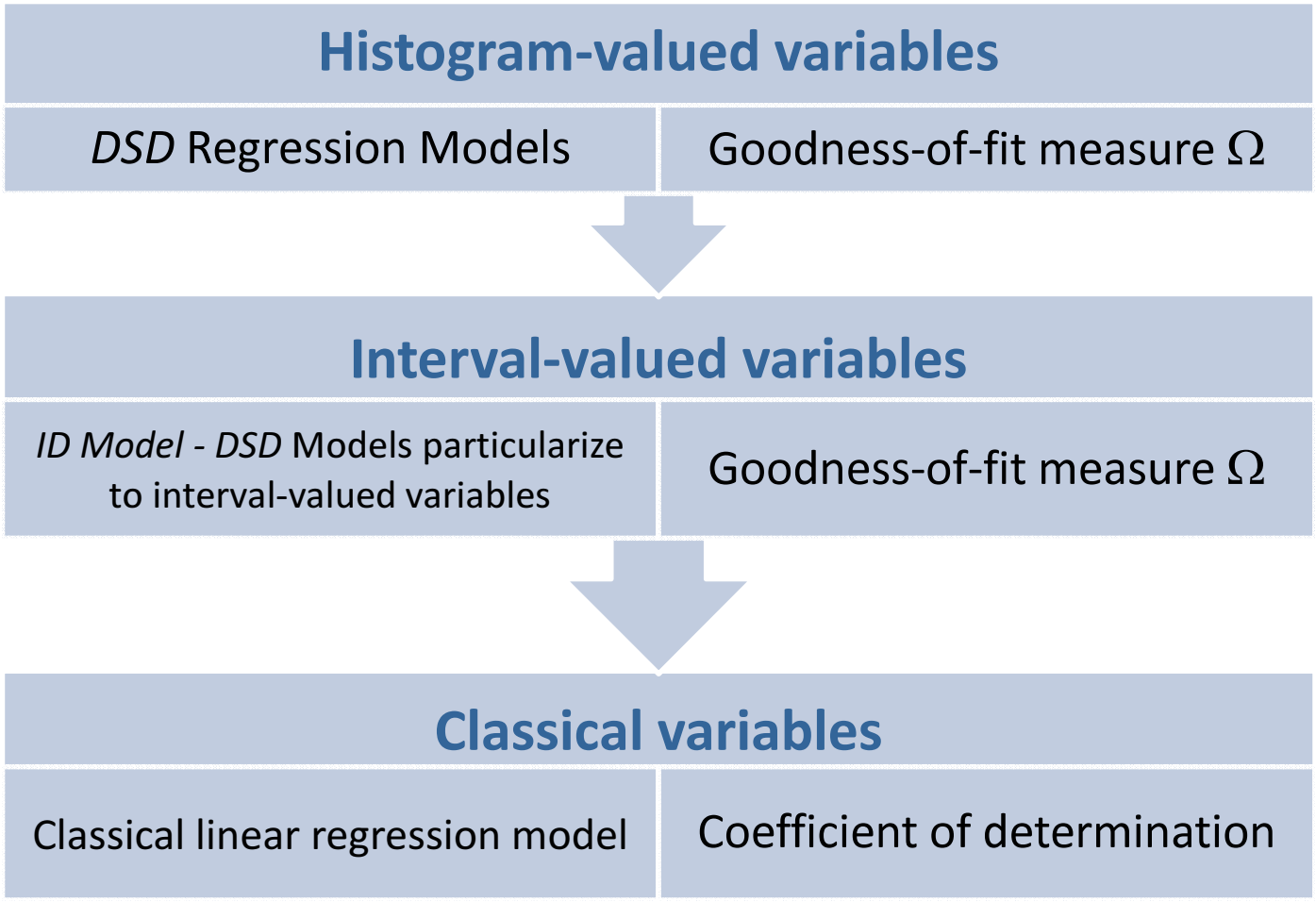
$$=\sum_{j=1}^{m}D^2{}_M\left(\psi^{-1}{}_{\hat{Y}(j)}(t),\bar{Y}\right)+\sum_{j=1}^{m}D^2{}_M\left(\psi^{-1}{}_{Y(j)}(t),\psi^{-1}{}_{\hat{Y}(j)}(t)\right)$$

the goodness-of-fit measure of the model is given by

$$\Omega=\frac{\sum_{j=1}^{m}D^2{}_M\left(\psi^{-1}{}_{\hat{Y}(j)}(t),\bar{Y}\right)}{\sum_{j=1}^{m}D^2{}_M\left(\psi^{-1}{}_{Y(j)}(t),\bar{Y}\right)}$$

$\bar{Y}$ is the symbolic mean of the histogram-valued variable Y

Linear Regression for histogram-valued variables
Distribution and Symmetric Distribution Models (DSD)



Linear Regression for histogram-valued variables

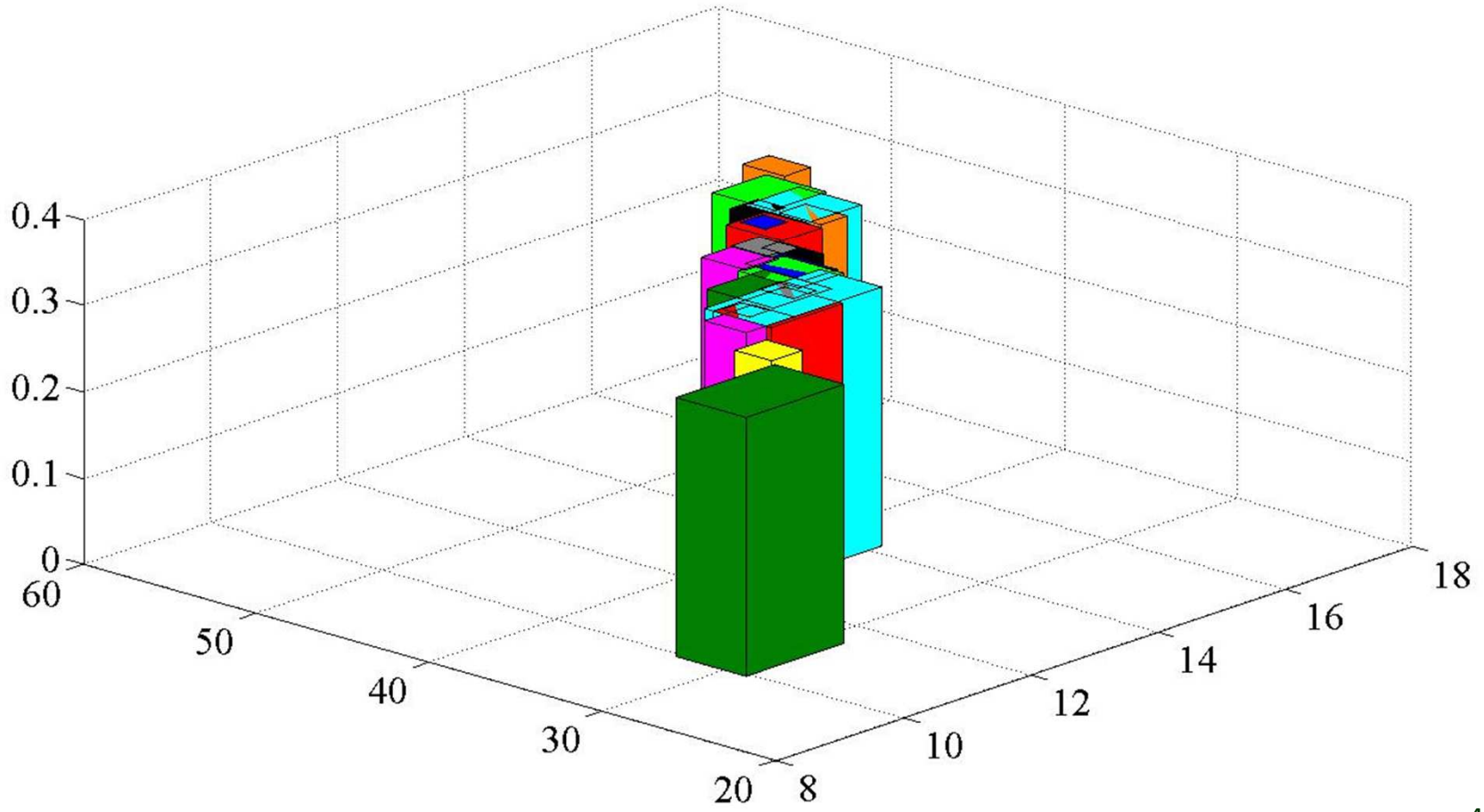
Example: Histogram data

Observations	Y	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}

Linear Regression for histogram-valued variables

Example: Histogram data

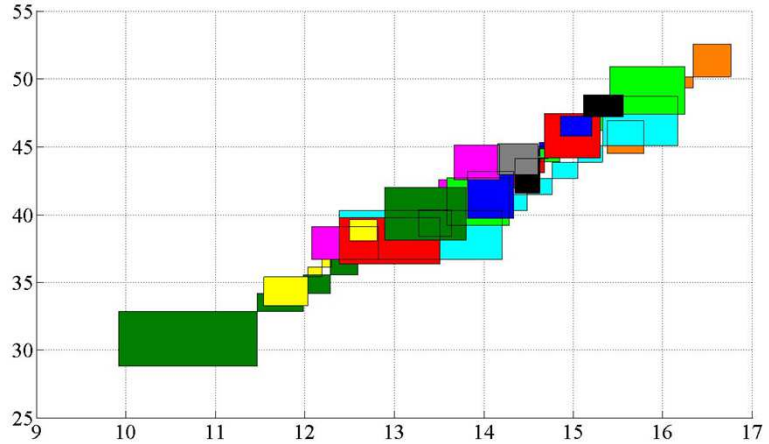
Scatter plot of the data



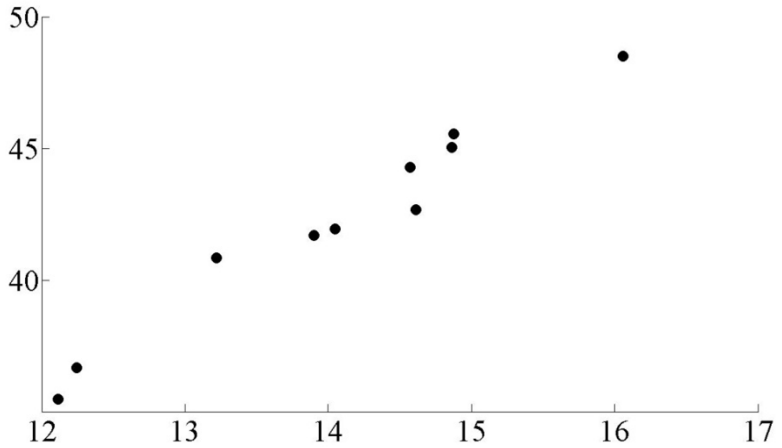
Linear Regression for histogram-valued variables

Example: Histogram data

VARIABLES X AND Y

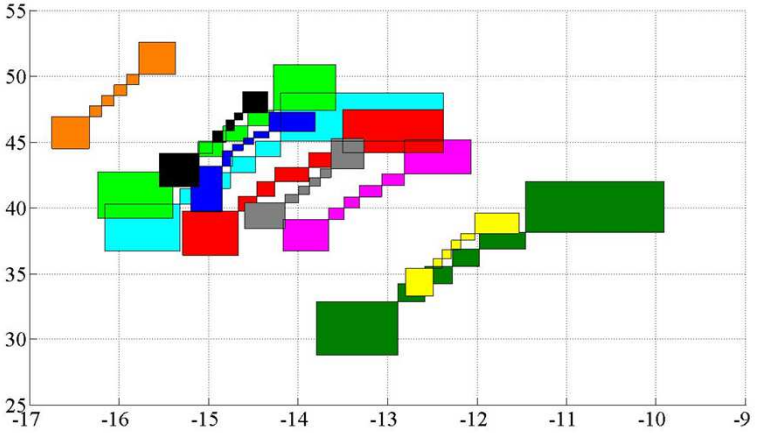


$$\psi^{-1}_{\hat{Y}(j)}(t) = -1.95 + 3.56\,\psi^{-1}{}_{X(j)}(t) - 0.41\,\psi^{-1}{}_{X(j)}(1-t)$$

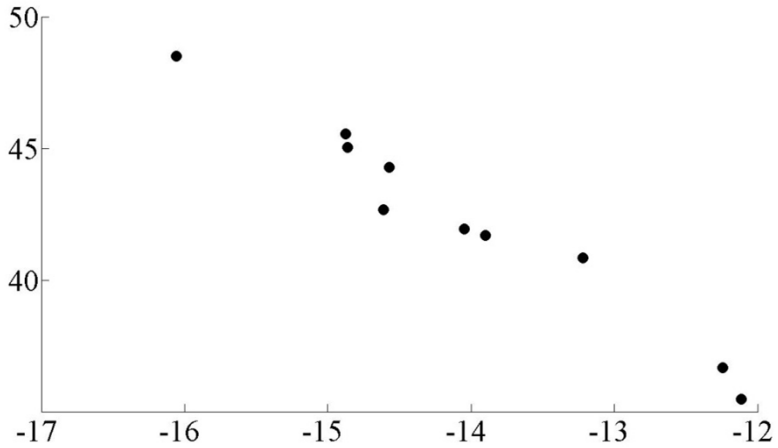


$$\overline{Y}(j) = -1.95 + (3.56 - 0.41)\,\overline{X}(j)$$

VARIABLES -X AND Y



$$\psi^{-1}_{\hat{Y}(j)}(t) = -1.95 + 0.41\,\psi^{-1}{}_{X(j)}(t) - 3.56\,\psi^{-1}{}_{X(j)}(1-t)$$

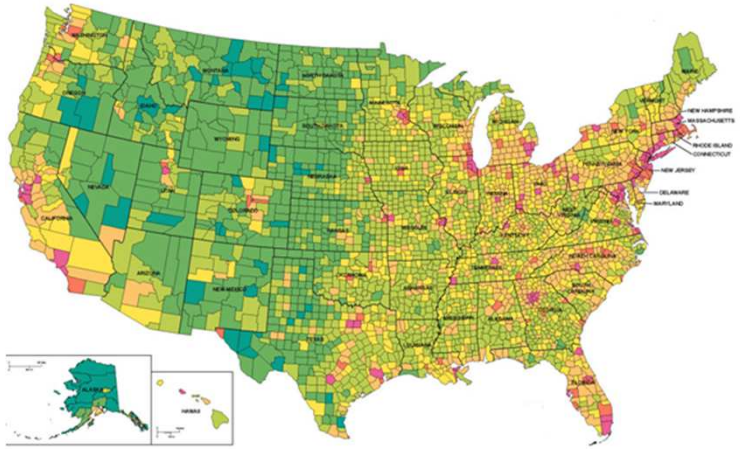


$$\overline{Y}(j) = -1.95 + (0.41 - 3.56)\,\overline{X}(j)$$

Linear Regression for histogram-valued variables
Example: Crimes in USA

Original data:

- Socio-economic data from the '90 Census
- Crime data from 1995



First level units: Cities of the USA states

Observations associated to each unit:

- The records (real values) associated to the cities in USA states.
- Only consider the states for which the number of records for all selected variables was higher than thirty – twenty states were considered.

Original variables

Response variable :

VC - violent crimes *(total number of violent crimes per 100 000 habitants)*

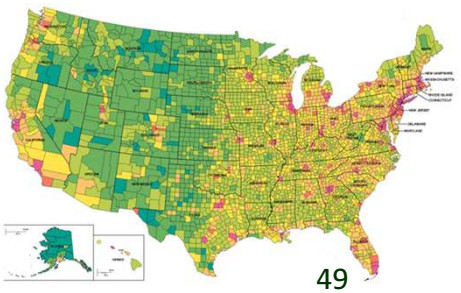
Four explicative variables:

LEd - percentage of people aged 25 and over with less than 9th grade education;
Emp - percentage of people aged 16 and over who are employed;
Div - percentage of population who are divorced;
Img - percentage of immigrants who immigrated within the last 10 years.

Linear Regression for histogram-valued variables

Example: Crimes in USA

Cities	States	LEd	Emp	Div	Img	VC
Selma	AL	16.59	46.94	13.35	73.86	2758.9
Bessemer	AL	16.97	46.83	14.46	18.39	1257.09
Dothan	AL	11.71	62.19	13.75	34.25	373.54
...	...	...	...	...	...	...
San Pablo	CA	14.03	55.94	16.57	62.3	374.07
Glendale	CA	11.54	60.04	11.12	60.4	644.75
...	...	...	...	...	...	...
Enfield	CT	6.55	68.24	8.38	27.01	78.65
Newingtont	CT	8.71	67.54	8.57	18.44	2127.02
New Haven	CT	11.86	56.71	12.44	46.52	53.2
...	...	...	...	...	...	...
Rockledge	FL	4.07	64.92	11.99	23.72	142.7
Ormond Beach	FL	4.72	51.11	10.31	13.69	339.96
Sebastian	FL	7.04	48.43	9.08	17.35	1981.45
...	...	...	...	...	...	...
Alpharetta	GA	2.21	75.34	12.84	53	958.15
Valdosta	GA	11.35	59.19	12.97	37.64	1358.47
...	...	...	...	...	...	...



Linear Regression for histogram-valued variables

Example: Crimes in USA

Contemporary aggregation for state.

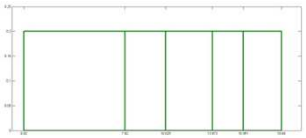
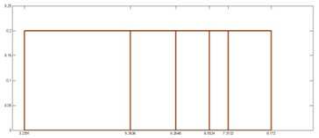
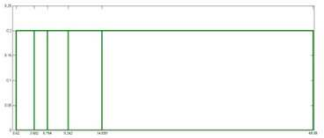
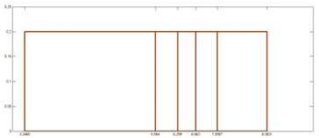
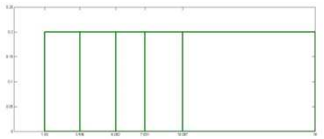
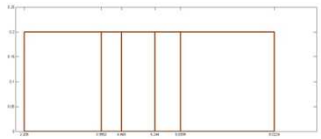
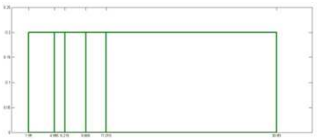
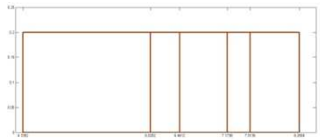
Higher level units: USA states

Observations associated to each unit:

•The distributions of the records of the cities of the respective state;

•In all observations, the subintervals of each histogram have the same weight (equiprobable) with frequency 0.20.

Goal: Study the criminality in the USA states

States	LEd	…	VC
AL		…	
CA		…	
CT		…	
FL		…	
…	…	…	…

Linear Regression for histogram-valued variables

Example: Crimes in USA

DSD Model I

$$\psi^{-1}_{\hat{LVC}(j)}(t) = 3.9321 + 0.0009\,\psi^{-1}_{LEd(j)}(t) - 0.0123\psi^{-1}_{Emp(j)}(1-t) + 0.2073\,\psi^{-1}_{Div(j)}(t) - 0.0353\psi^{-1}_{Div(j)}(1-t) + 0.0187\psi^{-1}_{Img(j)}(t), \quad t \in [0,1]$$

$$\Omega = 0.8680$$

DSD Model II

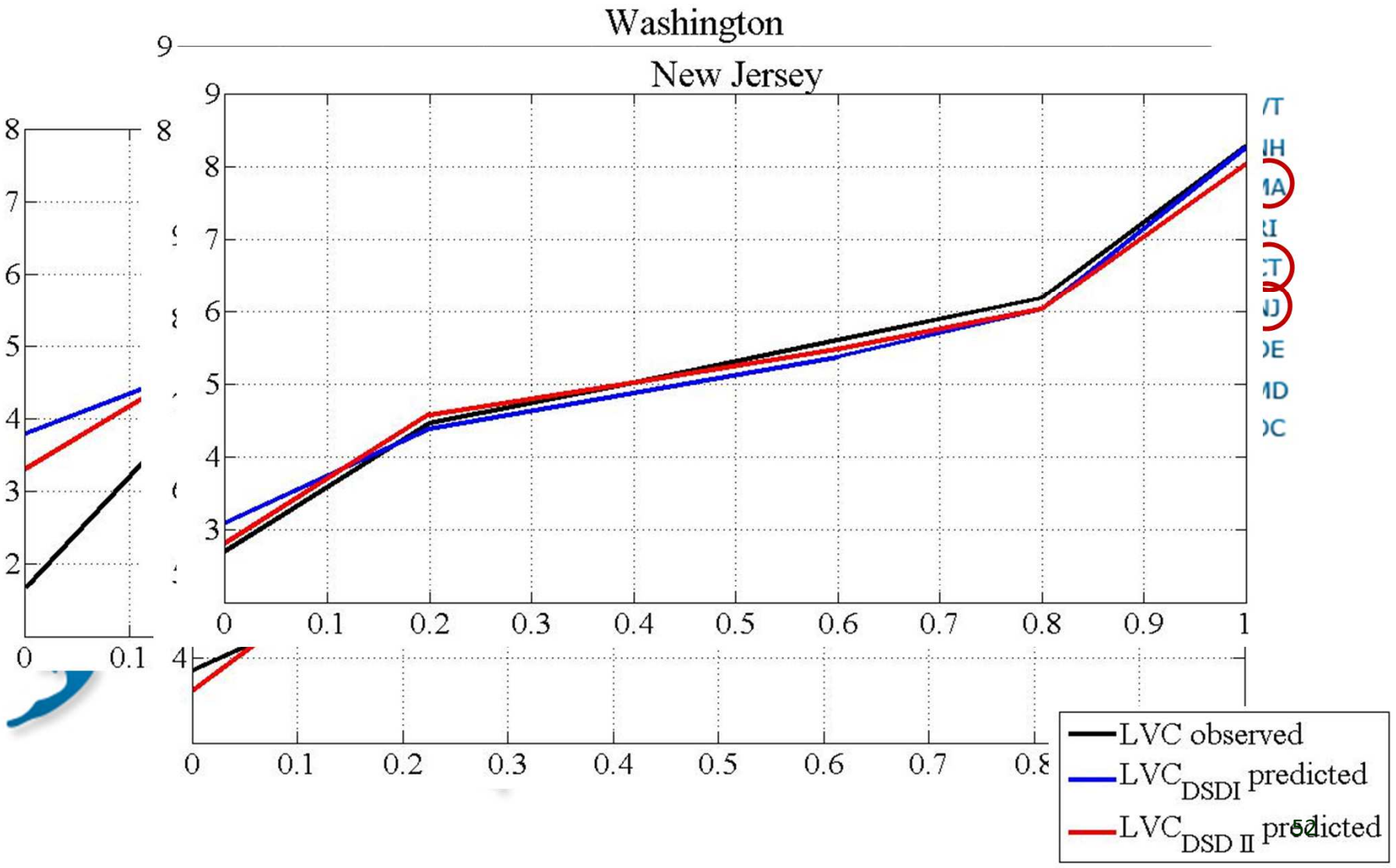
$$\psi^{-1}_{\hat{LVC}(j)}(t) = \psi^{-1}_{Constant}(t) + 0.016\,\psi^{-1}_{LEd(j)}(t) - 0.009\psi^{-1}_{Emp(j)}(1-t) + 0.155\,\psi^{-1}_{Div(j)}(t) + 0.019\psi^{-1}_{Img(j)}(t),$$

$$\psi^{-1}_{Constant}(t) = \begin{cases} 3.37 + \left(\frac{2t}{0.2} - 1\right) \times 0.42 & 0 \le t < 0.2 \\ 3.83 + \left(\frac{2(t-0.2)}{0.2} - 1\right) \times 0.04 & 0.2 \le t < 0.4 \\ 3.9 + \left(\frac{2(t-0.4)}{0.2} - 1\right) \times 0.03 & 0.4 \le t < 0.6 \\ 3.93 & 0.6 \le t \le 1 \end{cases}$$

$$\Omega = 0.8818$$

Linear Regression for histogram-valued variables

Example: Crimes in USA



Linear Regression for histogram-valued variables
Example: Crimes in USA

Predicted distribution of LVC:

$$\psi^{-1}_{\hat{LVC}(j)}(t) = 3.9321 + 0.0009\,\psi^{-1}_{LEd(j)}(t) - 0.0123\psi^{-1}_{Emp(j)}(1-t) + 0.2073\,\psi^{-1}_{Div(j)}(t) -$$

$$-0.0353\psi^{-1}_{Div(j)}(1-t) + 0.0187\psi^{-1}_{Img(j)}(t), \qquad t \in [0,1]$$

Interpretation:

- The variables *LEd*, *Div* and *Img* have a direct influence in the logarithm of the number of violent crimes. The percentage of employed people, *Emp*, has an opposite effect.

Property

For each unit *j*, let $\hat{Y}(j)$ be the distribution predicted by the *DSD Model I* and consider the parameters obtained for the optimal solution $b^*=(a_1^*, b_1^*, a_2^*, b_2^*, \ldots, a_n^*, b_n^*, v^*)$. The mean of the predicted histogram-valued variable $\overline{\hat{Y}}$ is given by $\overline{\hat{Y}} = \sum_{k=1}^{p}(a_k^* - b_k^*)\overline{X}_k + v^*$

- For the set of states to which the data refer, when the symbolic mean of the percentage of divorced population increases 1% and the other variables remain constant, the symbolic mean of the LVC increases 0.172.

Linear Regression for histogram-valued variables

Example: Crimes in USA

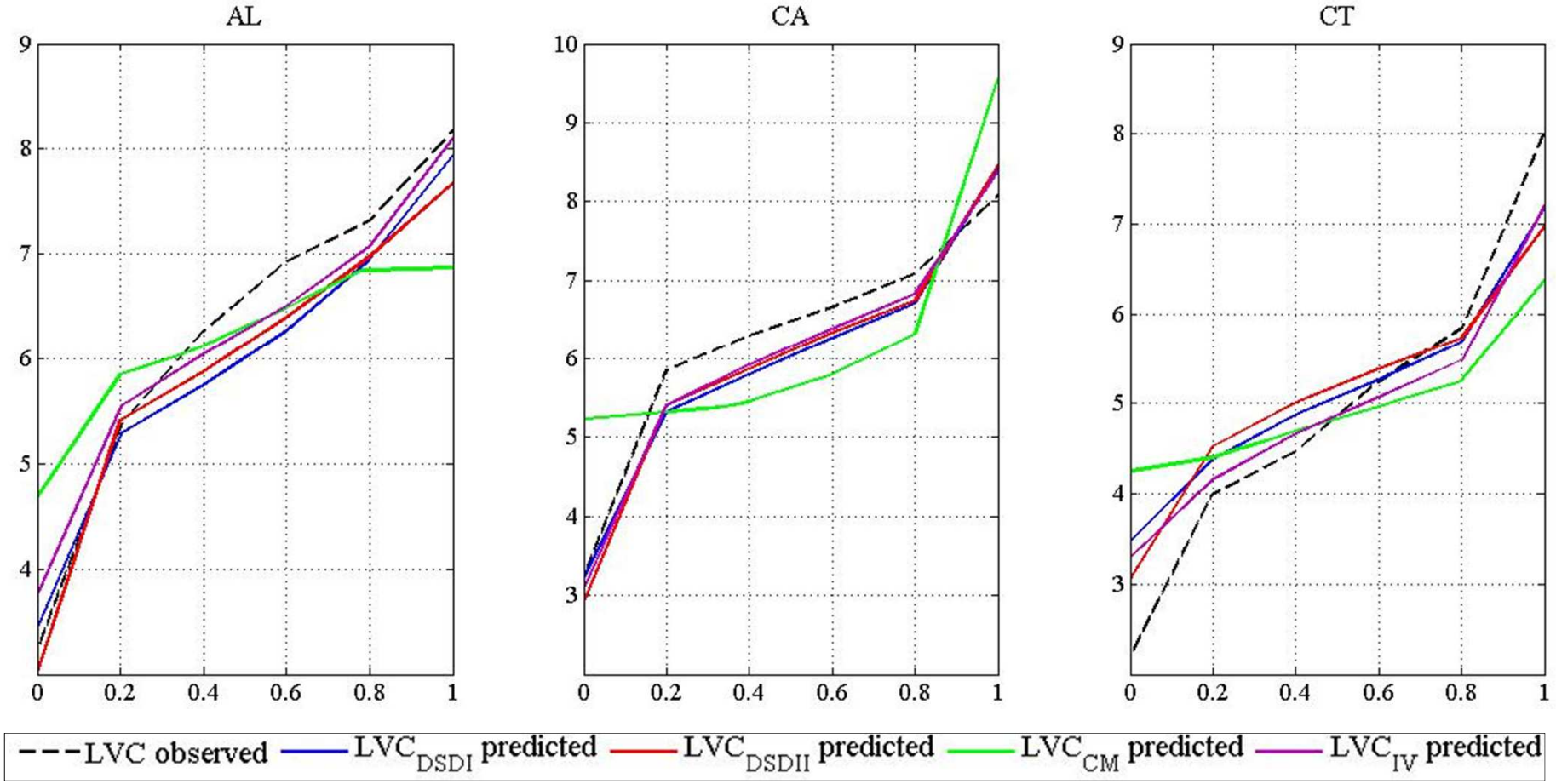
Evaluate the performance of the models

Models	$RMSE_L$	$RMSE_U$	$RMSE_M$
CM	0.9182	0.5617	0.6717
IV	0.5214	0.3444	0.3933
DSD I	0.5571	0.4233	0.4477
DSD II	0.5164	0.3992	0.4237

Linear Regression for histogram-valued variables

Example: Crimes in USA

Observed and predicted quantile functions considering different methods



References

- Billard and Diday (2000). *Regression analysis for interval-valued data*. In: Proc. of IFCS'00, Belgium, pp. 369-374, Springer.
- Billard, L., Diday, E. (2002). Symbolic Regression Analysis. In: *Proc.* IFCS'02, Poland, pp. 281-288, Springer.
- Billard, L., Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.
- De Carvalho, F.A.T., Neto, E.A.L. (2008). Centre and Range method for fitting a linear regression model to symbolic intervalar data. *Computational Statistics & Data Analysis*, 52 , pp. 1500-1515.
- De Carvalho, F.A.T., Neto, E.A.L., (2010). Centre and Range method for fitting a linear regression model to symbolic intervalar data. *Computational Statistics & Data Analysis*, 54 , pp. 333-347.
- Dias, 2014. *Linear regression with empirical distributions*. Ph.D thesis. Universidade do Porto, Portugal.
- Dias and Brito, 2015. *Linear Regression Model with histogram-valued variables*. Stat Anal Data Min 8(2), pp. 75-113.
- Lima Neto, Cordeiro, De Carvalho, 2011. Bivariate symbolic regression models for interval-valued variables. Journal of Statistical Computation and Simulation 81 (11), pp.1727-1744.
- Irpino, A. Verde, R., (2006). A New Wasserstein Based distance for the hierarchical clustering of histogram symbolic data. In: *Proc.* IFCS'06, Berlin, pp. 185-192, Springer, .
- Irpino and Verde, 2015. *Linear regression for numeric symbolic variables: an ordinary least squares approach based on Wasserstein Distance . Adv Data Anal Classif 9(1), pp. 81-106.*