Histogram data analysis based on Wasserstein distance

Rosanna Verde – Antonio Irpino

Department of European and Mediterranean Studies
Second University of Naples

Caserta - ITALY

Aims

Introduce:

- New distances to compare histogram (distribution) data in the framework of the Symbolic Data Analysis – especially the Wasserstein – Mallows' ℓ_2 distance

Show:

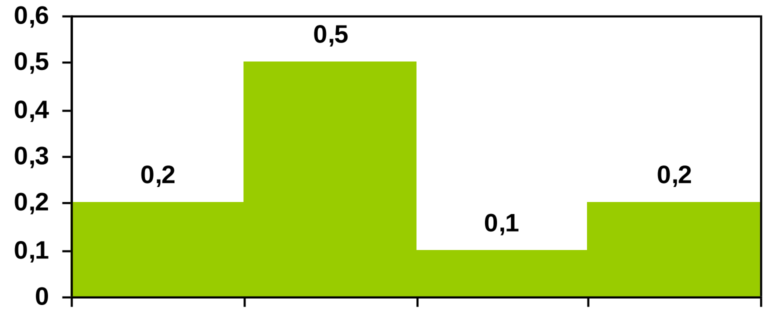
- Some properties of Wasserstein – Mallows' ℓ_2 distance

Present:

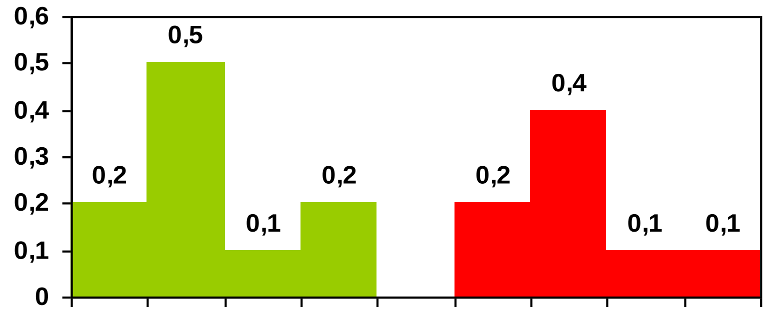
- Methodological approaches based on ℓ_2 distance
  - Basic Statistics for histogram data (including intervals and distributions as particular cases)
  - Clustering methods: DCA and hierarchical (Ward criterion)
  - Linear regression model: ℓ_2 as metric for Sum Square Errors in OLS estimation method

Main Sources of histogram data

- Results of summary/clustering procedures
  - From surveys
  - From large databases
  - From sensors
    - Temperatures
    - Pollutant concentration
    - Network activity



- Data streams
  - Description of time window data sequences
- Image analysis
  - Color bandwidths
- Confidentiality data
  - Summary data – non punctual



Histogram data as a particular case of modal symbolic descriptions [*Bock and Diday (2000)*]

- Histogram data is a model for representing the empirical distribution of a continuous variable Y partitioned into a set of contiguous I_h intervals (bins) with associated π_h weights.

  A histogram is represented by a set of H ordered pairs $(I_h,\pi_h)\}$ such that:

$$I_{hi} \equiv \left[\underline{y}_h ; \overline{y}_h\right] \quad ; \quad \underline{y}_h \leq \overline{y}_h \quad ; \quad \underline{y}_h, \overline{y}_h \in \Re$$

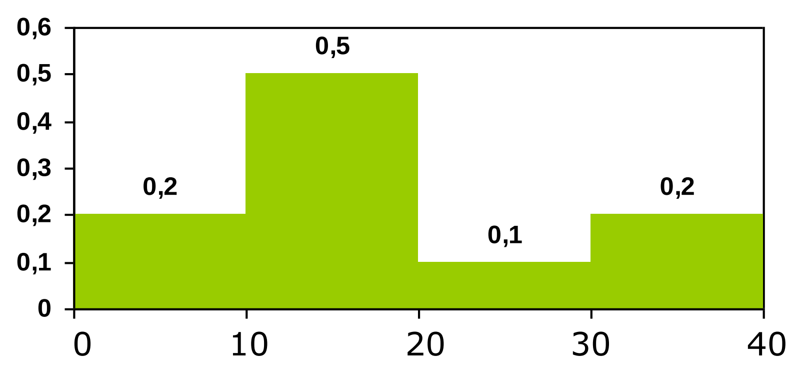
$$\bigcup_{h=1,\dots,H} I_{hi} = \left[\min_{h=1,\dots,H}\left\{\underline{y}_h\right\} ; \max_{h=1,\dots,H}\left\{\overline{y}_h\right\}\right]$$

$$h \neq h' \quad I_h \bigcap I_{h'} = \varnothing$$

$$\pi_h \geq 0$$

$$\sum_{h=1,\dots,H} \pi_h = 1$$

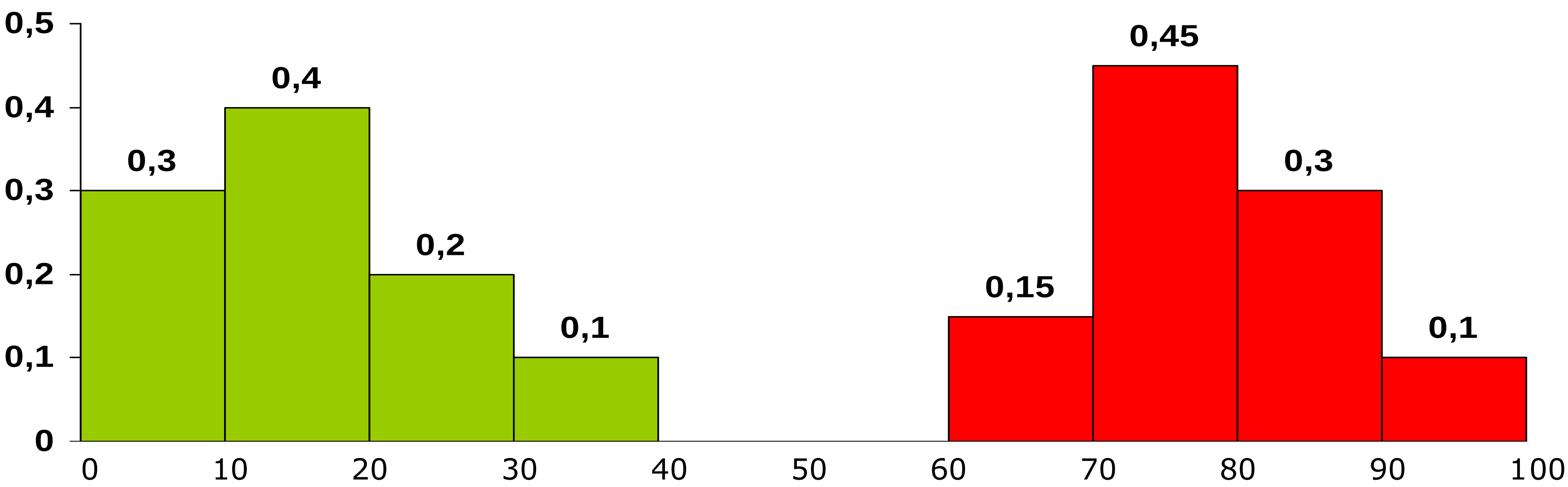
Y(i)=[([0-10],*0.2*);([10-20],*0.5*);([20-30],*0.1*);
([30-40],*0.2*)]



Comparison of two histogram data

- How do compare two units described by two histograms?

- A possibility is to use metrics developed for comparing probability distributions



Metrics used for the evaluation of the convergence of two probability measures (Gibbs and Su, 2002)

Given:

- a domain Ω on which is possible to define a Borel σ-algebra,
- two measure μ, ν
- the density functions f and g
- the corresponding distribution functions **F** and **G** and
- a subdominant measure λ, like: $\lambda=(\mu+\nu)/2$,

Gibbs and Su (2002) present a review of the most used dissimilarities:

Abbreviation	Metric
D	Discrepancy
H	Hellinger distance
I	Relative entropy (or Kullback-Leibler divergence)
K	Kolmogorov (or Uniform) metric
L	Lévy metric
P	Prokhorov metric
S	Separation distance
TV	Total variation distance
W	Wasserstein (or Kantorovich) metric
χ^2	χ^2 distance

A suitable measure to compute the distance between histograms: Wasserstein-Kantorovich metric

- we propose to use the Wasserstein-Kantorovich metric:

$$d_W(\mu,\upsilon)=\sqrt{\int_0^1 \left(F^{-1}(t)-G^{-1}(t)\right)^2 dt}$$

- in particular the derived ℓ_2 Mallow's distance between two quantile functions

$$d_W^2(\mu,\upsilon)=\int_0^1 \left(F^{-1}(t)-G^{-1}(t)\right)^2 dt$$

- The main difficulties to compute this distance is the analytical definition of the *quantile function*…

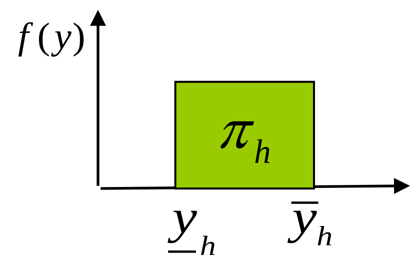
But in our case we treat especially with histogram data), indeed…

Histograms are locally uniform

Having represented an histogram Y(*i*) as $\{(I_1,\pi_1),\ldots,(I_h,\pi_h),\ \ldots,\ (I_m,\pi_H)\}$ (where *H* is the number of intervals of the support), we may define:

- the empirical density function *f*(*y*) as:

$$f_h = \begin{cases} \dfrac{\pi_h}{\underline{y}_h - \bar{y}_h} & \text{if } y \in [\underline{y}_h ; \bar{y}_h) \\ 0 & \text{otherwise} \end{cases}$$



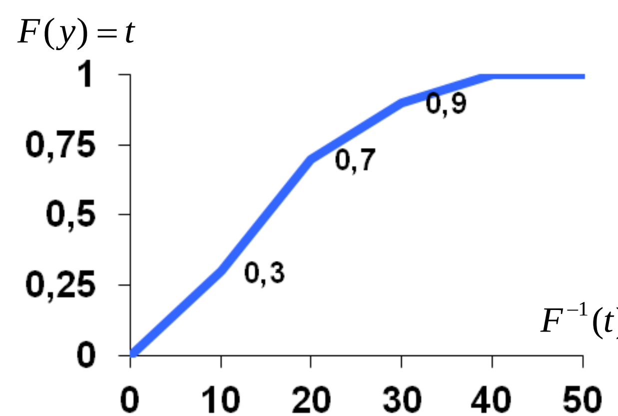
- the empirical *distribution function*

$$w_h = \begin{cases} 0 & \text{if } h = 0 \\ \sum_{k=1}^{h} \pi_k & \text{if } h = 1,\ldots,m \end{cases}$$

$$F(y) = w_h + \left(y - \underline{y}_h\right)\frac{w_h - w_{h-1}}{\bar{y}_h - \underline{y}_h} \quad \textit{iff} \quad \underline{y}_h \le y < \bar{y}_h$$

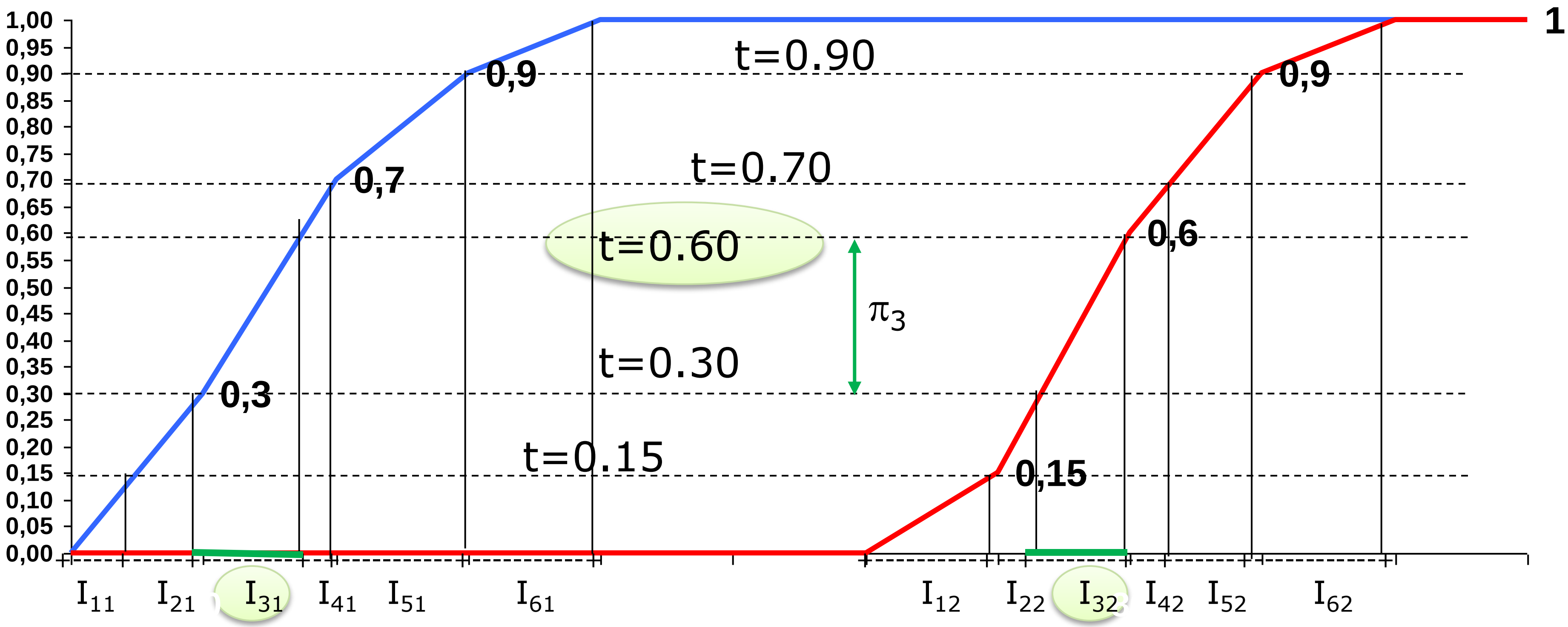
- hence the empirical *quantile function* (The inverse of the distribution function)

$$F^{-1}(t) = \underline{y}_h + \frac{t - w_{h-1}}{w_h - w_{h-1}}\left(\bar{y}_h - \underline{y}_h\right) \quad \text{where } 0 \le w_h \le t \le w_{h-1} \le 1$$



Geometric interpretation

The comparison of two quantile functions associated with two histograms requires the partition of the support of the two qf's into *m* intervals with associated uniform density



l	π_l	$I_{li}=\left[\underline{y}_{li},\overline{y}_{li}\right]$	$I_{lj}=\left[\underline{y}_{lj},\overline{y}_{lj}\right]$	$d^2_W(I_{il};I_{jl})$	$\pi_l\, d^2_W(I_{il};I_{jl})$
1	0.15	[0; 5]	[60,70]	3908.33	586.25
2	0.15	[5;10]	[70;73.3]	4115.46	617.32
3	0.30	[10; 16.6]	[73.3;80]	4013.22	1203.97
4	0.10	[16.6;20]	[80;83.3]	4013.22	401.32
5	0.20	[20;30]	[83.3;90]	3801.63	760.33
6	0.10	[30;40]	[90;100]	3600.00	360.00
					3929.18

$$d_W^2(Y_i,Y_j)=\sum_{l=1}^{m}\pi_l d^2(I_{li},I_{lj})=\sum_{l=1}^{m}\pi_l\left[(c_{li}-c_{lj})^2+\frac{1}{3}(r_{li}-r_{lj})^2\right]$$

With $I_{li}=\left[\underline{y}_{li},\overline{y}_{li}\right]$; $c_{li}=\dfrac{\underline{y}_{li}+\overline{y}_{li}}{2}$; $r_{li}=\dfrac{\overline{y}_{li}-\underline{y}_{li}}{2}$

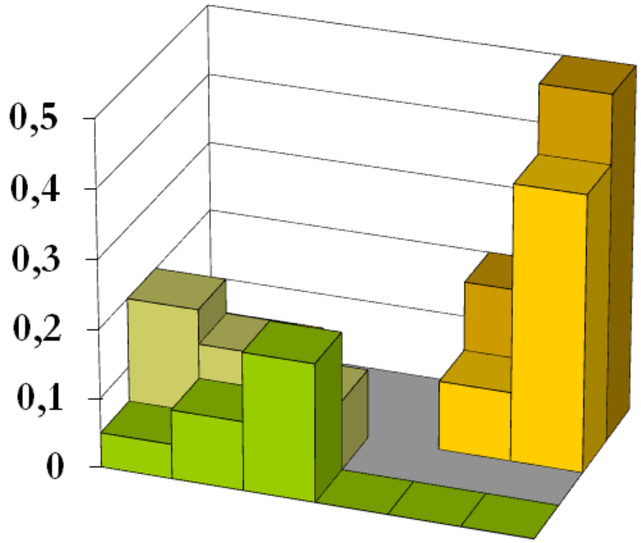
center and radius of I_{li}

Verde, Irpino, COMPSTAT2006

Multivariate distance

- Assuming independence among *p variables* we propose the an extension of $d^2{}_W$ distance to the multivariate case

$$d_W^2(Y(i),Y(j)) := \sum_{k=1}^{p}\sum_{l=1}^{m_k}\pi_l^{(k)}\left[\left(c_{li}^{(k)}-c_{lj}^{(k)}\right)^2+\frac{1}{3}\left(r_{li}^{(k)}-r_{lj}^{(k)}\right)^2\right].$$



p=2

Average histogram based on d_W^2 Wasserstein distance

- The *barycenter* (average) *histogram* $Y(b)$ of a set of histogram data $Y(i)$ $(i=1, \ldots, n)$ can be computed minimizing the **Sum of Square Distances**

$$f(Y(b^*)) = \sum_{i=1}^{n} d_W^2 (Y(i), Y(b))$$

(like for a cloud of points in classical data analysis)

$$\boxed{f(Y(b^*)) = \sum_{i=1}^{n} \sum_{l=1}^{m} \pi_l \left[(c_{li} - c_{lb})^2 + \frac{1}{3}(r_{li} - r_{lb})^2 \right]}$$

That is minimized when the usual first order conditions are satisfied:

$$\begin{cases} \dfrac{\partial f}{\partial c_{lb}} = -2\pi_l \sum_{i=1}^{n} (c_{li} - c_{lb}) = 0 \\ \dfrac{\partial f}{\partial r_{lb}} = -\dfrac{2}{3}\pi_l \sum_{i=1}^{n} (r_{li} - r_{lb}) = 0 \end{cases} \Rightarrow \quad c_{lb} = n^{-1} \sum_{i=1}^{n} c_{li} \quad ; \quad r_{lb} = n^{-1} \sum_{i=1}^{n} r_{li}$$

$$Y(b) = \left\{ \left([c_{1b} - r_{1b}; c_{1b} + r_{1b}], \pi_1 \right); \ldots; \left([c_{lb} - r_{lb}; c_{lb} + r_{lb}], \pi_l \right); \ldots; \left([c_{mb} - r_{mb}; c_{mb} + r_{mb}], \pi_m \right) \right\}$$

Variance of the set of histograms Y(i)

- Determined *Y*(*b*) through the minimization problem

$$b^* = \arg\min \sum_{i=1}^{n} d_W^2(Y(i), Y(b))$$

- The variance of the set of histogram data is:

$$\sigma^2 = \sum_{i=1}^{n} \sum_{l=1}^{m} \pi_l \left[(c_{li} - c_{lb})^2 + \frac{1}{3}(r_{li} - r_{lb})^2 \right]$$

Clustering methods for Histogram data based on Wasserstein distance

- Dynamic Clustering Algorithm
- Hierarchical method (Ward criterion)

Dynamic Clustering Algorithm

general schema of the "Nuées Dynamiques" (Diday 1972)

The algorithm aims to obtain a partition **P** of

- a set **E** of symbolic data in *k* clusters and
- a set **L** of *k* prototypes $\{G_1,\ldots,G_i,\ldots,G_k\}$ that best represent the clusters $\{C_1,\ldots,C_i,\ldots,C_k\}$ of the partition **P**

The algorithm optimizes a criterion Δ of fitting between prototypes and clusters:

$$\Delta(P^*,L^*) = Min\left\{\Delta(P,L) \mid P \in P_k, L \in L_k\right\}$$

where : P_k is the set of all the partitions of E into k clusters and L_k is the set of k prototypes.

The algorithm executes alternatively:

an *allocation step*

and a representation step

In this case ***E*** is a set of histogram data

Wasserstein distance for Clustering data according to Dynamic Clustering algorithm classical schema

(a) Initialization

k prototypes $Y(b_1),\ldots,Y(b_K)$ of *L are randomly chosen*

(b) Allocation step

For each histogram $Y(i)$ of *E* the allocation index ℓ to the clusters is computed and $Y(i)$ is assigned to the cluster C_ℓ where:

$$\ell = \arg\min_{k=1,\ldots,K} d_W^2(Y(i), Y(b_k))$$

Wasserstein ℓ_2 distance

(c) Representation step

For each cluster C_k is identified the prototype $Y(b_k)$ of *L* that minimizes

$$\Delta(C_k, Y(b_k)) = \sum_{i \in C_k} d_W^2(Y(i), Y(b_k))$$

Histogram prototype/ barycenter of histograms belonging to the cluster C_k

(b) and (c) are repeated until the convergence

Property of Wasserstein distance: Inertia decomposition

Being the prototype the barycenter and d_W^2 is a squared Euclidean distance, then

$$TI = \sum_{i=1}^{n} d_W^2 (Y(i), Y(b)).$$

is the inertia for grouped data

Then, according to the Huygens' theorem *TI* can be decomposed in Within and Between inertia

$$TI = WI + BI =$$

$$= \sum_{h=1}^{k} \sum_{i \in C_h} d_W^2 (Y(i), Y(b_h)) + \sum_{h=1}^{k} |C_h| d_W^2 (Y(b_h), Y(b)).$$

Some results
Application on US monthly temperatures data set

- We have considered a dataset constituted by the "Monthly Average Temperatures recorded in the 48 states of US from 1895 to 2004 (Hawaii and Alaska are not present in the dataset).
- The analysis consists of the following three steps:

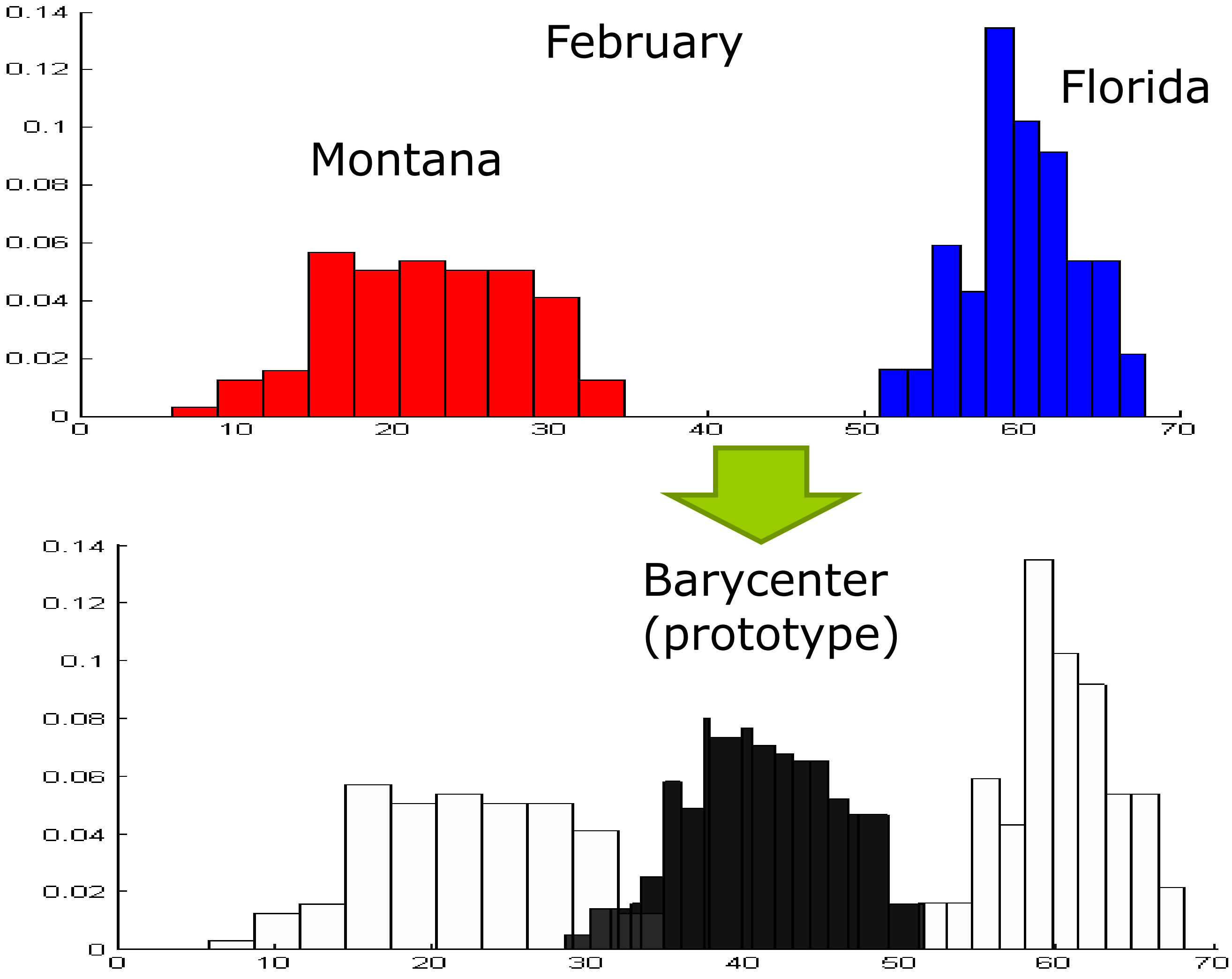
1. Representation of the distributions of temperatures of each State for each month by means of histograms;
2. Computing of the distance matrix using d^2_W;
3. DCA procedure to find the best partition P
4. Calinski Harabaz index is computed to compute the optimal number k of clusters
5. Hierarchical clustering procedure based on the Ward criterion

	A	B	C	D	E	F	G	H	I
1	State code	measure	Year	January	Febraury	March	April	May	June
2	1	2	1895	43,7	37,6	54,5	63,1	69,8	7
3	1	2	1896	44,1	47,9	52,5	67,9	75,7	
4	1	2	1897	42,6	51,2	60,2	62	68,6	8
5	1	2	1898	49,4	45,9	59	58,1	73,4	8
6	1	2	1899	44,4	39,9	55,2	61,6	75,8	7
7	1	2	1900	44	44,1	52,6	63,5	71,2	7
8	1	2	1901	46,1	43,1	53	57,4	70	7
9	1	2	1902	43,2	40,6	54,8	61,6	75,3	8
10	1	2	1903	43,6	48,2	59,2	60,7	69,4	7
11	1	2	1904	41,7	49,2	58,2	60	69,6	7
12	1	2	1905	39,1	39,5	59,5	63,5	74,4	7

The original dataset is freely available at the National Climatic Data Center

website http://www1.ncdc.noaa.gov/pub/data/cirs/drd964x.tmpst.txt

Histogram barycenter (prototype)



Experimental results

$$CH(k) = \frac{BI(k)/(k-1)}{WI(k)/(n-k)}$$

Dynamic Clastering algorithm results: The first 5 columns represent the Quality Partition index (min, max, means, median) on 100 iterations – the last one the best value of Calinski-Harabaz index

k	QPI min	QPI max	QPI mean	QPI median	QPI std dev	Best CH
2	0.6482	0.6555	0.6527	0.6527	0.0014	44.07
3	0.8069	0.8190	0.8125	0.8124	0.0029	70.94
4	0.8449	0.8666	0.8528	0.8522	0.0044	72.02
5	0.8494	0.8924	0.8663	0.8646	0.0080	**77.65**
6	0.8665	0.9086	0.8847	0.8811	0.0115	75.05
7	0.8577	0.9144	0.9000	0.9044	0.0126	69.63
8	0.8835	0.9233	0.9121	0.9135	0.0070	70.70
9	0.9028	0.9350	0.9174	0.9168	0.0048	64.05
10	0.8917	0.9360	0.9208	0.9202	0.0061	60.81

Hierarchical clustering – according to the “Ward criterion” - based on Wasserstein distance

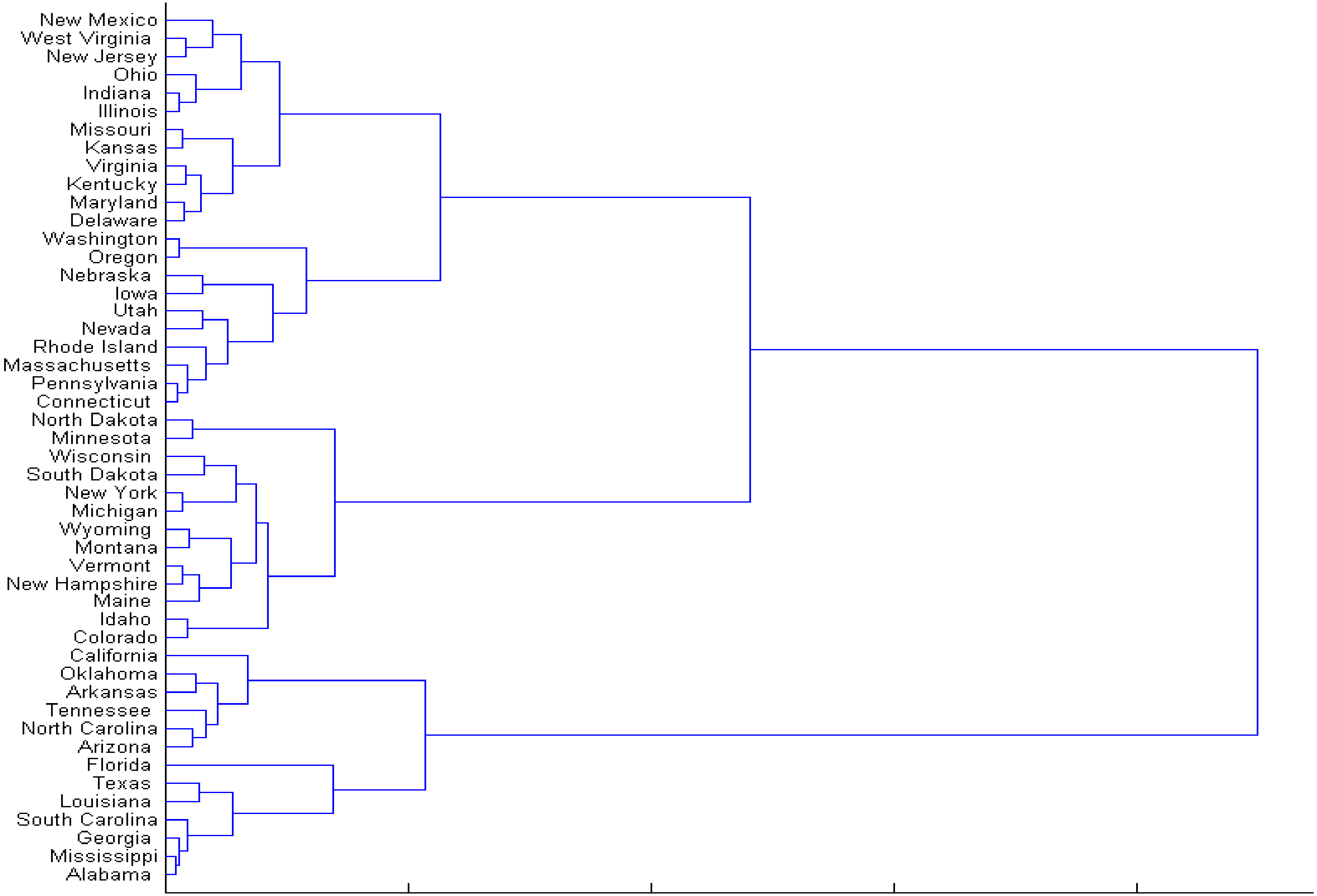
- For example, we can apply the Ward criterion for a hierarchical clustering of the set E of histogram data:

$$d_{Ward}(C_s, C_t) = \frac{|C_s||C_t|}{|C_s| + |C_t|} d_W^2(Y(b_s), Y(b_t))$$

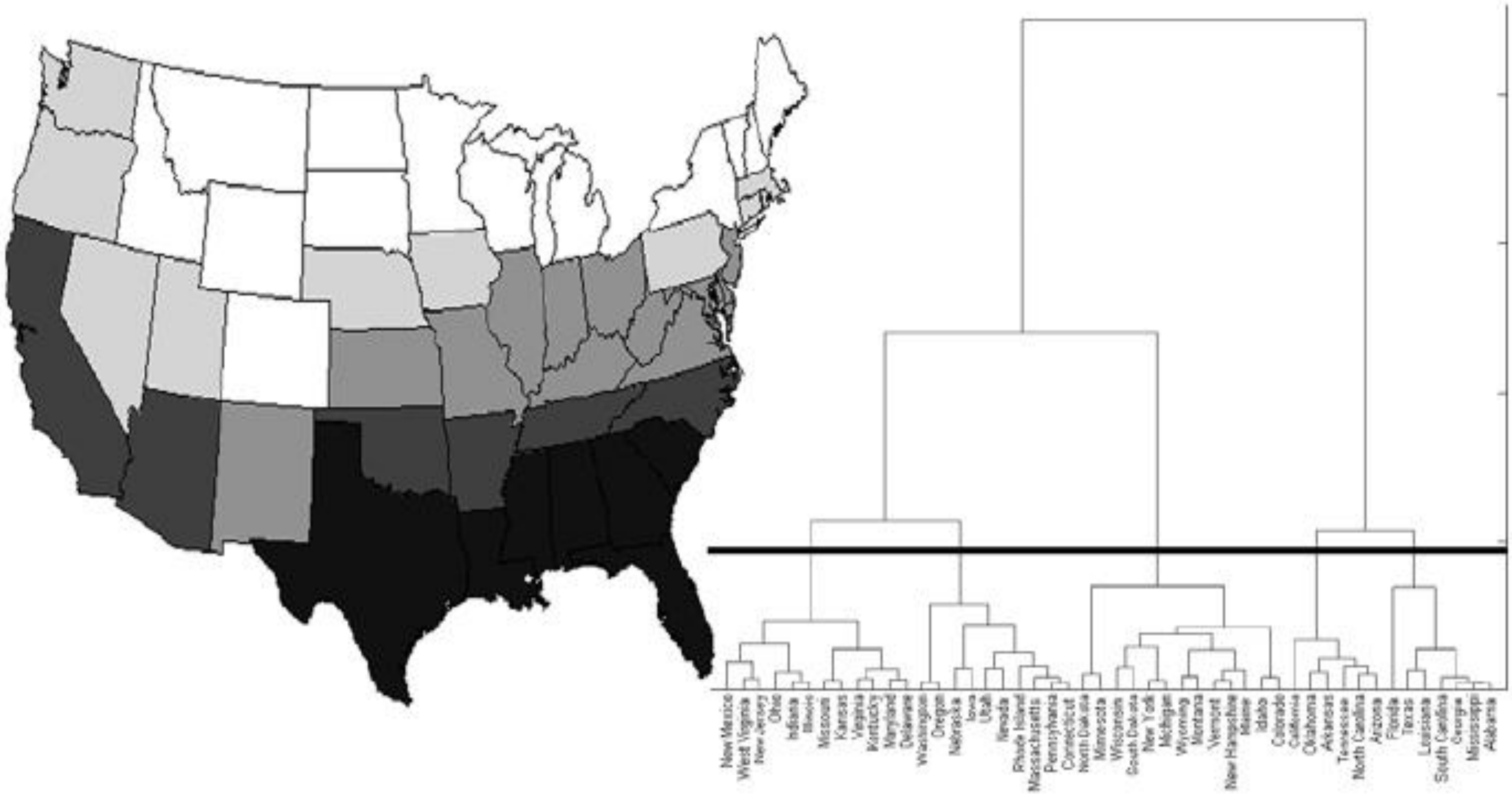
the procedure joins those two classes which minimize

$$TI(C_s \cup C_t) = TI(C_s) + TI(C_t) + \frac{|C_s||C_t|}{|C_s| + |C_t|} d_W^2(Y(b_s), Y(b_t))$$

Hierarchical tree



Final results and coloured map



Regression model for histogram variables based on Wasserstein distance

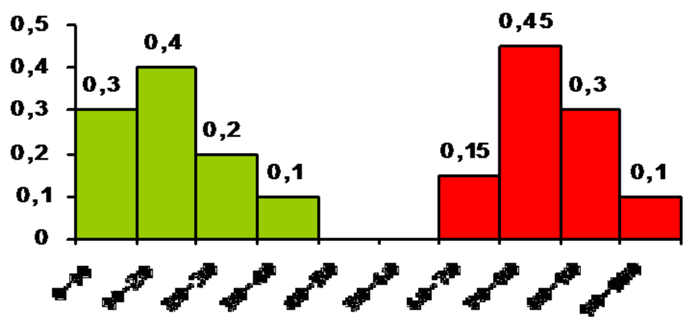
A Regression model for histogram data

Data = Model Fit + Residual

- Linear regression is a general method for estimating/describing association between a **continuous** outcome variable (dependent) and one or multiple predictors in one equation.

Easy conceptual task with classic data

But what does it means when dealing with histogram data?



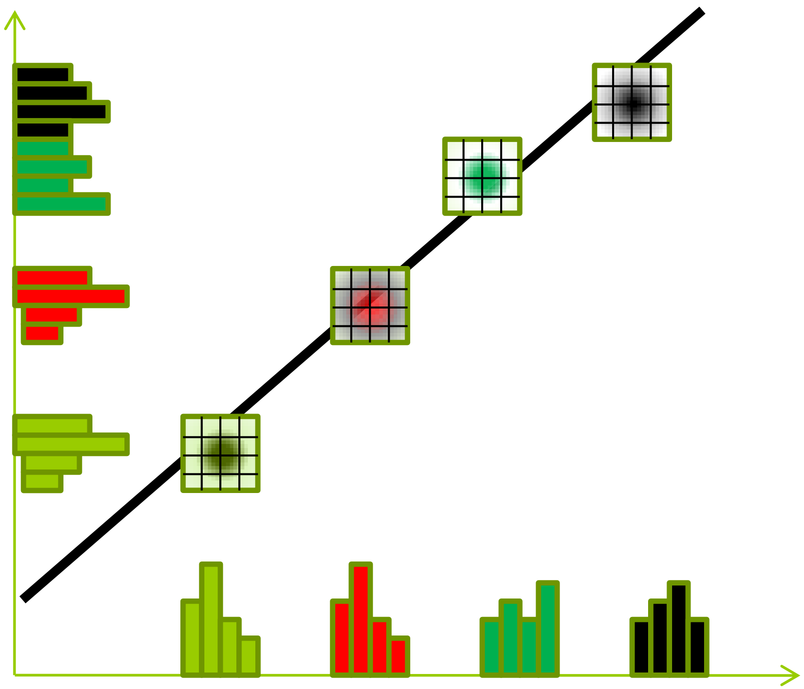
Billard, Diday, IFCS 2006
Verde, Irpino, COMPSTAT 2010; CLADAG 2011
Dias, Brito, ISI 2011

Regression with histograms variables: a proposal in SDA framework

A solution was given by Billard and Diday (2006)

- The model fit a linear regression line throught the mixture of the *n* bivariate distributions

- Given a punctual value of X it is possible to predict the punctual value of Y

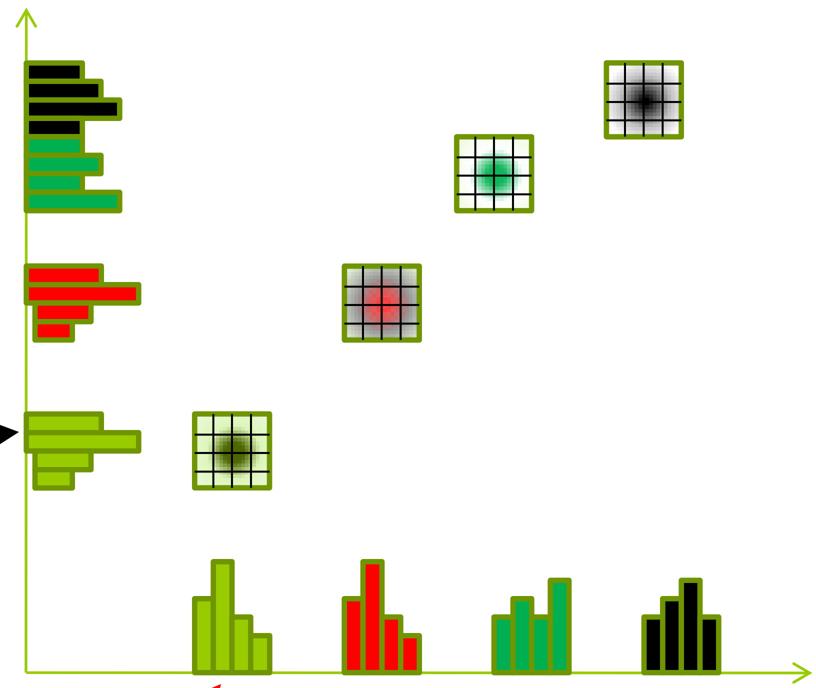


Linear Regression Model for histogram data:

our approach involves SDA as well as Functional Data Analysis (Verde, Irpino, 2010)

- Given a histogram variable X, we search for **a linear trasformation** of X which allows us to predict the histogram variable Y

- For example:
  given **the histogram of the temperatures observed in a region during a month**, is it possible to predict **the distribution of the temperature of another month** using a linear transformation of the histogram variable?



A histogram by a histogram

Multiple regression model for quantile functions

Our concurrent multiple regression model is:

$$y_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}(t) + \varepsilon_i(t)$$

Quantile functions associated to histogram/ distribution data

in matrix notation:

$$Y(t) = X(t)\beta + \varepsilon(t)$$

This formulation is analogous to the functional linear model (Ramsay, Silverman, 2003) except for the constants β parameters and for the functions $y_i(t)$, $x_{ij}(t)$ which are <u>quantile functions</u>

while each $\varepsilon_i(t)$ is a residual function (distribution?) for all $i=1, \ldots, n$.

Parameters estimation - LS method using Wasserstein distance

According to the nature of the variables, for the parameters estimation, we propose to extend the Least Squares principle to the functional case using a typical metric between quantile functions:

$$d_W^2\left(x_i, x_j\right) = \int_0^1 \left(F_i^{-1}(t) - F_j^{-1}(t)\right)^2 dt$$

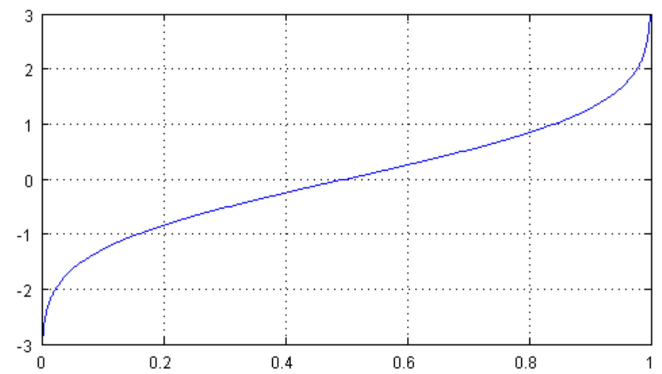
Wasserstein l2 distance between two quantile functions

Interpretative decomposition of the distance in three components related to the *location – size* and *shape* parameters

$$d_W^2(x_i, x_j) := \int_0^1 \left(F_i^{-1}(t) - F_j^{-1}(t) \right)^2 dt =$$

$$= \underbrace{\left(\bar{x}_i - \bar{x}_j \right)^2}_{Location} + \underbrace{\left(\sigma_i - \sigma_j \right)^2}_{Size} + \underbrace{2\sigma_i \sigma_j (1 - \rho(x_i, x_j))}_{Shape}$$

In *general* case of distributions

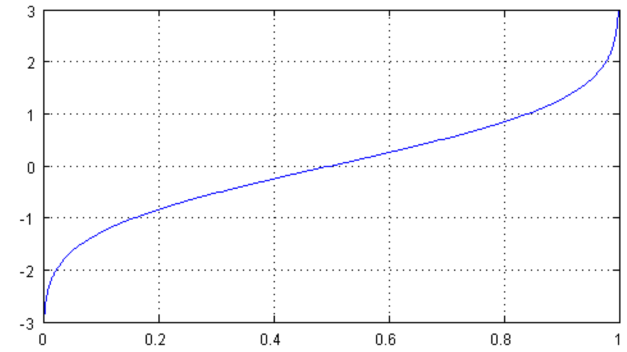


If the two distributions have the same shape:

$$d_W^2(x_i, x_j) := \underbrace{\left(\bar{x}_i - \bar{x}_j \right)^2}_{Location} + \underbrace{\left(\sigma_i - \sigma_j \right)^2}_{Size}$$

If they have the same size and shape: $d_W^2(x_i, x_j) := \underbrace{\left(\bar{x}_i - \bar{x}_j \right)^2}_{Location}$

Notations



$x_i(t) = F_i^{-1}(t)$ quantile function of x_i

Mean and variance of the quantile function:

$$\bar{x}_i = \int_0^1 x_i(t)dt$$

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) \quad \forall t \in [0,1]; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 x_i(t)dt = \int_0^1 \bar{x}(t)dt$$

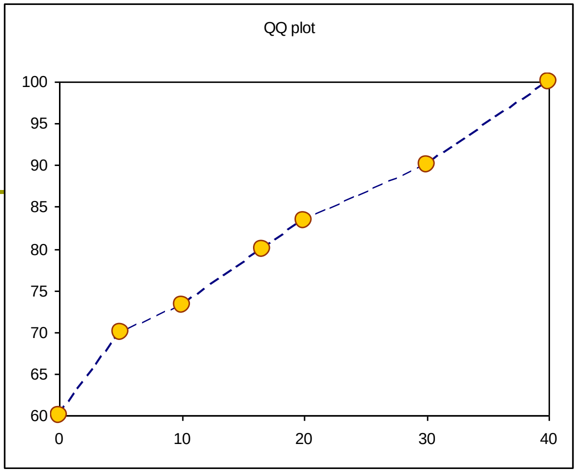
$$\sigma_{x_i}^2 = \int_0^1 [x_i(t)]^2 dt - [\bar{x}_i]^2 \Rightarrow \int_0^1 [x_i(t)]^2 dt = \sigma_{x_i}^2 + [\bar{x}_i]^2$$

Correlation between quantile functions (x_i , x_j)

$$\rho(x_i, x_j) = \frac{\int_0^1 x_i(t)x_j(t)dt - \bar{x}_i\bar{x}_j}{\sigma_{x_i}\sigma_{x_j}} \Rightarrow \int_0^1 x_i(t)x_j(t)dt = \rho(x_i, x_j)\sigma_{x_i}\sigma_{x_j} + \bar{x}_i\bar{x}_j$$

If data are histograms the mean, variance and correlation formulae are:

Empirical quantile function



Mean and variance of the quantile function:

$$\bar{x}_i = \int_0^1 x_i(t)dt \quad \Leftrightarrow \quad \bar{x}_i = \sum_l \pi_l c_{il}$$

$$\sigma_{x_i}^2 = \int_0^1 [x_i(t)]^2 dt - [\bar{x}_i]^2 \quad \Leftrightarrow \quad \sigma_{x_i}^2 = \sum_l \pi_l \left(c_{il}^2 + \frac{1}{3} r_{il}^2 \right) - \bar{x}_i^2$$

Correlation between quantile functions (x_i , x_j)

$$\rho(x_i, x_j) = \frac{\int_0^1 x_i(t)x_j(t)dt - \bar{x}_i \bar{x}_j}{\sigma_{x_i}\sigma_{x_j}} \Leftrightarrow \rho(x_i, x_j) = \frac{\sum_l \pi_l \left[c_{il}c_{jl} + \frac{1}{3} r_{il} r_{jl} \right] - \bar{x}_i \cdot \bar{x}_j}{\sigma_i \cdot \sigma_j}$$

Interpretation of the Linear Regression model

- The regression model is here proposed to find the best linear transformation of the $X_j(t)$'s in order to predict $Y(t)$

$$y_i(t) = \beta_0 + \beta_1 x_{1i}(t) + \ldots + \beta_p x_{pi}(t) + \varepsilon_i(t)$$

$y_i(t), x_{1i}(t), \ldots, x_{pi}(t)$ are quantile functions and the estimated response variable $\hat{y}_i(t)$ is again a quantile function = linear combination of quantile functions - according to the aim of symbolic data analysis:

Input Symbolic data ⇔ *Numerical/Symbolic method* ⇔ Output Symbolic data

(*same nature of the input data*)

Fitting linear regression model

- Find a linear transformation of the quantile functions of x_{ij} *(for $j=1,\dots,p$)* in order to predict the quantile function of y_i i.e.:

$$\hat{y}_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}(t) \quad \forall t \in [0,1]$$

It is worth of noting the linear transformation <u>is unique</u>: the parameters β_0 and β_j are estimated for all the x_{ij} and y_i distributions

- A first problem:

  Only if $\beta_j > 0$ a quantile function $\hat{y}_i(t)$ can be derived.

  In order to overcome this problem, we propose a solution based on the decomposition of the Wasserstein distance and on the NNLS algorithm.

Solution

- The quantile function can be decomposed as:

$$x_{ij}(t) = \bar{x}_{ij} + x_{ij}^c(t) \quad where$$

$$x_{ij}^c(t) = x_{ij}(t) - \bar{x}_{ij} \quad is\ the\ centered\ quantile\ function$$

- Then, we propose the following regression model:

$$y_i(t) = \underbrace{\beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j x_{ij}^c(t)}_{\hat{y}_i(t)} + \epsilon_i(t) \quad 0 \le t \le 1$$

- Using the Wasserstein distance it is possible to set up a OLS method that returns the two sets of coefficients $(\beta_0, \beta_j;\ \gamma_j)$.

The error term:

a property of the Wasserstein distance decomposition

- The squared error can be written according to the two components

$$\varepsilon_i^2 = d_W^2(y_i, \hat{y}_i) = \int_0^1 \left(y_i(t) - \hat{y}_i(t) \right)^2 dt =$$

$$= \boxed{\left(\bar{y}_i - \bar{\hat{y}}_i \right)^2} + \boxed{d_W^2(y_i^c, \hat{y}_i^c)}$$

Least Squares parameters estimation

$$\arg\min_{\beta_j,\gamma_j} f(\beta_j,\gamma_j) = \sum_{i=1}^{n} \varepsilon_i^{\,2}(t) = \sum_{i=1}^{n} d_W^2\left(y_i(t), \hat{y}_i(t)\right)$$

$$SSE = f(\beta_j,\gamma_j) = \sum_{i=1}^{n} \int_0^1 \left[\bar{y}_i + y_i^c(t) - \beta_0 - \sum_{j=1}^{p} \beta_j \bar{x}_{ij} - \sum_{j=1}^{p} \gamma_{ij} x_i^c(t) \right]^2 dt$$

Matrix notation:

$$SSE = \int_0^1 \left[\bar{Y} + Y^c(t) - \bar{X}\mathrm{B} - X^c(t)\Gamma \right]^2 dt$$

The estimated parameters

Correlation between quantile functions x_i y_i

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \bar{x}_i \bar{y}_i - n\bar{y}\,\bar{x}}{\sum_{i=1}^{n} \bar{x}_i^2 - n\bar{x}^2}; \quad \hat{\gamma}_1 = \frac{\sum_{i=1}^{n} \rho(x_i, y_i)\sigma_{x_i}\sigma_{y_i}}{\sum_{i=1}^{n} \sigma_{x_i}^2}$$

- It is easy to see that:

$$\hat{\beta}_0, \hat{\beta}_1 \in \Re \ \ and \ \ \hat{\gamma}_1 \geq 0$$

Multiple regression estimated parameters

According to the properties of the Wasserstein distance between quantile functions, the elements of the product matrix $\left(\int_0^1 X^c(t)' X^c(t)\right)$ computed according to our definition of a inner product operator between two quantile functions x_{ij} and $x_{i'k}$:

$$\int_0^1 x_{ij}(t) x_{i'k}(t) dt = \rho(x_{ij}, x_{i'k}) \sigma_{x_{ij}} \sigma_{x_{i'k}}$$

Then, the parameters are estimated under the constraint $\hat{\gamma}_j \geq 0$ (j=1,…, p) using the NNLS (Lawson , Hanson, 1974).

$$\hat{\mathrm{B}} = \left(\bar{X}'\bar{X}\right)^{-1}\bar{X}'\bar{Y}$$

$$\hat{\Gamma} = \left[\int_0^1 \left(\tilde{X}^c(t)'\tilde{X}^c(t)\right) dt\right]^{-1} \left[\int_0^1 \left(\tilde{X}^c(t)' Y^c(t)\right) dt\right]$$

$$\tilde{X}^c = \begin{cases} X^c \text{ if } \gamma_j > 0 \\ \textit{otherwise} \text{ it is a transformed matrix by NNLS} \end{cases}$$

Interpretation of the parameters

- Regression parameters for the distribution mean locations

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \in \Re$$

- Shrinking factors for the variability

$$\hat{\gamma}_1, \ldots, \hat{\gamma}_p \in \Re^+$$

- > 1 (< 1) the $\hat{y}_i$ histogram has a greater (smaller) variability than the x_{ij} histogram.

Tools for the interpretation

- The sum of squares of Y is

$$SS(Y) = \sum_{i=1}^{n} d_W^2\left(y_i(t), \bar{y}(t)\right) = \sum_{i=1}^{n} \int_0^1 \left[y_i(t) - \bar{y}(t)\right]^2 dt$$

In classical regression model, the SS(Y) is constituted by:

$$SS_{Error} + SS_{Regression}$$

Decomposition of SS(Y)

- Being: $$\hat{y}_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j \overline{x}_{ij} + \sum_{j=1}^{p} \gamma_j x_{ij}^c(t)$$
- we obtain

$$SS(Y) = \sum_{i=1}^{n} d_W^2\left(y_i(t), \overline{y}(t)\right) = \underbrace{\sum_{i=1}^{n} \int_0^1 [\hat{y}_i(t) - y_i(t)]^2 \, dt}_{SS_{Error}} +$$

$$+ \underbrace{\sum_{i=1}^{n} \int_0^1 [\overline{y}(t) - \hat{y}_i(t)]^2 \, dt}_{SS_{Regression}} \underbrace{-2n \int_0^1 \overline{y}(t) \overline{e}(t) dt}_{Bias}$$

$$\overline{e}(t) = \frac{1}{n} \sum_{i=1}^{n} (y_i(t) - \hat{y}_i(t)) \quad \forall t \in [0,1]$$

Average error function

The bias

- The bias is due to different shapes of distributions:

$$bias = -2n\left[\sigma_{\bar{y}}^2 - \sum_{j=1}^{p} \hat{\gamma}_j \rho(\bar{x}_j^c(t), \bar{y}(t)) \sigma_{\bar{x}_j^c} \sigma_{\bar{y}}\right]$$

Correlation between the average quantile functions

- *bias=0* when all the histograms data have the same shape

That represents the *incapacity* of the linear transformation of fitting distributions that are very different in shape

A measure of fitting

- Pseudo R^2

  Considering that

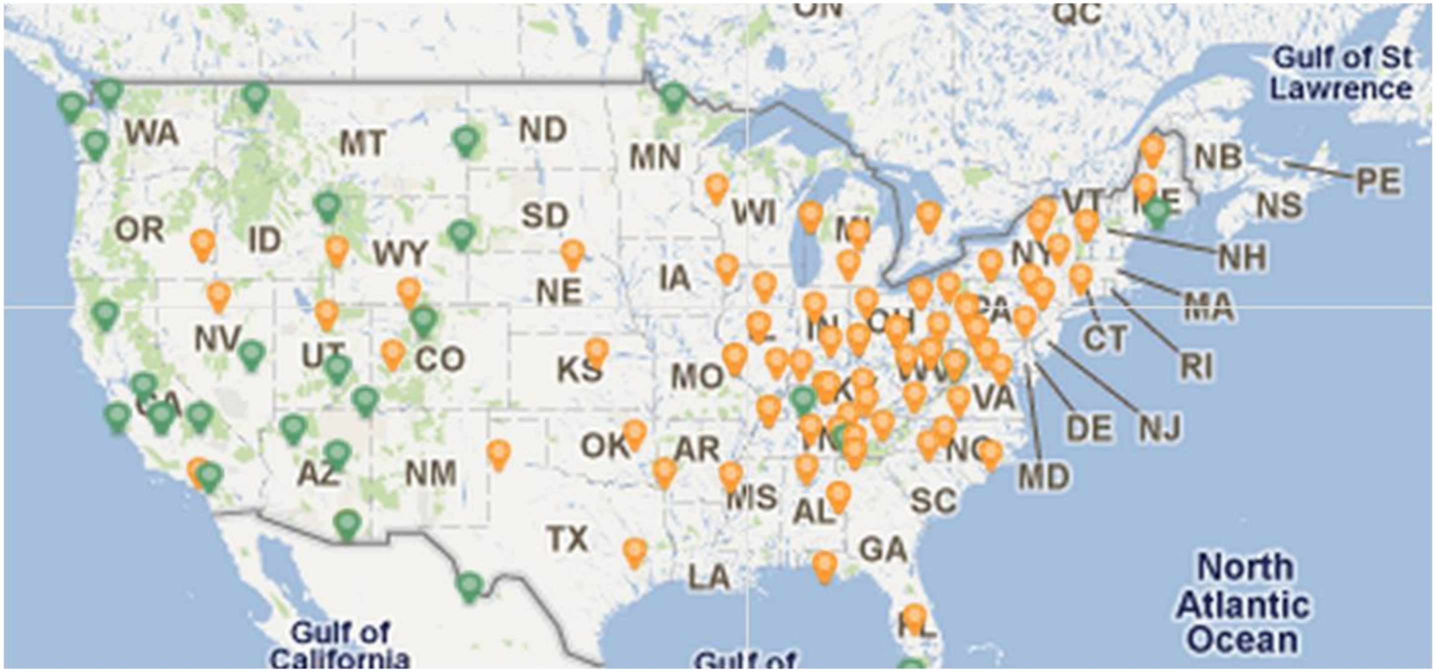
$$SS_{Regression} = \sum_{i=1}^{n}\left(\overline{y}_i - \overline{\hat{y}}_i\right)^2 + \sum_{i=1}^{n}\int_0^1\left[\overline{y}^c(t) - \hat{y}_i^c(t)\right]^2 dt - \int_0^1 \overline{y}(t)\overline{e}(t)dt$$

  We propose the following pseudo R^2

$$PseudoR^2 = \min\left[\max\left[0;1-\frac{SS_{Error}}{SS(Y)}\right];1\right].$$

An application on the Ozone concentration in 78 USA sites (http://java.epa.gov/castnet/)

Hourly monitoring of Ozone ground-level concentration and other meteorological variables



Forecasting OZONE levels

Ozone is a gas that can cause respiratory deseases.
In the literature there exists studies that relates the OZONE level to the Temperature, the Wind speed and the Solar radiation.

Given the distribution of **TEMPERATURE (C°)**, the distribution of **WIND SPEED** (meters per second) and the distribution of **SOLAR RADIATION** (watt per square meter), the main objective is to predict the distribution of **OZONE CONCENTRATION (**particles per billion**)** using a linear model.

We have chosen as period of observation the Summer seasons of 2010 and the central hours of the days (10 a.m. – 5 p.m.).
The model is:

$$OZONE(t) = \beta_0 + \beta_1 \overline{TEMP} + \beta_2 \overline{WSPEED} + \beta_3 \overline{SORAD} + \\ + \gamma_1 TEMP^c(t) + \gamma_2 WSPEED^c(t) + \gamma_3 SORAD^c(t) + \varepsilon(t)$$

The estimated model

Wasserstein based LS (WASS-LS)
(the current proposal)
In parentesis the 95% Bootstrap C.I.

$$OZONE(t) =$$

$$\underbrace{\mathbf{2.9272}}_{[-12.4;17.5]} - \underbrace{\mathbf{0.3456}}_{[-0.7888;0.1964]} \overline{TEMP} + \underbrace{\mathbf{0.3948}}_{[-1,28;2,34]} \overline{WSPEED} + \underbrace{\mathbf{0.0704}}_{[0,049;0,092]} \overline{SORAD} +$$

$$+ \underbrace{\mathbf{0.9153}}_{[0,49;1,36]} TEMP^{c}(t) + \underbrace{\mathbf{1.8867}}_{[1,08;3,20]} WSPEED^{c}(t) + \underbrace{\mathbf{0.0183}}_{[0,0118;0,024]} SORAD^{c}(t)$$

Billard (2006) regression (SREG)

$$OZONE = \underset{[19.17;34.74]}{\mathbf{26,49}} + \underbrace{\mathbf{0,2358}}_{[-0.03;0.51]} TEMP + \underset{[0.78;3.5]}{\mathbf{1,555}} WSPEED + \underbrace{\mathbf{0,0086}}_{[0.0052;0.0117]} SORAD$$

Diagnostics

Diagnostics	WASS-LS	SREG
Rsquare	NA	0.08
Pseudo Rsquare	0.57	0.08*
Ω (Dias-Brito 2011)	0.74	NA
RMSE_W	7.00	9,93*
RMSE_L2	1.06	1.58*

$$\Omega = \frac{\sum_{1=i}^{n} d_W^2\left(\hat{y}_i(t),\bar{Y}\right)}{\sum_{1=i}^{n} d_W^2\left(y_i(t),\bar{Y}\right)} \quad where\ \bar{Y} = n^{-1}\sum_{i=1}^{n}\bar{y}_i$$

$$RMSE_W = \sqrt{n^{-1}\sum_{i=1}^{n} d_W^2\left(y_i(t),\hat{y}_i(t)\right)}$$

$$RMSE_{L2} = \sqrt{n^{-1}\sum_{i=1}^{n} d_2^2\left(F_i(y),\hat{F}_i(y)\right)}$$

$$d_2^2 = \int_{-\infty}^{+\infty}\left(F_i(y)-\hat{F}_i(y)\right)^2 dy$$

* The Billard model does not allow to compute directly a distribution. Given a set of input distributions, a Montecarlo experiment is needed to compute an output distribution.

Main references

- BILLARD, L. and DIDAY, E. (2006): Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley Series in Computational Statistics. John Wiley & Sons.
- BOCK, H.H. and DIDAY, E. (2000): Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data. Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.
- CUESTA-ALBERTOS, J.A., MATRAN, C., TUERO-DIAZ, A. (1997): Optimal transportation plans and convergence in distribution. Journ. of Multiv. An., 60, 72–83.
- GIBBS, A.L. and SU, F.E. (2002): On choosing and bounding probability metrics. Intl. Stat. Rev. 7 (3), 419–435.
- IRPINO, A., LECHEVALLIER, Y. and VERDE, R. (2006): Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A., Vichi, M. (eds.) COMPSTAT 2006. Physica-Verlag, Berlin, 869–876.
- VERDE, R. and IRPINO, A.(2008): Comparing Histogram data using a Mahalanobis–Wasserstein distance. In: Brito, P. (eds.) COMPSTAT 2008. Physica–Verlag, Springer, Berlin, 77–89.
- VERDE, R. and IRPINO, A.(2010): Ordinary Least Squares for Histogram Data based on Wasserstein Distance COMPSTAT 2010
- DIAS, S. and BRITO P.,(2011) A new linear regression model for histogram-valued variables, ISI 2011

Thank you

- Rosanna Verde rosanna.verde@unina2.it
- Antonio Irpino irpino@unina.it

$$\rho(X_i, X_j) = \frac{\sum_l \pi_l \left[c_{il} c_{jl} + \frac{1}{3} r_{il} r_{jl} \right] - \bar{x}_i \cdot \bar{x}_j}{s_i \cdot s_j}$$