GeoXp : an R package for interactive exploratory spatial data analysis. Illustration with a data set of schools in Midi-Pyrénées.

Thibault Laurent, Anne Ruiz-Gazen and Christine Thomas-Agnan

Gremaq (Groupe de Recherche en Economie Mathématique et Quantitative), UT1 and LSP (Laboratoire de Statistique et Probabilité), UT3

15 June 2006

Introduction

Authors of GeoXp : Christine Thomas-Agnan, Anne Ruiz-Gazen and Yves Aragon (Gremaq, Toulouse). Version S-plus (2001), Matlab (2003) and R (2005).

Introduction

Authors of GeoXp : Christine Thomas-Agnan, Anne Ruiz-Gazen and Yves Aragon (Gremaq, Toulouse). Version S-plus (2001), Matlab (2003) and R (2005).

Tool for researchers in **spatial statistics**, **geography**, **ecology**,... for anyone who possesses a data set of variables measured at **geographical sites** or on **geographical zones** (cities, counties, countries,...)

Introduction

Authors of GeoXp : Christine Thomas-Agnan, Anne Ruiz-Gazen and Yves Aragon (Gremaq, Toulouse). Version S-plus (2001), Matlab (2003) and R (2005).

Tool for researchers in **spatial statistics**, **geography**, **ecology**,... for anyone who possesses a data set of variables measured at **geographical sites** or on **geographical zones** (cities, counties, countries,...)

Main objective : **exploratory spatial data analysis** and **coupling** between a **map** and a **statistical graph**.

Principle

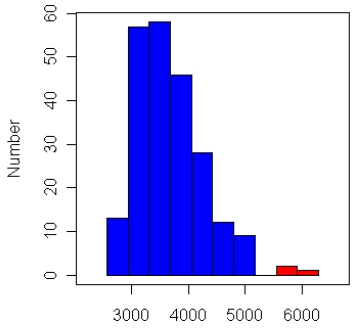
GeoXp links dynamically **statistical plots** like boxplot, histogram, scatterplot,... with a **map**.

Selection of a zone on the **map** results in the automatic highlighting of the **corresponding points** on the **statistical graph**.

Selection of a portion of the graph results in the automatic highlighting of the **corresponding sites** on the **map**.

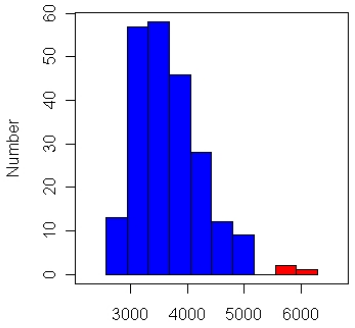
Principle

Example : sites selected by mouse clicking on bars of histogram are represented in red on the map.

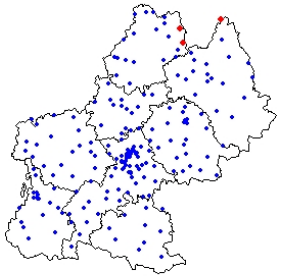


Principle

Example : sites selected by mouse clicking on bars of histogram are represented in red on the map.

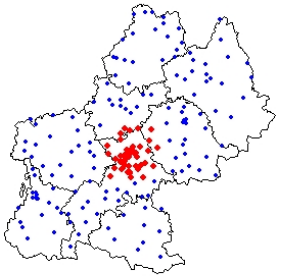


$\Rightarrow$



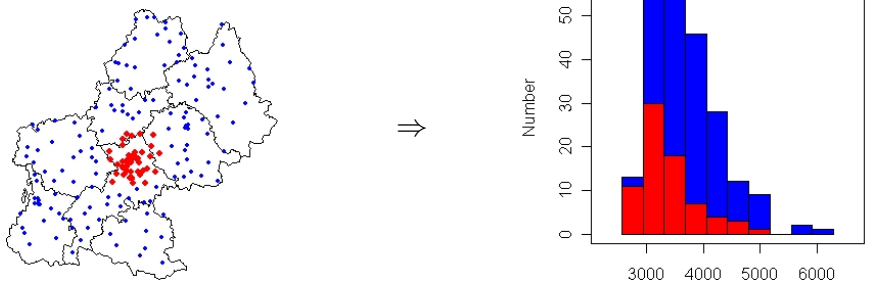
Principle

Example : sites selected by points or polygon on the map are represented in red on the histogram.



Principle

Example : sites selected by points or polygon on the map are represented in red on the histogram.



Functionality

Descriptive functions : univariate or bivariate graphs such as histogram, barplot, scatterplot,...

Functionality

Descriptive functions : univariate or bivariate graphs such as histogram, barplot, scatterplot,...

Geostatistic functions : angle plot, drift plot,...

Functionality

Descriptive functions : univariate or bivariate graphs such as histogram, barplot, scatterplot,...

Geostatistic functions : angle plot, drift plot,...

Econometric functions : Moran plot, neighbour plot,...

Functionality

Descriptive functions : univariate or bivariate graphs such as histogram, barplot, scatterplot,...

Geostatistic functions : angle plot, drift plot,...

Econometric functions : Moran plot, neighbour plot,...

Multivariate functions : principal component analysis, cluster analysis....

Use of GeoXp

Point pattern analysis : a site is represented on the map by a point. Coordinates are included into **two vectors** of **numeric values**.

Use of GeoXp

Point pattern analysis : a site is represented on the map by a point. Coordinates are included into **two vectors** of **numeric values**.

Data set is a **vector** or a **matrix** of **numerical** or **categorical** variables, associated to each **spatial unit**.

Use of GeoXp

Point pattern analysis : a site is represented on the map by a point. Coordinates are included into **two vectors** of **numeric values**.

Data set is a **vector** or a **matrix** of **numerical** or **categorical** variables, associated to each **spatial unit**.

Some options are common to all functions (spatial contours, bubbles,...) and others **depend on** the function (number of bars for histogram, use of colors for barplot,...).

Use of GeoXp

Point pattern analysis : a site is represented on the map by a point. Coordinates are included into **two vectors** of **numeric values**.

Data set is a **vector** or a **matrix** of **numerical** or **categorical** variables, associated to each **spatial unit**.

Some options are common to all functions (spatial contours, bubbles,...) and others **depend on** the function (number of bars for histogram, use of colors for barplot,...).

Remark : **no lattice/area data analysis** and **no use of Spatial Classes** such as in spdep library (Roger Bivand).

Examples

Example 1 (univariate analysis) :
histomap(*latitude*, *longitude*, *var*, *opt*1, *opt*2*a*)

Examples

Example 1 (univariate analysis) :
histomap(*latitude*, *longitude*, *var*, *opt*1, *opt*2*a*)

Example 2 (bivariate analysis) :
scattermap(*latitude*, *longitude*, *var*1, *var*2, *opt*1, *opt*2*b*)

Examples

Example 1 (univariate analysis) :
histomap(*latitude*, *longitude*, *var*, *opt*1, *opt*2*a*)

Example 2 (bivariate analysis) :
scattermap(*latitude*, *longitude*, *var*1, *var*2, *opt*1, *opt*2*b*)

Example 3 (multivariate analysis) :
pcamap(*latitude*, *longitude*, *dataset*, *opt*1, *opt*2*c*)

Examples

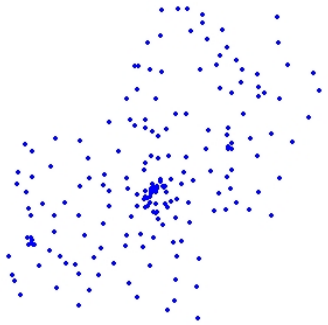
Example 1 (univariate analysis) :
histomap(*latitude*, *longitude*, *var*, *opt*1, *opt*2*a*)

Example 2 (bivariate analysis) :
scattermap(*latitude*, *longitude*, *var*1, *var*2, *opt*1, *opt*2*b*)

Example 3 (multivariate analysis) :
pcamap(*latitude*, *longitude*, *dataset*, *opt*1, *opt*2*c*)

Example 4 (spatial econometric analysis) :
moranplotmap(*latitude*, *longitude*, *var*, *W*, *opt*1, *opt*2*d*)

Common options



On the map :

Common options

On the map :

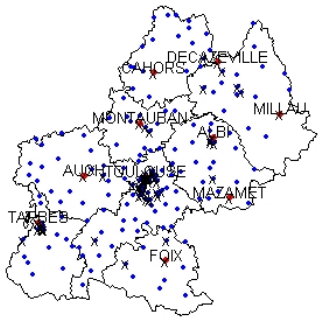
- Possibility to draw **spatial contours**.

Common options

On the map :

- Possibility to draw **spatial contours**.
- Possibility to cross out specific sites.

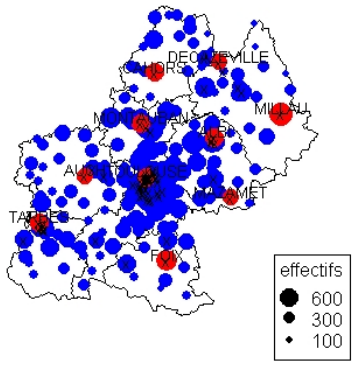
Common options



Possibility to improve the map :

- Possibility to draw **spatial contours**.
- Possibility to cross out specific sites.
- Possibility to print labels.

Common options

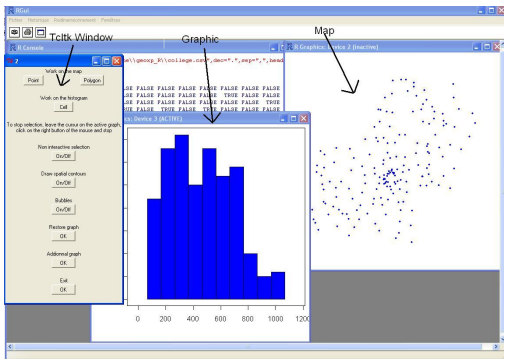


Possibility to improve the map :

- Possibility to draw **spatial contours**.
- Possibility to cross out specific sites.
- Possibility to print labels.
- Possibility to draw bubbles.

Interface

Call of a function ⇒ drawing of a graph, a map and creation of a **tcltk Window** (library `tcltk`), tool for **selecting** a zone on the map (or on the graph) and **using options**.



Use of GeoXp : data set considered

Spatial units s_j $(j = 1, ...226)$: **public schools** in French Midi-Pyrénées region. Spatial position of a school is represented by the **centroid of the "commune"** where the school is located.

Use of GeoXp : data set considered

Spatial units s_j $(j = 1, ...226)$: **public schools** in French Midi-Pyrénées region. Spatial position of a school is represented by the **centroid of the "commune"** where the school is located.

Observed variables : number of students, age of staff, different fields of study, status of teacher,... during the 2003-2004 school year.

Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
Principle
Use of GeoXp
Options
Interface
Description of Data

Use of GeoXp : data set considered

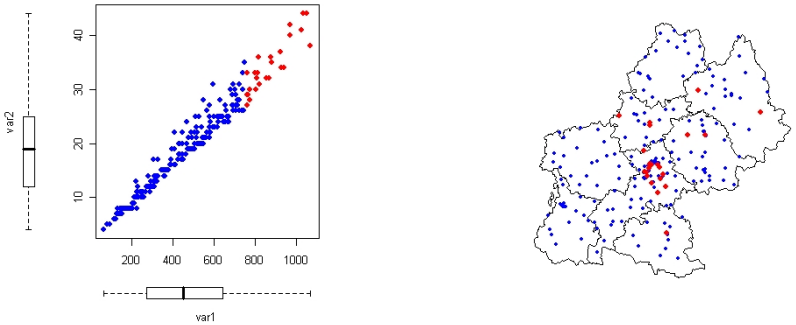
Spatial units s_j $(j = 1, ...226)$: **public schools** in French Midi-Pyrénées region. Spatial position of a school is represented by the **centroid of the "commune"** where the school is located.

Observed variables : number of students, age of staff, different fields of study, status of teacher,... during the 2003-2004 school year.

Aim : determine characteristic of schools according to their localization in rural, periurban and urban area (Insee classification).

Scatter plot

scattermap(X,Y,var1,var2,carte=coord,listvar=dataset)

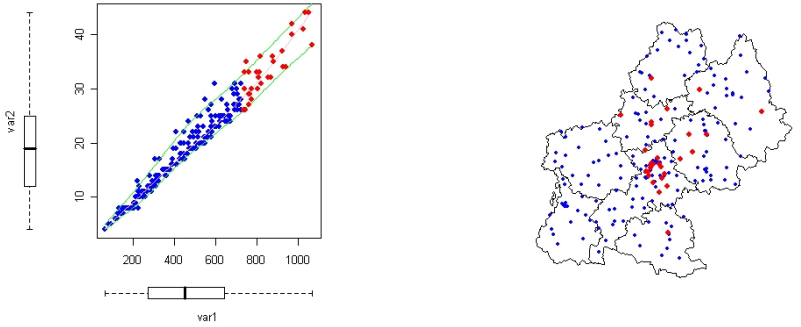


Number of classroom $\sim$ number of students

Thibault Laurent
Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
GeoXp
Scatterplot
Gini Map
Kernel Density Function estimate

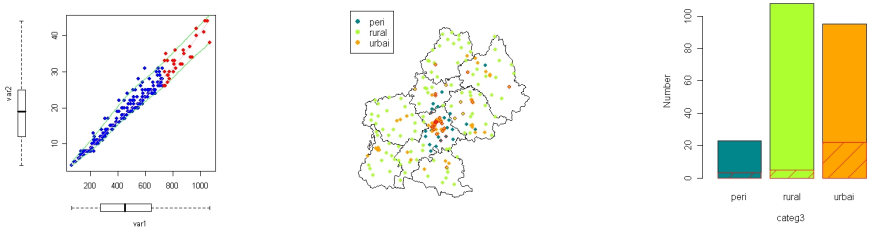
Scatter plot

Possibility to draw conditionnal quantile regression for a given list of quantiles orders (*ex : quantiles=c(0.05,0.5,0.95)*).



Scatter plot

Possibility to **add a graph** among **histogram**, **scatterplot** and **barplot**, by selecting variable(s) (given in *listvar*) on `tcltk` window .

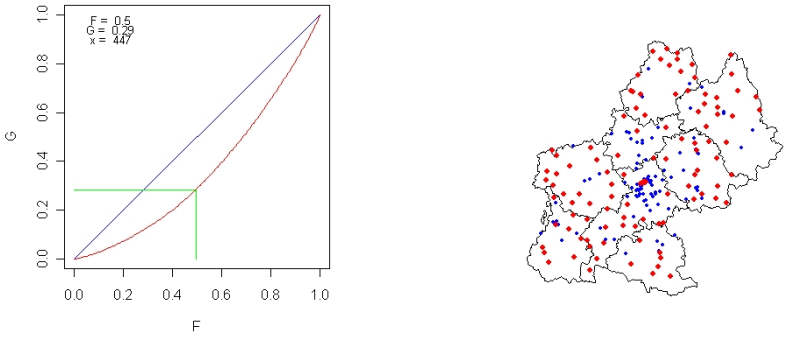


The schools selected are mainly included in urban area.

Gini Map

ginimap(lat,long,var,carte=coord,listvar=dataset)

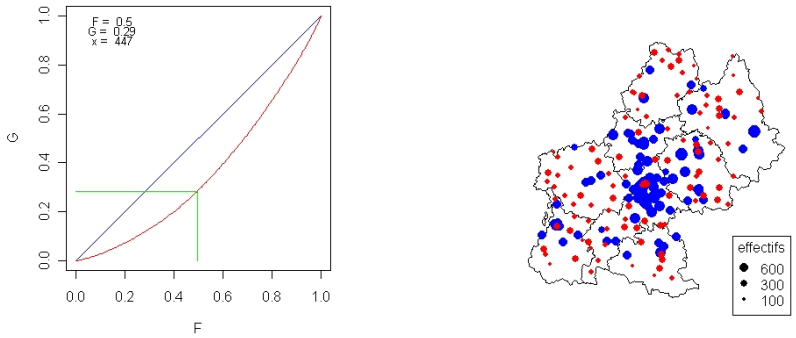
Computes a **Lorentz curve** from 'var' and calculates the **Gini Index** associate.



Gini Index = 0.28 ; the 50% schools with lowest number of students contain only 29% of the student population.

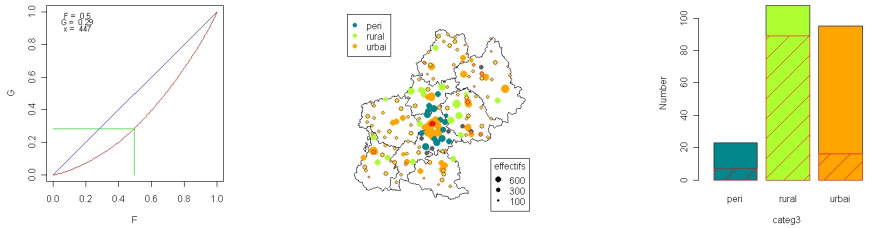
Gini Map

Possibility to draw bubbles by selecting on `tcltk` window a numerical variable among *listvar*.



Gini Map

The schools selected are mainly included in rural area.

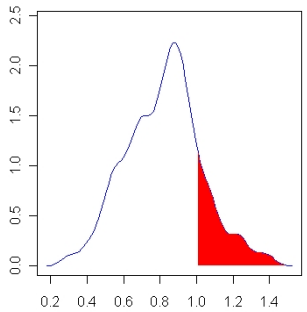


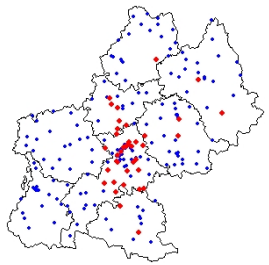
Kernel Density function estimate

densitymap(lat,long,var,carte=coord,listvar=dataset)

Use of function `bkde.R` (library `Kernsmooth`) with option *kernel=triweight*.

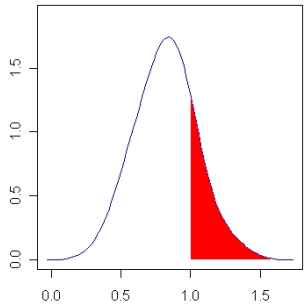
Possibility to choose an interval by **mouse clicking** on the graph on the **extremities** of interval or by **directly specifying values**.

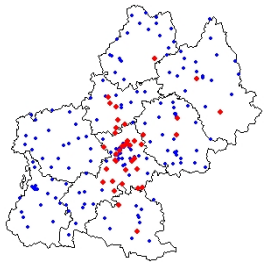




Kernel Density function estimate

Possibility to change **smoothing parameter** α with a **cursor** on `tcltk` window.



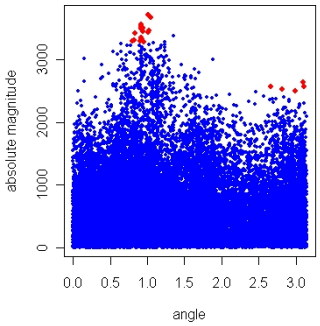


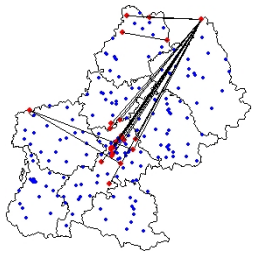
Urban schools with a high coefficient are close to Toulouse.

Angle plot

angleplotmap(lat,long,var,carte=coord,listvar=dataset)

Represents the **absolute difference** between the value of *var* at **two sites** as a function of the **angle** between **vector** $\overrightarrow{s_i s_j}$ and **the x-axis**.

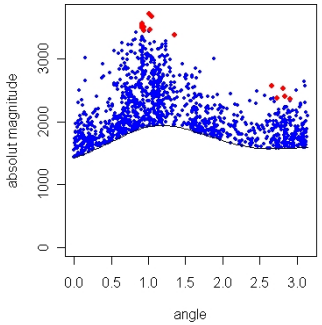


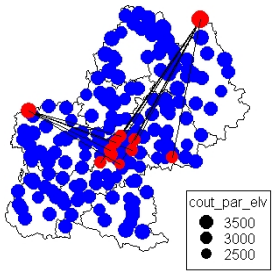


Variable average cost per student.

Angle plot

Possibility to draw only couple of sites whose absolute difference is **larger** than the 95% **regression quantile smooth spline**.

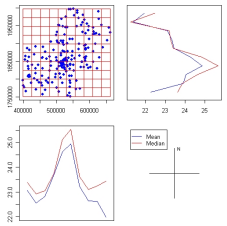




2 schools in the North and in the west with high average cost per student.

Drift map function

Creates a **grid** on the map and calculates **the mean and median for each cell**. The right plot (resp. left plot) represents row (resp. column) means and median.



Average number of student per class reaches a maximum in the center of the region which corresponds to the surroundings of Toulouse.

Spatial weight matrix

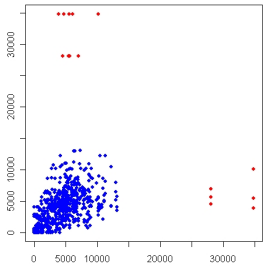
Possibility to create spatial weight matrix based on a **threshold distance** or based on a given number of **nearest neighbors**.

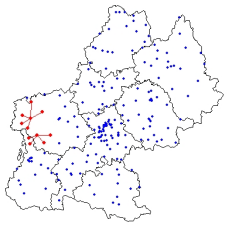
ex :

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Neighbour plot

Scatterplot of the values of a variable at **neighbouring sites** for a **neighbourhood structure** given by a **spatial weight matrix**.



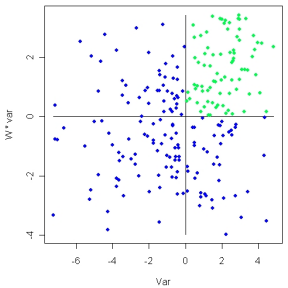


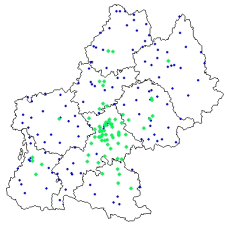
2 schools with a **high difference** with its neighbours for variable *average distance between school and home*. W built with 3 nearest neighbors.

Moran plot

On x-axis, is represented $(Var - \bar{Var})$ and on y-axis, is represented $W(Var - \bar{Var})$.

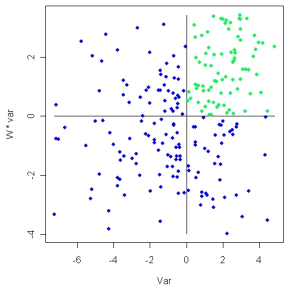
It also calcultes **Moran's I statistic** (see nonnormoran.r) and gives a **p-value** associated to the **gaussian test** or to the **permutation test**.

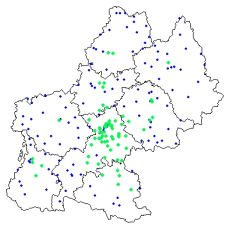


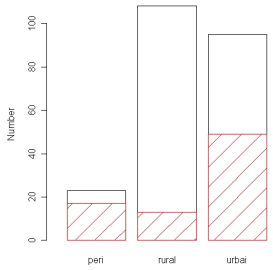


Moran plot

High spatial autocorrelation for *average number of student by class* : Moran'I statistic = 0.19 with $p - value < 0.0001$.





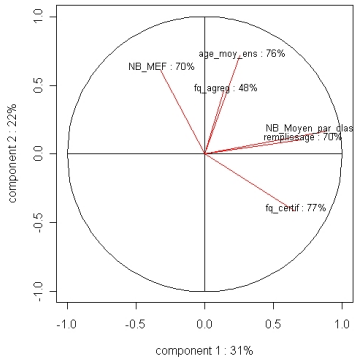


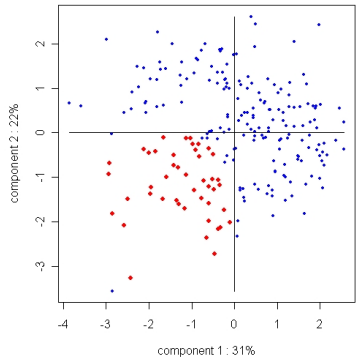
Schools with high values and whose neighbours have also high values are mainly included in urbain area.)

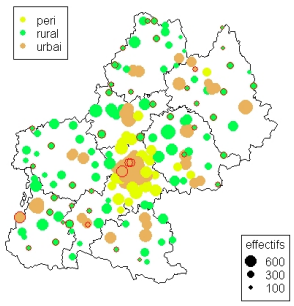
PCA

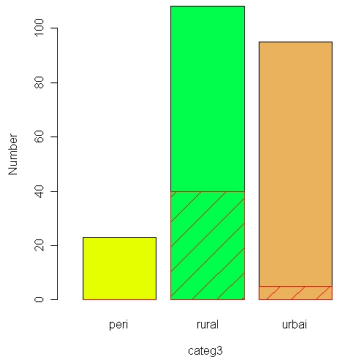
pcamap(lat,long,dataset)

Draws the plots summarizing a **generalized Principal Component Analysis** (PCA).







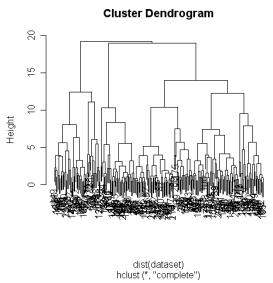


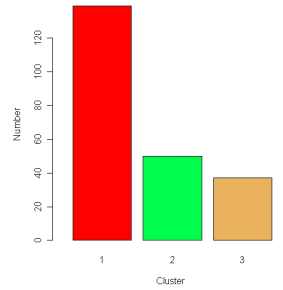
Cluster Analysis

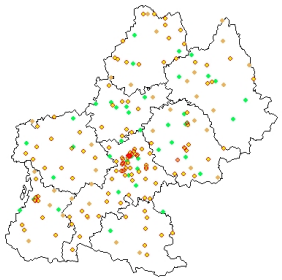
clustermap(lat,long,dataset,number.of.class)

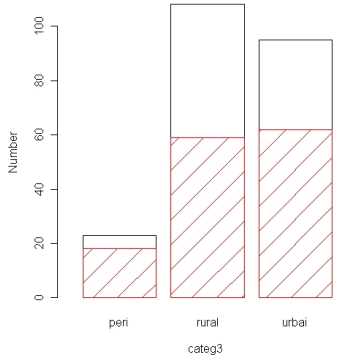
Performs a **classification** of the sites from the variables included in 'dataset' and **computes a bar plot** of the **clusters calculated**.

Two methods : *Hierarchical Cluster Analysis* (see hclust.R) or *k-means clustering* (see kmeans.R)









There seem to exist no link between calculated clusters and geographical area.

Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
Principal component analysis
Cluster analysis

Perspective

Writing of a GeoXp user manual.

Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
Principal component analysis
Cluster analysis

Perspective

Writing of a GeoXp user manual.

To include others functions as scatterplot 3D, sliced inverse regression, projection pursuit, spatial autoregressive models (SAR, SEM,...).

Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
Principal component analysis
Cluster analysis

Perspective

Writing of a GeoXp user manual.

To include others functions as scatterplot 3D, sliced inverse regression, projection pursuit, spatial autoregressive models (SAR, SEM,...).

Integration of `Micromaps` software (Carr and Symanzik).

Conclusion

You can download Geoχp on

http ://w3.univ-tlse1.fr/GREMAQ/Statistique/

Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
Principal component analysis
Cluster analysis

Conclusion

You can download Geoχp on

http ://w3.univ-tlse1.fr/GREMAQ/Statistique/

Don't try to ask me a demonstration

Presentation of GeoXp
Descriptive analysis
Geostatistic functions
Spatial Econometrics
Multivariate analysis
Principal component analysis
Cluster analysis

Conclusion

You can download Geoχp on

http ://w3.univ-tlse1.fr/GREMAQ/Statistique/

Don't try to ask me a demonstration

Merci de votre attention ! !