Spatial exploratory analysis of the Guerry's data with GeoXp

Thibault Laurent

Spatial exploratory analysis of the Guerry's data with GeoXp

Thibault Laurent[1]

[1] *Toulouse School of Economics, 21 allées de Brienne, 31000 Toulouse, France*
thibault.laurent@univ-tlse1.fr

Résumé. En 1833, André-Michel Guerry a écrit son célèbre Essai sur la statistique morale de la France. Récemment, Friendly (2007), Dray et Jombard (2011) et Filzmoser et al. (2012) ont appliqué sur ces données des méthodes statistiques modernes, issues de l'analyse multivariée ou de l'analyse de données spatiales. L'objectif de ce document est d'appliquer sur ces données historiques, les outils de l'analyse exploratoire interactive de données spatiales, en utilisant le package **GeoXp** (Laurent et al., 2012), disponible sur le logiciel R.

Mots-clés. André-Michel Guerry, analyse exploratoire de données spatiales, logiciel R, librairie **GeoXp**.

Abstract. In 1833, André-Michel Guerry wrote his famous Essai sur la statistique morale de la France. This data set has been recently studied in Friendly (2007), Dray and Jombard (2011), Filzmoser et al. (2012). The aim of this document is to present some tools of the R package **GeoXp** (Laurent et al., 2012) to analyze this data set.

Keywords. André-Michel Guerry, spatial exploratory data analysis, R software, package **GeoXp**.

1 Introduction

In 1833, André-Michel Guerry wrote his famous Essai sur la statistique morale de la France. This data set has been recently studied in Friendly (2007), Dray and Jombard (2011), Filzmoser et al. (2012). It is available within R (R Development Core Team, 2012) in the **Guerry** package including the Stéphane Dray's vignette "Spatial multivariate analysis of Guerry's data in R" (2009). The Guerry data frame comprises a collection of "moral variables" on the 86 departments of France around 1830.

The aim of this document is to present some tools of the R package **GeoXp** (Laurent et al., 2012) to analyze this data set. We focus essentially on the variables "crimes against property" and "crimes against persons". In the **Guerry** package there are several forms of the data. We consider here the `gfrance85` data which does not include the Corse department. The R packages used in this document are **GeoXp** and **Guerry** :

```
> require("GeoXp")
> require("Guerry")
```

Contrary to the original data, we consider the number of crimes against persons for 1 000 inhabitants rather than the population per crime against persons. Indeed, Guerry used the value 2 200 if 1 person has been accused for 2 200 people. In this study, we will rather use the value 0.4545 crimes for 1 000 inhabitants. Indeed, the latest statistics on crime given by the French statistical institute are expressed using this measure[1]. For that reason, we apply the following transformation to the variables of interest:

```
> gfrance85@data$Crime_pers<-1000/gfrance85@data$Crime_pers
> gfrance85@data$Crime_prop<-1000/gfrance85@data$Crime_prop
```

2 Non parametric estimates of the variables crimes

To represent the non parametric estimates of two variables density, we use the function `dbledensitymap`. On figure 1, we select the 20% highest values of the variable Crimes against persons by clicking on the tail of the non parametric distribution (left panel). The corresponding selected departments are shown on the map (right panel) in red; they correspond to the departments of South and East of the France. The non parametric estimate of the selected departments is represented by the dotted red curve on the graphic of the variable Crimes against property (middle panel). We notice that the initial distribution is not very different from the distribution of the selected departments which means that the highest values of the crimes against persons do not correspond to the highest value of the crimes against property. Moreover, the Pearson correlation coefficient is very low (10%).

```
> row.names(gfrance85@data)<-as.character(gfrance85@data$Department)
> dbledensitymap(gfrance85, c("Crime_pers","Crime_prop"),
   xlab=c("Crimes against persons","Crimes against property"), identify=TRUE)
```

3 Is there any geographical drift for the variables crimes?

To answer to this question, we use the function `driftmap` which cross rules the map into 6 rows and 6 columns and represents the mean and median by row (on the top-right panel) and by column (on the bottom-left panel). The figure 2 shows that the Crimes against persons present a positive drift in the direction North-South and also in the direction West-Est (left panel). Concerning the variable Crimes against property (right panel), there is a negative tendency in the direction North-South, but this drift does not seem very strong. There is no tendency in the direction West-Est.

```
> driftmap(gfrance85,"Crime_pers")
> driftmap(gfrance85,"Crime_prop")
```

[1] See `http://www.cartocrime.net/Cartocrime2/index.jsf`

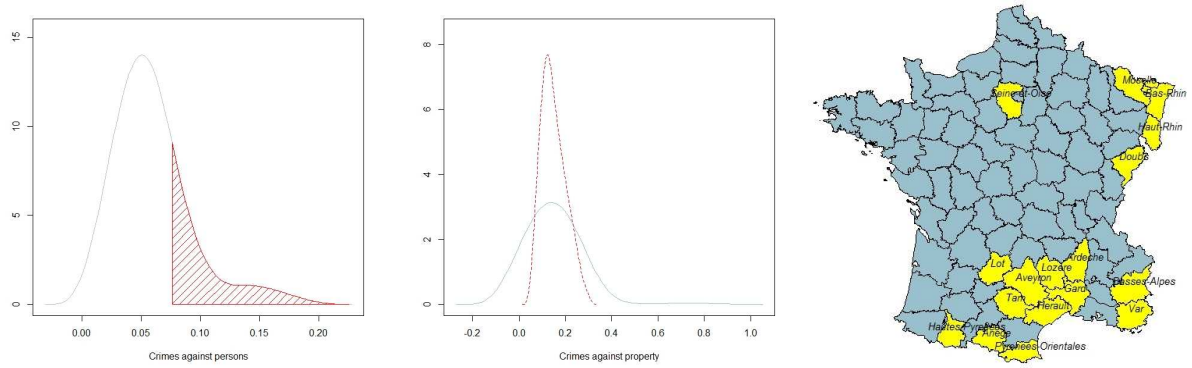

Figure 1: Example of use of the function `dbledensitymap` applied on the variables Crimes. This version of GeoXp (upper than 1.6.0) permits to represent the polygons of the selected units rather than the centroïds.

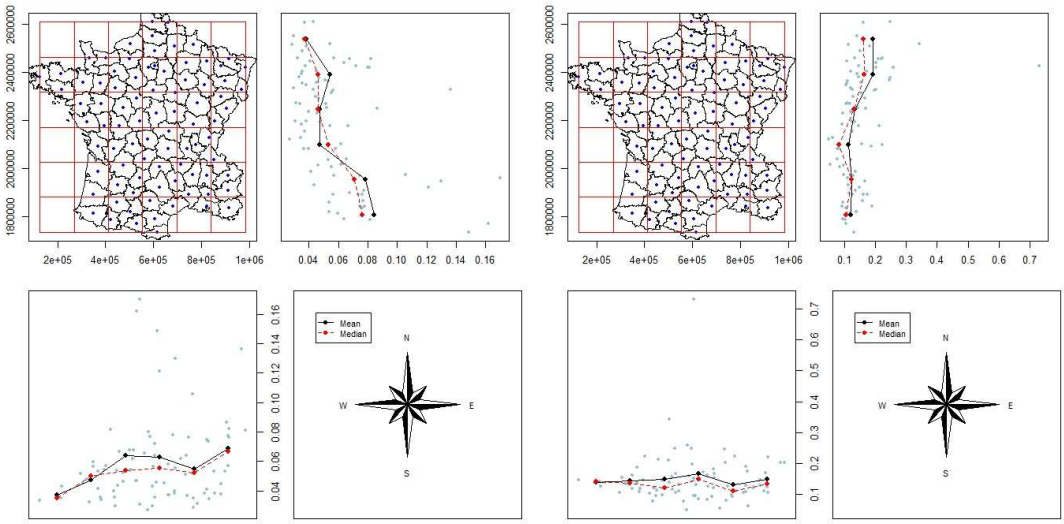

Figure 2: Example of use of the function driftmap applied to the variables Crimes.

4 Is there some spatial autocorrelation in the variable crimes?

To compute the Moran's I statistic, we need a spatial weight matrix W. For this, we use the function `poly2nb` in **spdep** package (Bivand et al., 2008) which computes a spatial contiguity matrix. W is row standardised (sums over all links to n). Then, we use the function `moranplotmap` to represent the Moran's scatterplot and computes the Moran's I statistic. The Moran's I statistic is equal to 0.3786 for the crimes against persons and 0.1932 for the Crimes against property and is significant for each of the variable.

Thus, there is a strong spatial autocorrelation (essentially for Crimes against persons). This is confirmed by the figure 3 which represent the Moran's scatterplot (for the crimes against person only). Different colors are given to the department depending on the quadrant they belong to (High-High, High-Low, Low-High and Low-Low). The size of the circles depend on the LISA (Local Indicator of Spatial Autocorrelation). The clusters are different depending on the kind of crimes: South and West contain the high values for Crimes against persons.

```
> lw <- nb2listw(poly2nb(gfrance85),style="W")
> moranplotmap(gfrance85, "Crime_pers", lw)
```

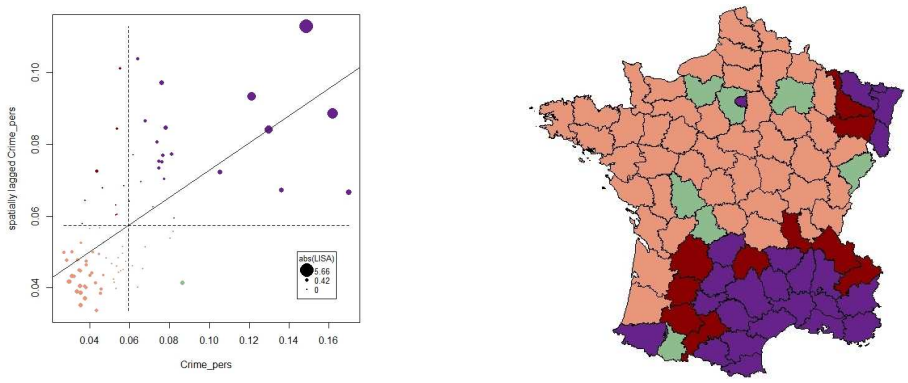

Figure 3: example of use of the function of the function moranplotmap (here, for the crimes against persons). The size of the departments on the map is proportional to the crimes.

5 Is there a link between Crimes and some categorical variables?

On figure 4, we use the function `polyboxplotmap` which represents the parallel boxplots of the variable crimes against property depending on the Literacy which have been cutting into 4 classes. We notice that in the highest classes $]40\%; 60\%]$ and $]60\%; 74\%]$ of military conscripts reading and writing), crimes against property are a little bit higher than in the small classes. Moreover, these high classes correspond to departments mostly located in the North of France.

```
> gfrance85@data$Literacy_b <- cut(gfrance85@data$Literacy,breaks=c(1,20,40,60,85))
> polyboxplotmap(gfrance85, c("Literacy_b", "Crime_prop"), identify=TRUE)
```

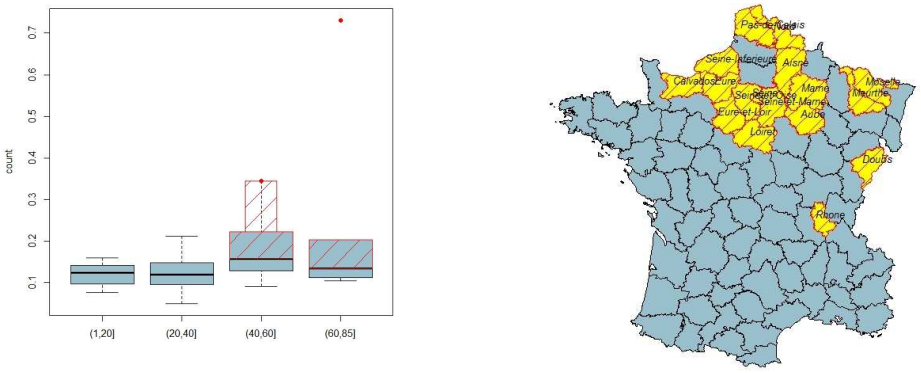

Figure 4: Example of use of the function polyboxplotmap applied on the variable Crimes against property and Literacy.

6 Is there a link between the Crimes and numeric variables?

The variable the most linearly correlated to Crimes against persons is the variable distance to Paris (45%). On figure 6, we represent the scatter plot of the Crimes against persons depending on the distance to Paris. The link between the 2 variables is obvious and positive. We select the departments which are the most "influent" in the simple linear regression. These departments are underestimated by the model and mostly located in the South of France. The variables the most linearly correlated to Crimes against property are the variables Prostitutes (85%), Infants (76%) and Suicides (70%).

```
> scattermap(gfrance85, c("Distance","Crime_pers"),  identify=TRUE)
```

7 Conclusion

Only a few tools of the package **GeoXp** have been presented in this document. We could also have used the function `ginimap` to represent the Lorentz curve of the variable Donations to the poor or the function `pcamap` to compute a principal component analysis applied to all the numeric variables.

We can conclude that the two variables related to Crimes are spatially different. Crimes against persons are more important in the South and East of France whereas Crimes against property is more important in the North near the department of Seine (Paris) and also in departments containing big cities. Each variable is spatially autocorrelated which confirms that Crimes are more present in specific areas and not distributed randomly. Finally, crimes against property are related to characteristics such as the number

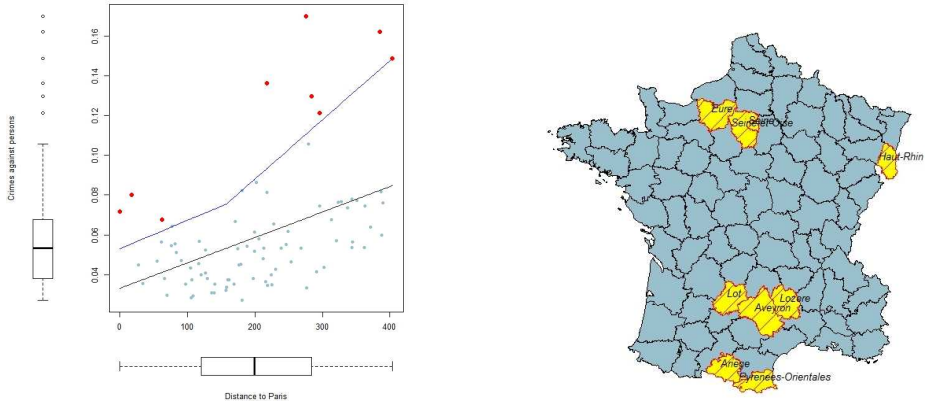

Figure 5: Example of use of the scattermap function. On the top, we represent the crimes against persons depending on the distance to Paris and on the bottom, the crimes against property depending on the Suicides.

of Suicides, prostitutes, illegitimate birth whereas crimes against persons are related to the distance to Paris. [2]

Bibliography

[1] Bivand, R.S., Pebesma, E.J. and Gomez-Rubio, V. (2008). *Applied spatial data analysis with R*. London, Springer-Verlag.
[2] Dray, S. and Jombard, T. (2009), Revisiting Guerry's data : introducing spatial constraints in multivariate analysis, *The Annals of Applied Statistics*, **5**, 2278-2299.
[3] Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C. (2012). Identification of local multivariate outliers, submitted.
[4] Friendly, M. (2007). A.-M. Guerry's moral statistics of France : challenges for multivariable, *Statistical Science*, **22**, 368-399.
[5] Laurent, T., Ruiz-Gazen, A. and Thomas-Agnan, G. (2012). GeoXp : an R package for exploratory spatial data analysis, *Journal of Statistical Software*, **47**.
[6] R Development Core Team (2011). *R : a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[2]This work was supported by the agence nationale de la recherche through the ModULand project (ANR-11-BSH1-005).