Quadtree anonymization of point data

Raymond Lagonigro, Ramon Oller, Joan Carles Martori

2023-07-13

Contents

1 Introduction

The AQuadtree package provides an automatic aggregation tool to anonymise point data. The framework proposed seeks the data accuracy at the smallest possible areas preventing individual information disclosure. Aggregation and local suppression of point data is performed using a methodology based on hierarchical geographic data structures. The final result is a varying size grid adapted to local area population densities described in Lagonigro, Oller, and Martori (2017).

The grid is created following the guidelines for grid datasets of the GEOSTAT project (GEOSTAT 2014) and the INSPIRE grid coding system is adopted as defined in the INSPIRE Data specifications (INSPIRE 2010). Geospatial specifications use the European Terrestrial Reference System 89, Lambert Azimuthal Equal Area (ETRS89-LAEA) projection (Annoni et al. 2003), although other Coordinate Reference Systems (CRS) and projections are also be used with the package. In the definition of the grid dataset, each cell is identified by a code composed of the cell's size and the coordinates of the lower left cell corner in the ETRS89-LAEA system. The cell's size is denoted in meters ("m") for cells' sizes up to 1000 meters, or kilometers ("km") for cells' sizes from 1000 meters and above. To reduce the length of the string, values for northing and easting are divided by 10n (where "n" is the number of zeros in the cell size value measured in meters).

> The cell code "1kmN2599E4695" identifies the 1km grid cell with coordinates of the lower left corner: Y=2599000m, X=4695000m.

The aggregation algorithm implemented in the package builds an initial regular grid of a given cell size, identifying each cell with the corresponding cell code. Each initial cell is recursively subdivided in quadrants where each new cell is assigned a second identifier containing a sequence of numbers to indicate the position of the cell in the disaggregation scheme. For instance, the sequence identifier corresponding to the right top cell in the right image in Figure 1 would be 416, i.e. fourth cell in the first division, and sixteenth cell in the second division.

To ensure data privacy, a cell is only split if all the resulting subdivisions satisfy the threshold restriction on the number of points. In cases of very irregular point pattern, this restriction results in less accuracy

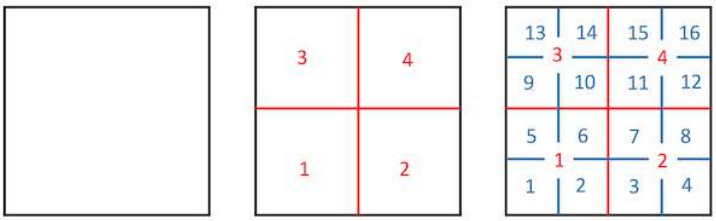
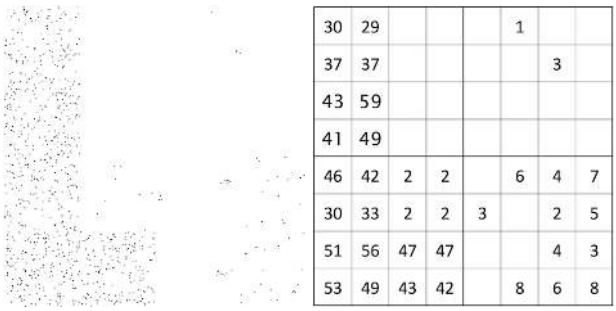


Figure 1: Three level quadtree splitting cell numbering example. Initial cell on the (left); first quadtree subdivision (center); second quadtree subdivision (right)

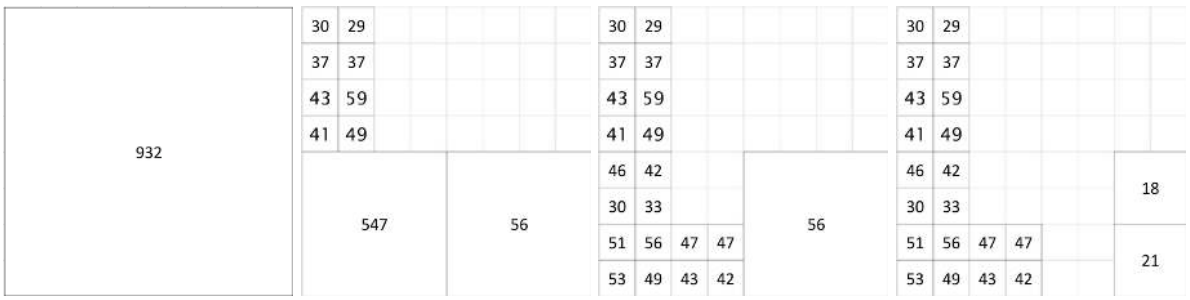
on the cell resolution. For instance, Figure 2a presents a pattern of 932 points unevenly distributed on a 1km cell and Figure 2b shows the corresponding grid of 62.5m cells with no threshold restrictions (the total number of points aggregated in each cell is shown).



(a) (b)

Figure 2: Set of spatial points (a) and the corresponding 62.5m grid with no threshold restrictions (b) (the numbers indicate the points aggregated in each cell).

If we define an anonymity threshold of 17, the cell in Figure 2a can not be subdivided because one of the four resulting quadrants contains only 4 points. The privacy mechanism aggregates all the points, as presented in Figure 3a, and covers an irregular spatial distribution. The AQuadtree algorithm contemplates the suppression of some points before continuing the disaggregation. For instance, suppressing the 4 points in the top right quadrant of Figure 2b results in the disaggregation shown in Figure 3b, which clearly is much more accurate to the underlying spatial distribution. Moreover, the elimination of more data points would lead to further disaggregation (Figure 3c and Figure 3d).



(a) (b) (c) (d)

Figure 3: Disaggregation examples with threshold value 17. No disaggregation and no loss (a); disaggregation with suppression of 4 points (b) ; more disaggregation with suppression of 12 points (c); maximum disaggregation with suppression of 29 points (d).

In order to balance information loss and resolution accuracy on the process of splitting a cell, the method computes the Theil inequality measure (Theil 1972) for the number of points in the possible quadrants as well as the percentage of points needed to be suppressed to force the division. In those cases where the anonymity threshold value prevents disaggregation, high values on the inequality measure may suggest the

need for further subdivision, while high values on the loss rate may suggest to stop this subdivision. The algorithm uses default limits for both measures: 0.25 and 0.4. respectively (both values can be defined between 0 and 1). Thus, if there exists any sub-cell with a number of points lower than the anonymity threshold and the inequality measure is higher than 0.25, then the disaggregation process continues by suppressing those points as long as the loss rate is lower than 0.4. Hence, following with example in Figure 2, the default disaggregation produced by the method would be the one shown in Figure 3b.

All the suppressed points during the process are aggregated in a cell with the initial dimension so their information does not disappear. This cell is marked as a residual cell. Following with the example in Figure 2, if the number of suppressed points overcome the anonymity threshold, as for instance, in Figure 3d, the 29 suppressed points are aggregated in a cell of the initial given dimension, which will be marked as a residual cell (see Figure 4).

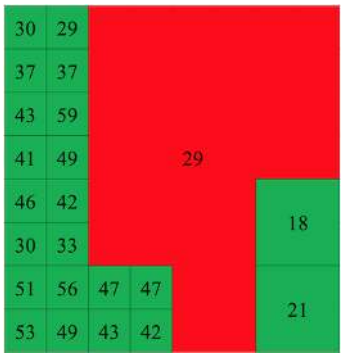


Figure 4: Example of a residual cell.

2 The AQuadtree Class

An AQuadtree class object is a spatial dataset representing a varying size grid and is created performing an aggregation of a given set of points considering a minimum threshold for the number of points in each cell. The AQuadtree main function of the package creates the AQuadtree object from *SpatialPoints* or *SpatialPointsDataFrame* objects.

```
example.QT <- AQuadtree(CharlestonPop)
class(example.QT)
## [1] "AQuadtree"
## attr(,"package")
## [1] "AQuadtree"
```

The AQuadtree class proposes a collection of methods to manage the generated objects and overrides the generic methods *show*, *print*, *summary* and *[* (subsetting) for the AQuadtree signature. The *plot* method overrides the generic function for plotting R objects with an extra parameter to specify if residual cells should be plotted. The *spplot* function overrides the lattice-based plot method from sp package (Pebesma and Bivand 2005), with two extra parameters to control if residual cells should be displayed, and wether attributes should be divided by the cell areas to make different zones comparable. The *merge* method merges data from an input data frame to the given AQuadtree object. An AQuadtree object can be coerced to a SpatialPolygonsDataFrame using the generic method *as* from methods package.

```
bcn.QT <- AQuadtree(BarcelonaPop)
plot(bcn.QT)
spplot(bcn.QT, by.density = TRUE)
```

2.1 Controlling the grid resolution

The characteristics of the AQuadtree object can be adjusted with various parameters. First, the *dim* parameter defines the size in meters of the highest scale cells and the *layers* parameter indicates the number of disaggregation levels. Thus, specifying the parameters *dim=10000* and *layers=4* would create

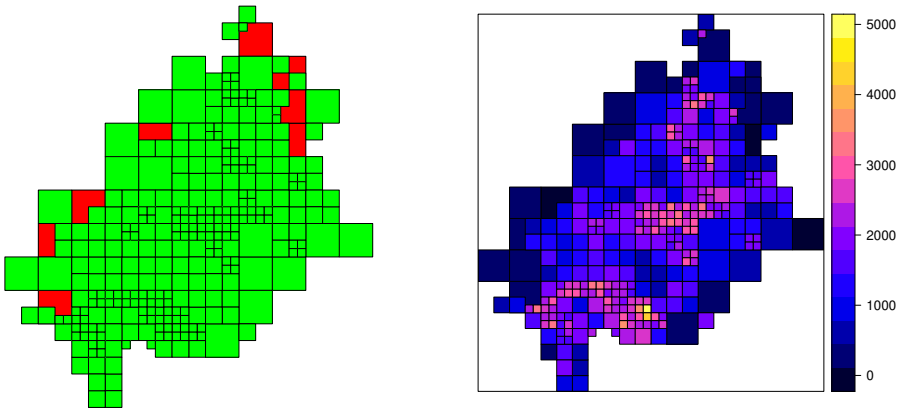


Figure 5: AQuadtree plot and spplot

a grid with cells of sizes between 10km and 1.25km. The default values establish an initial size of 1000 meters and 3 levels of disaggregation.

```
charleston.QT <- AQuadtree(CharlestonPop, dim = 10000, layers = 4)
summary(charleston.QT)
## Object of class "AQuadtree"
## 180 grid cells with sizes between 10km and 1.25km
## Coordinates:
##       min     max
## x 2060000 2160000
## y  110000  220000
## Is projected: TRUE
## proj4string:
## +proj=lcc +lat_0=36.6666666666667 +lon_0=-98.5 +lat_1=38.5666666666667
## +lat_2=37.2666666666667 +x_0=400000 +y_0=400000 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0
## +units=m +no_defs
## Initial Cell Size: 10km
## Number of valid grid Cells: 176
## Number of residual grid Cells: 4
## Data attributes:
##      total
##  Min.   : 100.0
##  1st Qu.: 157.8
##  Median : 226.5
##  Mean   : 292.2
##  3rd Qu.: 370.2
##  Max.   :2281.0
```

2.2 Summarizing data

The *colnames* parameter specifies the columns on the original dataset to summarize in the resulting grid. An extra attribute *total*, containing the number of points in each cell is automatically created and added to the dataframe. On the aggregation process, attributes specified in *colnames* parameter will be summarized using the *'sum'* function. A list of alternative summarizing functions can can be provided with the *funs* parameter. If any attribute indicated in the *colnames* parameter is a factor, the function creates a new attribute for each label of the factor. For instance, an attribute sex with two labels, *man* and *woman*, would be deployed into the two attributes *sex.man* and *sex.woman*.

```
class(BarcelonaPop$sex)
## [1] "factor"
levels(BarcelonaPop$sex)
## [1] "man"   "woman"
```

```
bcn.QT <- AQuadtree(BarcelonaPop, colnames = names(BarcelonaPop), funs = c("mean",
    "sum"))
summary(bcn.QT)
## Object of class "AQuadtree"
## 321 grid cells with sizes between 1km and 125m
## Coordinates:
##       min     max
## x 3659000 3670000
## y 2062500 2074500
## Is projected: TRUE
## proj4string:
## +proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80
## +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
## Initial Cell Size: 1km
## Number of valid grid Cells: 313
## Number of residual grid Cells: 8
## Data attributes:
##      total           age           sex.man        sex.woman
##  Min.   : 100   Min.   :35.28   Min.   : 40.0   Min.   : 44.0
##  1st Qu.: 139   1st Qu.:42.37   1st Qu.: 64.0   1st Qu.: 73.0
##  Median : 177   Median :44.42   Median : 83.0   Median : 95.0
##  Mean   : 248   Mean   :44.16   Mean   :117.1   Mean   :130.9
##  3rd Qu.: 328   3rd Qu.:46.18   3rd Qu.:158.0   3rd Qu.:170.0
##  Max.   :1288   Max.   :51.18   Max.   :626.0   Max.   :662.0
```

2.3 Specifying a threshold and threshold fields

The package applies a default anonymity threshold value of 100 and it can be changed with the *threshold* parameter. If nothing else is indicated, the threshold restriction is applied only to the total number of points aggregated in each cell (i.e. the *total* attribute added to the resulting dataset). When some of the attributes include confidential information, the threshold restriction can be applied to various properties with the *thresholdField* parameter, indicating the list of attributes from the resulting dataset that must satisfy that given threshold.

```
bcn.QT <- AQuadtree(BarcelonaPop, colnames = c("age", "sex"), funs = c("mean", "sum"),
    threshold = 17, thresholdField = c("sex.man", "sex.woman"))
summary(bcn.QT)
## Object of class "AQuadtree"
## 730 grid cells with sizes between 1km and 62.5m
## Coordinates:
##       min     max
## x 3659000 3670000
## y 2062000 2075000
## Is projected: TRUE
## proj4string:
## +proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80
## +towgs84=0,0,0,0,0,0,0 +units=m +no_defs
## Initial Cell Size: 1km
## Number of valid grid Cells: 713
## Number of residual grid Cells: 17
## Data attributes:
##      total           age           sex.man        sex.woman
##  Min.   : 34.0   Min.   :32.63   Min.   : 17.00   Min.   : 17.00
##  1st Qu.: 64.0   1st Qu.:41.52   1st Qu.: 30.25   1st Qu.: 33.00
##  Median :103.0   Median :43.87   Median : 49.00   Median : 54.00
##  Mean   :110.5   Mean   :43.71   Mean   : 52.25   Mean   : 58.21
##  3rd Qu.:140.0   3rd Qu.:46.08   3rd Qu.: 65.75   3rd Qu.: 74.00
##  Max.   :807.0   Max.   :53.46   Max.   :371.00   Max.   :436.00
```

2.4 Balancing information loss and accuracy

In order to control the disaggregation process, two more parameters set the thresholds on the inequity and loss rate. The extra parameter *ineq.threshold*, a rate between 0 and 1, specifies a threshold to force disaggregation when there is high inequality between sub-cells. The Theil entropy measure as computed in the *ineq* package (Zeileis, Kleiber, and Zeileis 2009) is used to measure inequality for each cell. The *ineq.threshold* parameter defaults to 0.25. Lower values in the *ineq.threshold* produce grids with smaller cells (see Figure 6).

```
bcn.QT <- AQuadtree(BarcelonaPop, threshold = 5, ineq.threshold = 0.01)
plot(bcn.QT)
bcn.QT <- AQuadtree(BarcelonaPop, threshold = 5, ineq.threshold = 0.5)
plot(bcn.QT)
```

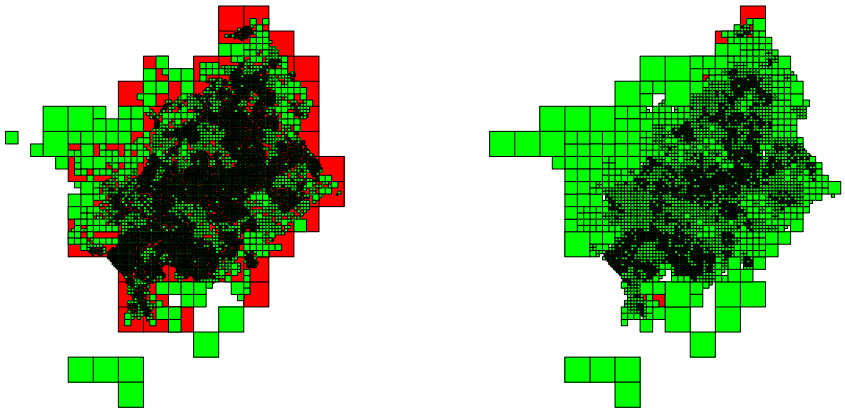


Figure 6: Examples of the effect of the ineq.threshold parameter.

On the other side, the parameter *loss.threshold*, also a rate between 0 and 1, indicates a rate of loss to prevent disaggregation of cells. A low value states that lower loss is preferred on the resulting grid so less disaggregation is obtained.

2.5 AQuadtree object structure

A call to the AQuadtree function will return an AQuadtree class object with six slots indicating the parameters used on the creation of the grid:

- *dim*: scale in meters of the highest level cells
- *layers*: number of subdivision levels
- *colnames*: attribute names summarized in the resulting grid
- *threshold*: the value used for anonymization
- *thresholdField*: attribute names to which the threshold restriction has been applied
- *loss*: number of points discarded during the process of disaggregation because of the threshold

```
bcn.QT <- AQuadtree(BarcelonaPop)
slotNames(bcn.QT)
##  [1] "dim"            "layers"         "colnames"       "threshold"
##  [5] "thresholdField" "loss"           "data"           "polygons"
##  [9] "plotOrder"      "bbox"           "proj4string"
```

The data slot contains a dataframe with the information comprised in each cell:

- *total*: number of points grouped in the cell.
- *level*: scale of disaggregation of the cell.
- *residual*: logical value indicating if the cell contains only residual points. Residual points are those that have been suppressed on the disaggregation process to get better accuracy, but can be grouped at the highest scale cell as it overcomes the given threshold.
- *cellCode*: cell's size and the coordinates of the lower left cell corner in the ETRS89-LAEA system at the highest aggregation level.

- *cellNum*: sequence of numbers indicating the position of the cell in the disaggregation scheme.

```
names(bcn.QT)
## [1] "cellCode" "cellNum"  "level"    "residual" "total"
head(bcn.QT)
## An object of class "AQuadtree" with 6 grid cells with sizes between 1km and 125m
##        cellCode cellNum level residual total
## 1 1kmN2064E3665             1    FALSE   313
## 2 1kmN2065E3666             1    FALSE   317
## 3 1kmN2066E3659             1    FALSE   109
## 4 1kmN2066E3660             1    FALSE   434
## 5 1kmN2066E3666             1    FALSE   919
## 6 1kmN2066E3667             1    FALSE   564
```

3 Provided data

The package includes two *SpatialPointsDataFrame* objects: *BarcelonaPop* for the city of Barcelona (Spain) and *CharlestonPop* for the Charleston, SC metropolitan area (USA). Both objects contain random point data with the distributions of real data acquired at census scale from different sources. The package also provides two *SpatialPolygons* objects with the spatial boundaries for each region. *BarcelonaCensusTracts* and *CharlestonCensusTracts* contain, respectively, the census tracts spatial limits for the city of Barcelona, and the census tracts spatial limits for the Charleston, SC metropolitan area.

BarcelonaPop comprises 81,359 sample points in the city of Barcelona, Spain. The original information was obtained from the statistics department of the Ajuntament de Barcelona, providing population data at the census tract level for the year 2018 (Ajuntament de Barcelona. Departament d'Estadística 2018). The points were generated and distributed randomly in space, maintaining unchanged the information at each census tract. To reduce the file size, only a sample of 7% of the points have been maintained.

```
data("BarcelonaPop", package = "AQuadtree")
summary(BarcelonaPop)
## Object of class SpatialPointsDataFrame
## Coordinates:
##       min     max
## x 3655447 3669871
## y 2059179 2074546
## Is projected: TRUE
## proj4string :
## [+proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80
## +towgs84=0,0,0,0,0,0,0 +units=m +no_defs]
## Number of points: 81359
## Data attributes:
##       age            sex
##  Min.   :  0.00   man  :38472
##  1st Qu.: 27.00   woman:42887
##  Median : 43.00
##  Mean   : 43.94
##  3rd Qu.: 61.00
##  Max.   :100.00
```

In a similar way, the *CharlestonPop* object, with 54,619 random sample points, was created using the information in the dataset Charleston1 from the 2000 Census Tract Data for the Charleston, SC metropolitan area (USA) (Geoda Data and Lab 2019). To reduce the file size, only a sample of 10% of the points have been maintained.

```
data("CharlestonPop", package = "AQuadtree")
summary(CharlestonPop)
## Object of class SpatialPointsDataFrame
## Coordinates:
```

```
##         min       max
## x 2047824.7 2187604.2
## y  102892.3  219129.3
## Is projected: TRUE
## proj4string :
## [+proj=lcc +lat_0=36.6666666666667 +lon_0=-98.5 +lat_1=38.5666666666667
## +lat_2=37.2666666666667 +x_0=400000 +y_0=400000 +ellps=GRS80
## +towgs84=0,0,0,0,0,0,0 +units=m +no_defs]
## Number of points: 54619
## Data attributes:
##        ID             origin          sex             age
##  56     : 1274   asian   :  761   male  :26801   under16:12629
##  22     : 1248   black   :16643   female:27818   16_65  :36314
##  59     : 1128   hisp    : 1333                  over65 : 5676
##  67     : 1076   multi_ra:  669
##  30     : 1032   white   :35213
##  70     : 1019
##  (Other):47842
```

4 Session info

Here is the output of session_info("AQuadtree") on the system on which this document was compiled:

```
devtools::session_info("AQuadtree")
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.3.0 (2023-04-21)
##  os       macOS Monterey 12.6.7
##  system   x86_64, darwin20
##  ui       X11
##  language (EN)
##  collate  C
##  ctype    en_US.UTF-8
##  tz       Europe/Berlin
##  date     2023-07-13
##  pandoc   2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (via rmarkd
##
## - Packages -------------------------------------------------------------------
##  package    * version date (UTC) lib source
##  AQuadtree  * 1.0.4   2023-07-13 [1] local
##  cli          3.6.1   2023-03-23 [3] CRAN (R 4.3.0)
##  dplyr      * 1.1.2   2023-04-20 [3] CRAN (R 4.3.0)
##  fansi        1.0.4   2023-01-22 [3] CRAN (R 4.3.0)
##  generics     0.1.3   2022-07-05 [3] CRAN (R 4.3.0)
##  glue         1.6.2   2022-02-24 [3] CRAN (R 4.3.0)
##  lattice      0.21-8  2023-04-05 [3] CRAN (R 4.3.0)
##  lifecycle    1.0.3   2022-10-07 [3] CRAN (R 4.3.0)
##  magrittr     2.0.3   2022-03-30 [3] CRAN (R 4.3.0)
##  pillar       1.9.0   2023-03-22 [3] CRAN (R 4.3.0)
##  pkgconfig    2.0.3   2019-09-22 [3] CRAN (R 4.3.0)
##  R6           2.5.1   2021-08-19 [3] CRAN (R 4.3.0)
##  rlang        1.1.1   2023-04-28 [3] CRAN (R 4.3.0)
##  sp         * 2.0-0   2023-06-22 [3] CRAN (R 4.3.0)
##  tibble       3.2.1   2023-03-20 [3] CRAN (R 4.3.0)
##  tidyselect   1.2.0   2022-10-10 [3] CRAN (R 4.3.0)
##  utf8         1.2.3   2023-01-31 [3] CRAN (R 4.3.0)
##  vctrs        0.6.2   2023-04-19 [3] CRAN (R 4.3.0)
```

```
##  withr        2.5.0   2022-03-03 [3] CRAN (R 4.3.0)
##
##  [1] /private/var/folders/cf/7cn764jj39x3yyjcpd863ml00000gn/T/Rtmp7UwgGM/Rinst16a865263b1f
##  [2] /private/var/folders/cf/7cn764jj39x3yyjcpd863ml00000gn/T/Rtmp9WtbdT/temp_libpath16842bd6e52e
##  [3] /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/library
##
## ------------------------------------------------------------------------------
```

References

Ajuntament de Barcelona. Departament d'Estadística. 2018. "Població segons padró d'habitants." https://www.bcn.cat/estadistica/catala/dades/tpob/pad/padro/a2018/.

Annoni, Alessandro, Claude Luzet, Erich Gubler, and Johannes Ihde. 2003. "Map Projections for Europe." Ispra, Italy: EUR 20120 EN. European Comission. Joint Research Centre. https://ec.europa.eu/eurostat/documents/4311134/4366152/Map-projections-EUROPE.pdf.

Geoda Data and Lab. 2019. "Sample data referenced in the tutorials for GeoDa, GeoDaSpace, and CAST." Illinois: Center for Spatial Data Science. University of Chicago. https://geodacenter.github.io/data-and-lab/.

GEOSTAT. 2014. "ESSnet project GEOSTAT 1B – Representing 2011 Census data on grid." The European Forum for GeoStatistics. https://www.efgs.info/geostat/1b/.

INSPIRE. 2010. "INSPIRE Specification on Geographical Grid Systems – Guidelines (D2.8.I.2)." March. INSPIRE Infrastructure for Spatial Information in Europe: European Commission. https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_Specification_GGS_v3.0.1.pdf.

Lagonigro, Raymond, Ramon Oller, and Joan Carles Martori. 2017. "A quadtree approach based on European geographic grids: Reconciling data privacy and accuracy." *SORT* 41 (1).

Pebesma, E, and RS Bivand. 2005. "S classes and methods for spatial data: the sp package." *R News*.

Theil, Henry. 1972. "Statistical decomposition analysis." Amsterdam: North Holland.

Zeileis, Achim, Christian Kleiber, and Maintainer Achim Zeileis. 2009. "Package 'ineq'." *Tech. Rep.*