Class intervals for thematic mapping: implementations in R

European Colloquium on Theoretical and Quantitative Geography 2023, Braga (Portugal)

Roger Bivand

15 September 2023

Introduction

- We will review briefly the background in applied cartography for using class intervals,
- and describe the implementation of methods available in functionality available in R, in particular in `classInt` and associated packages.
- Materials available on https://github.com/rsbivand/eqc23_talk, https://rsbivand.github.io/eqc23_talk/

Classed or unclassed

- From Dickinson (1973) to Tyler (2010), via Slocum et al. (2005), key handbooks on statistical mapping and thematic cartography present similar lists of ways of creating class intervals.
- These authorities stress the need for the creator of the thematic map to consider the message(s) being conveyed.
- One of the choices facing us is whether to use class intervals or unclassed maps, representing the variability of the values being mapped by pseudo-continuous variations in shading or colour (Tobler 1973), or rather to choose class intervals.
- The colour rendering technology of the output device will determine the discrete colur count available in the pseudo-continuous case.

Class intervals

- When class intervals are chosen, the legend constitutes a look-up table, from which the observer can readily read off which shading or colour corresponds to which value interval.
- Such quantising of course transforms interval or ratio-level data to the ordinal level, while retaining interval or ratio-level labels, as in a histogram
- For nominal or ordinal data, the starting point for class intervals may be given by the definition of the discrete variables being mapped, but here again input categories may need to be merged to reach a sensible number of classes.

Cartograms

- Neither choice avoids the basic challenge of the stronger visual impact of irregular polygonal observations with larger surface areas, but often administrative data is only available for such aggregates.
- When gridded data is available, or when cartograms are presented in place of maps of polygon boundaries, this challenge may be addressed but not eliminated.
- Because these adaptations also need to handle the visual representation of observed variables, they constitute a distinctly different problem, and will not be considered here.

Border lines

- Similarly, while the display of boundary lines between the regular or irregular polygon borders that show the support of the data being visualized is an important question, it will not be considered here.
- We just note that without some even minimal border line, neighbouring observations with the same class (or value for unclassed maps) will be perceived as a single entity.

Multivariate graphics and micromaps

- Class intervals also matter in multivariate graphics, for example with a conditioning variable (Perpiñán Lamigueiro 2018; Sarkar 2008; Cleveland 1993); in this context `shingles` are class intervals that overlap
- Wilkinson (2006) proposed a grammar of graphics, implemented by Wickham (2016) in `ggplot2` and for mapping in `tmap` (Tennekes 2018)
- These extensions are also taken up in `micromap` (Payton et al. 2015), and described by Carr, Pickle, and micromapST Author Team (2010), and implemented in `micromapST` (Pickle, Pearson, and Carr 2015)
- These aspects will not be considered here

Applied thematic cartography

Applied thematic cartography

- The authorities cited in the introduction (Dickinson 1973; Slocum et al. 2005; Tyler 2010) write for users of thematic cartography rather than for those studying the underlying principles
- Naturally, the underlying perceptual and semiotic principles are important, but are not central to this discussion in applied thematic cartography
- Many users simply go with the default, and perhaps resent being asked to choose how to construct class intervals, despite there being no obvious automatic route
- Software providers like ESRI provide some helpful guidance: https://pro.arcgis.com/en/pro-app/latest/help/mapping/layer-properties/data-classification-methods.htm, referring on to: https://www.spatialanalysisonline.com/HTML/classification_and_clustering.htm

Class intervals in applied thematic cartography

- As part of an online course: https://magrit-formations.github.io/, a useful distinction is drawn in https://magrit-formations.github.io/discretisation
- This is based on the univariate classification of variables with many unique values, as clearly discrete variables with few categories do not require further discretisation (unless categories need grouping).
- The important distinction is between symmetrical and skewed distributions, where variables displaying uniform and symmetrical distributions can use equal-width intervals, and symmetrical distributions can also use intevals based on multiples of standard deviations from the mean; unclassed maps may work best for uniform distributions
- Skewed and multi-modal distributions should preferably not use these kinds of intervals, but rather quantiles (equal-count) or observed (natural) thresholds, and skewed distributions can also use geometric progression

Example data set: Lot Departément, France

- Giraud and Pecout (2023) use a data set for the Lot Departément in south-western France stored here as a GeoPackage file, with a number of layers.
- We will use the 313 commune layer with administrative boundaries from 2020, and population and economic activity data also from 2020.
- This is a vector example data set; it would be interesting in parallel to consider a raster data set, but here we use only the vector representation.

Lot Departément geopackage

We'll use `sf` to read the Lot GeoPackage file:

```
com <- sf::st_read("data/lot46.gpkg", layer = "commune")
```

```
## Reading layer `commune' from data source
##   `/home/rsb/presentations/eqc23_talk/data/lot46.gpkg'
##   using driver `GPKG'
## Simple feature collection with 313 features and 12 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 539668.5 ymin: 6346290 xmax: 637380.9 ymax: 6439668
## Projected CRS: RGF93 v1 / Lambert-93
```

and subset labels for the largest observations by population:

```
bigcom <- com[com$POPULATION > 2000,]
bigcom <- bigcom[rev(order(bigcom$NOM_COM)),]
```

Variables: population density and economically active population share

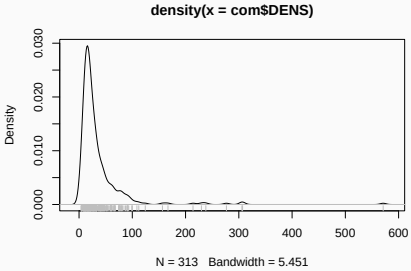
Population density is taken as the population count divided by observation area in square km; percentage economically active of total population by summing women and men in four sectors:

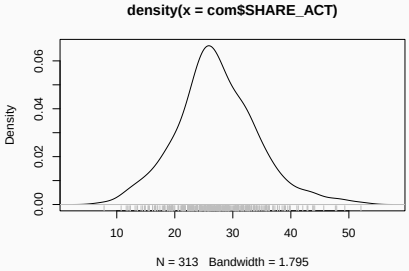
```
(com$POPULATION / (sf::st_area(com) |> units::set_units(km^2))) |>
    as.numeric() -> com$DENS
vars <- c("AGR_H", "AGR_F", "IND_H", "IND_F", "BTP_H", "BTP_F", "TER_H", "TER_F")
com |> subset(select=vars, drop=TRUE) |> apply(1, sum) |>
    unname() -> com$POP_ACT
com$SHARE_ACT <- 100 * com$POP_ACT / com$POPULATION
```

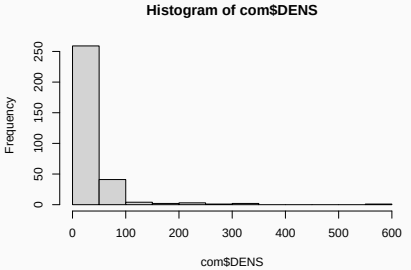
A Dougenik cartogram is also constructed:

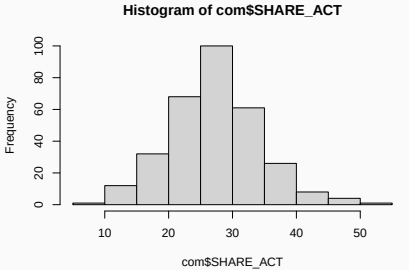
```
ccom <- cartogram::cartogram_cont(com, weight="POPULATION", prepare = "none")
```

DENS and SHARE_ACT histograms and density plots









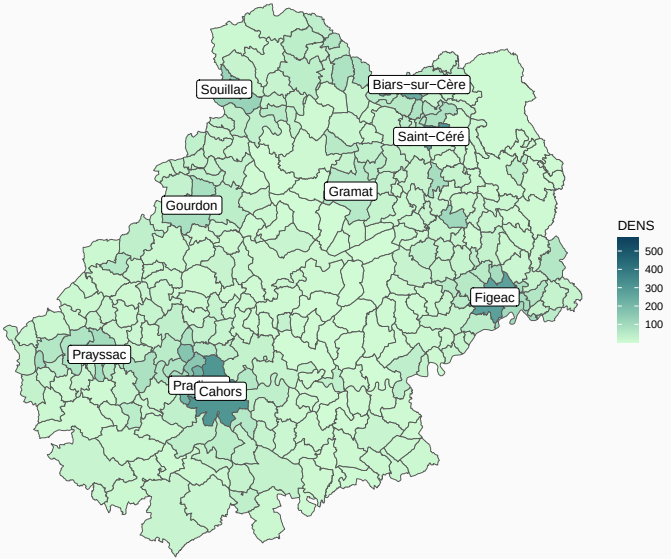
DENS and SHARE_ACT distributions

We can try to distinguish between symmetrical and skewed distibutions, but most tests for diverenge from the Normal lose power when the distribution is not Normal, DENS is obviously skewed, SHARE_ACT is symmetrical:

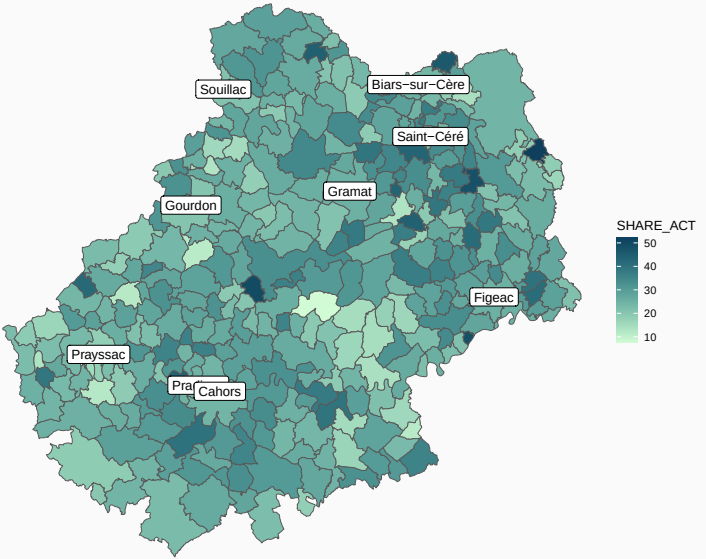
```
c(DENS=moments::agostino.test(com$DENS)$p.value < 0.005,
    SHARE_ACT=moments::agostino.test(com$SHARE_ACT)$p.value < 0.005)
```

```
##      DENS SHARE_ACT
##      TRUE     FALSE
```

No class intervals, population density map, Lot, 2020



No class intervals, economically active population share map, Lot, 2020



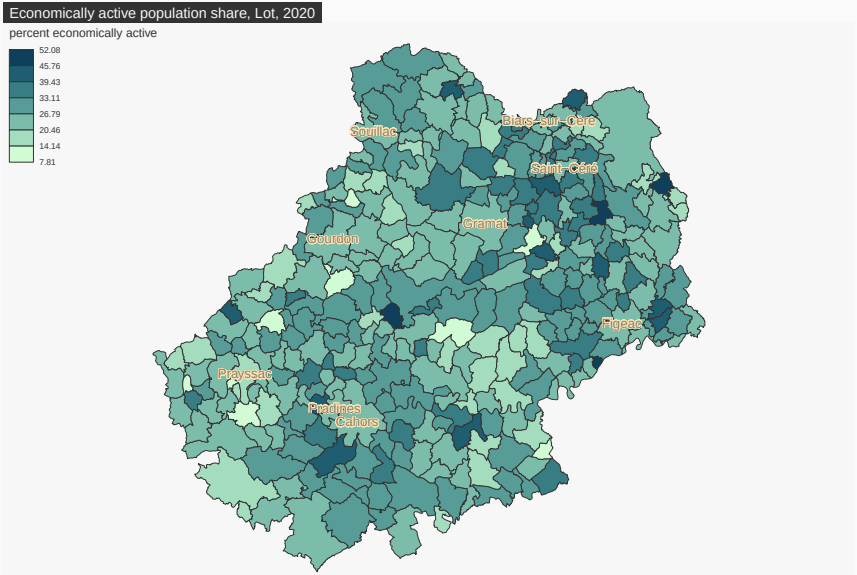
Kinds (styles) of class intervals

- Books like Tyler (2010) offer similar lists of kinds (styles) of class intervals driven by data rather than fixed in advance by the user
- Fixed class intervals may for example represent regulatory or customary limits, so the output maps shows where the chosen limits were exceeded
- Where the data diverge from a fixed point, often zero in for example regression residuals, choosing ways of representing the fixed point in otherwise data-driven interval construction may matter
- Both interval closure (left or right) and the data rounding/precision used in creating breaks do matter; more rounding may improve legend legibility

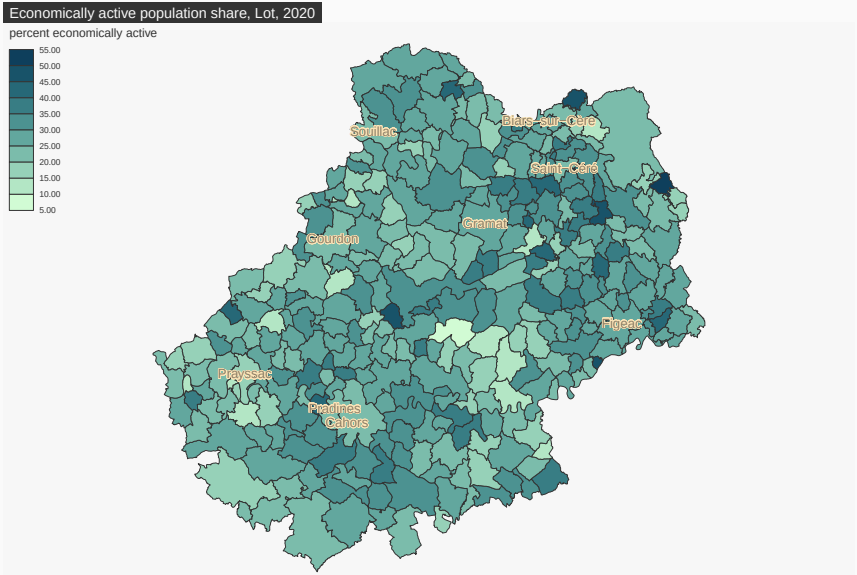
Equal-width class intervals

- Equal-width (equal-step, `equal`) intervals take the range of the data and cut into k intervals; the range may be buffered out a little; the break points may not be *pretty* numbers
- `pretty` intervals may not yield exactly k intervals, but instead of unrounded width values in `equal` intervals, the steps/intervals are `pretty`, 1, 2 or 5 times a power of 10; this approach was initially used for positioning and labelling ticks on scatterplot axes (Becker and Chambers 1984)
- Both of these styles are data-driven but neither are compute-intensive, so are also suited to larger data sets; however, they should not be used with data other than those with uniform or symmetric distributions

Equal-width class intervals, economically active population share map, Lot, 2020



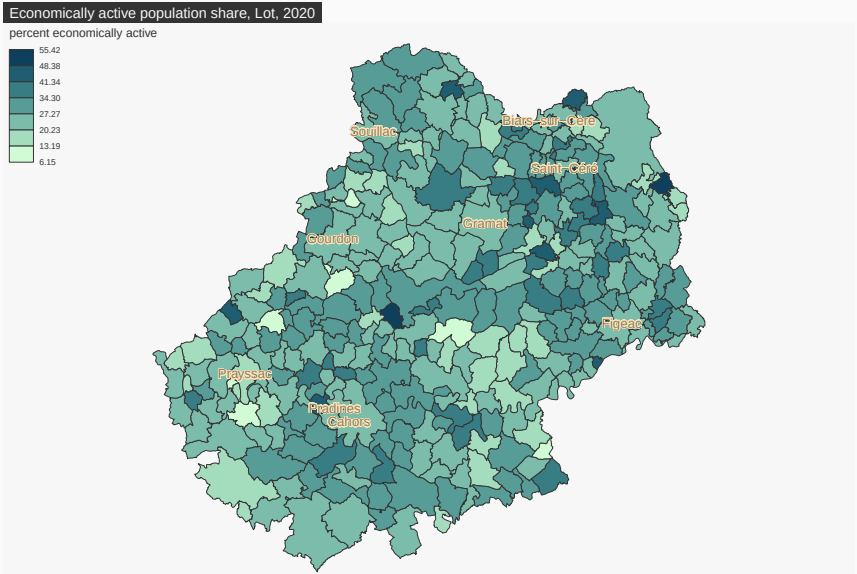
“Pretty” class intervals, economically active population share map, Lot, 2020



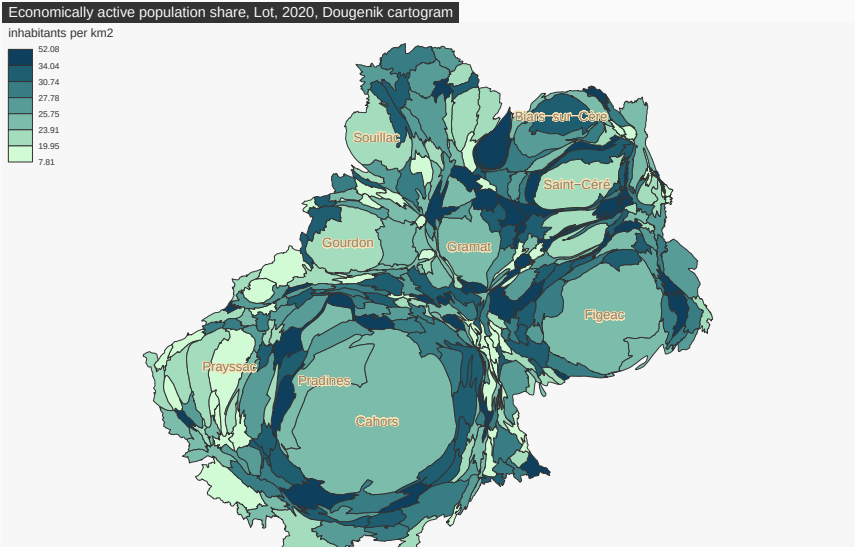
Standard-deviation class intervals

- If the data are roughly symmetric, the class intervals may be constructed to cover the data range by $\bar{x} \pm m\hat{\sigma}$ where $\bar{x}$ is the sample mean, $\hat{\sigma}$ the standard deviation, and m a sequence of intervals around zero
- It is possible for example to set m as `c(-Inf, -2, -1, 0, 1, 2, +Inf)`, or `c(-Inf, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, +Inf)`
- The `sd` style in `classInt::classIntervals` uses `pretty` run on the centred and scaled data to set m by default
- This style is data-driven but not compute-intensive, so also suited to larger data sets; however, it should not be used with data other than those with symmetric distributions

Standard-deviation class intervals, economically active population share map, Lot, 2020



Standard-deviation class intervals, economically active population share cartogram by population, 2020

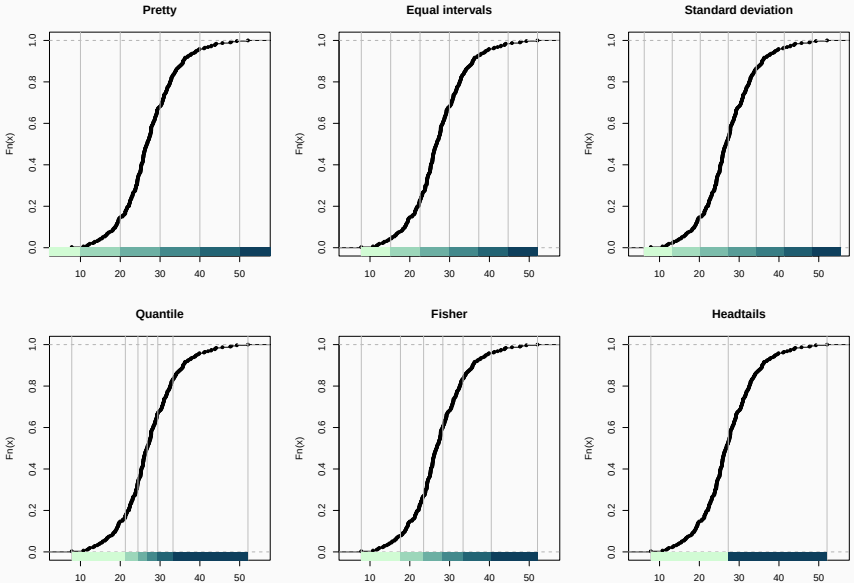


Comparing class intervals (symmetrical)

Table: Standard deviation by Pretty

sd		1	2	3	4	5	6	All
1	N	1	6	0	0	0	0	7
2	N	0	38	1	0	0	0	39
3	N	0	0	119	0	0	0	119
4	N	0	0	48	57	0	0	105
5	N	0	0	0	30	2	0	32
6	N	0	0	0	0	9	0	9
7	N	0	0	0	0	1	1	2
All	N	1	44	168	87	12	1	313

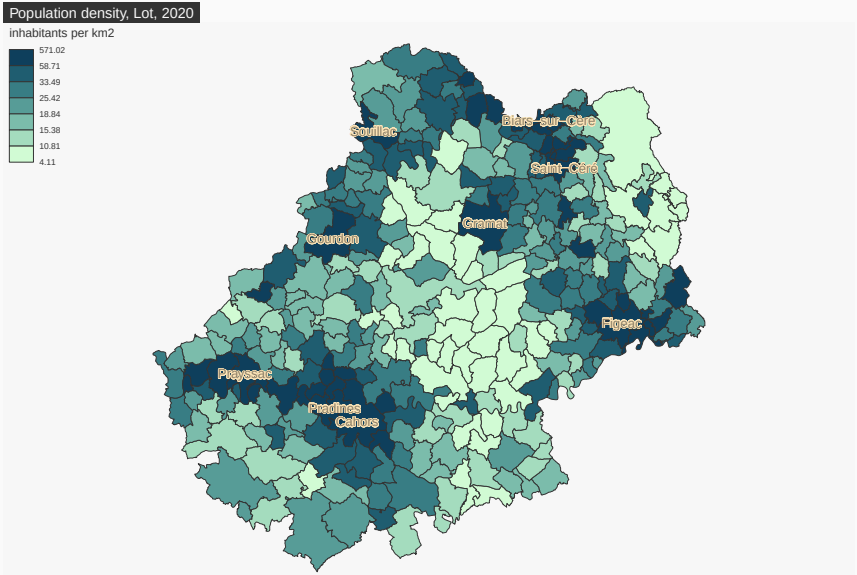
Comparing class intervals, empirical CDF (economically active share, symmetrical)



Quantile class intervals

- Quantiles cut the data into classes with equal counts as far as possible, and may be used with all kinds of data
- It is worth noting that quantiles may be calculated in many ways; some earlier algorithms have been superceded (Hyndman and Fan 1996), see also https://en.wikipedia.org/wiki/Quantile
- `classInt::classIntervals` uses default `type=7`, but `type=8` might be preferable; this can be passed through in `classInt::classIntervals` and `tmap::tm_fill`, but not in `mapsf::mf_map` or `sf::plot.sf`

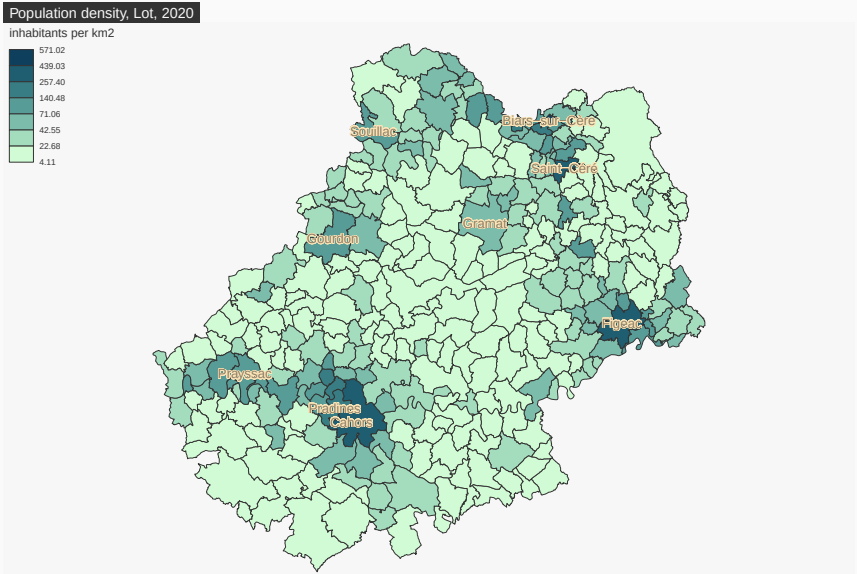
Quantile class intervals, population density, Lot, 2020



Observed threshold/natural breaks class intervals

- Jenks and Caspall (1971) `jenks` and Fisher (1958) `fisher` are legacy algorithms for data-driven univariate clustering, and are relatively compute-intensive; if there are many observations, a random sample may be used to calculate the breaks anchored with `set.seed`
- `stats::hclust` provides several hierarchical clustering methods that work in a univariate setting, defaulting to complete clustering, but requires a dense distance matrix so is not suitable for large data sets
- `stats::kmeans` can be used in a univariate setting, anchored with `set.seed` (but could be extended for larger data sets with `cluster::clara` which takes samples internally)
- Bagged clustering may also be used `e1071::bclust`, combining `kmeans` and hierarchical clustering (Leisch 1999; Dolnicar and Leisch 2003, 2004)

Fisher class intervals, population density, Lot, 2020

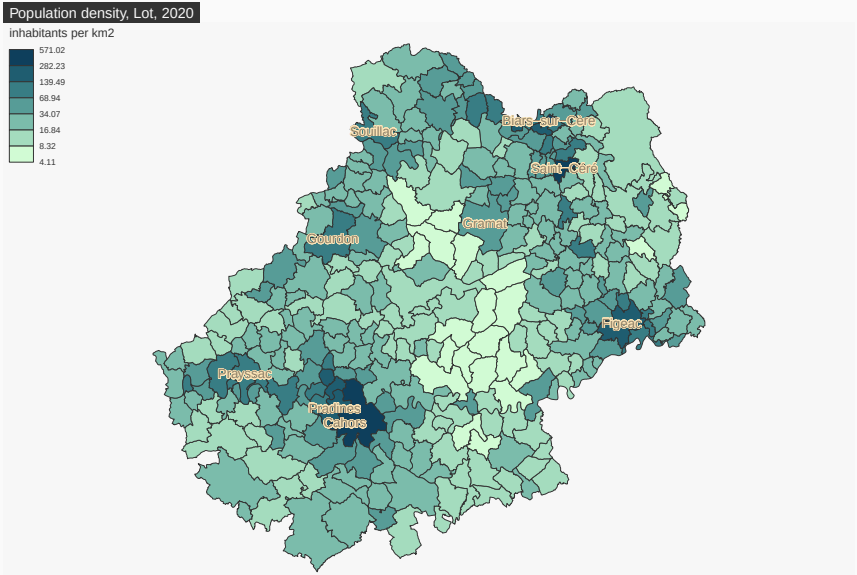


Fisher class

Geometric progression and other class intervals

- The `geom` style in `mapsf` is based on a geometric progression, all variable values must be strictly greater than zero.
- `headtails` based on Jiang (2013) has been contributed by Diego Hernangómez
- `box` was also contributed by Diego Hernangómez using the boxplot as a template
- `maximum` was contributed by Josiah Parry

Geometric progression class intervals, population density, Lot, 2020

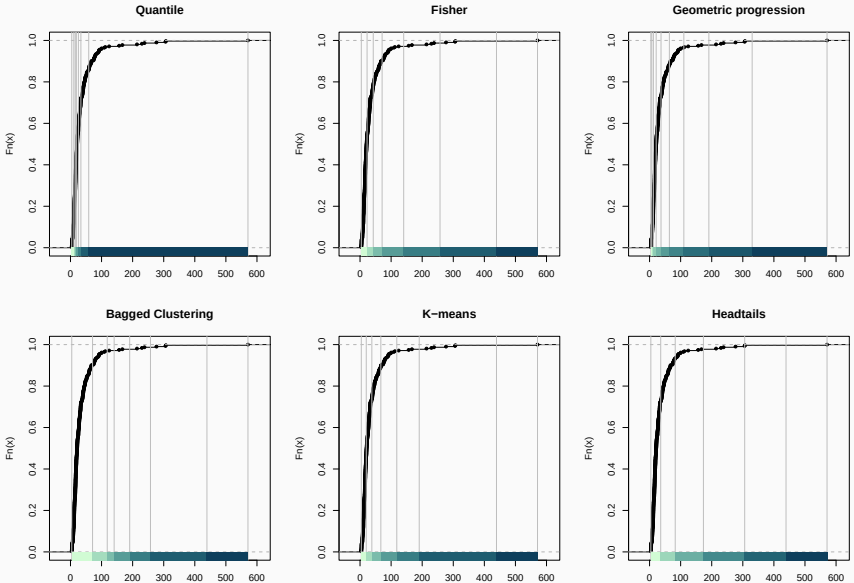


Comparing class intervals (skewed) ii

Table: Fisher by Geometric progression

fisher		1	2	3	4	5	6	7	8	9	All
1	N	14	49	93	12	0	0	0	0	0	168
2	N	0	0	0	61	15	0	0	0	0	76
3	N	0	0	0	0	31	7	0	0	0	38
4	N	0	0	0	0	0	20	2	0	0	22
5	N	0	0	0	0	0	0	2	3	0	5
6	N	0	0	0	0	0	0	0	3	0	3
7	N	0	0	0	0	0	0	0	0	1	1
All	N	14	49	93	73	46	27	4	6	1	313

Comparing class intervals, empirical CDF (population density, skewed)



Comparing class intervals (skewed) iii

Table: Fisher by K-means

fisher		1	2	3	4	5	6	7	All
1	N	145	23	0	0	0	0	0	168
2	N	0	62	14	0	0	0	0	76
3	N	0	0	35	3	0	0	0	38
4	N	0	0	0	21	1	0	0	22
5	N	0	0	0	0	2	3	0	5
6	N	0	0	0	0	0	3	0	3
7	N	0	0	0	0	0	0	1	1
All	N	145	85	49	24	3	6	1	313

Concluding remarks

How to document available choices?

- A useful discussion around a github issue raised about the R `classInt` package (https://github.com/r-spatial/classInt/issues/41) highlighted the desirability of deeper reflection about the role played by class intervals in contemporary applied thematic cartography.
- In Pebesma and Bivand (2023) (https://r-spatial.org/book/08-Plotting.html#sec-classintervals), only a brief paragraph is devoted to this topic, despite the use of the `classInt::classIntervals` function in thematic mapping in the sf, stars, tmap, mapsf, and other R packages.

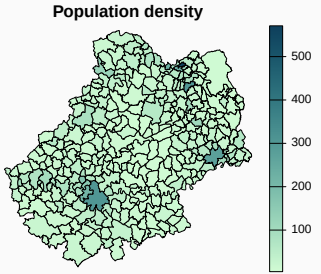
The `sf`, `stars` and `mapsf` packages

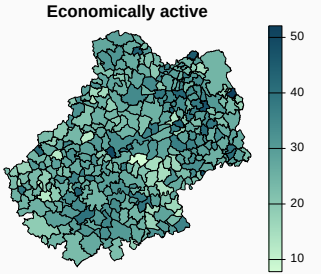
- The sf package documentation provides figures showing consequences of different choices: https://r-spatial.github.io/sf/articles/sf5.html#class-intervals (using `pretty` as default), supplemented for stars by https://r-spatial.github.io/stars/reference/plot.html (using `quantile` as default).
- The mapsf package has more detailed documentation, including https://riatelab.github.io/mapsf/articles/mapsf.html#choropleth-map and `mapsf::mf_get_breaks` extending `classInt::classIntervals` (https://riatelab.github.io/mapsf/reference/mf_get_breaks.html); `mapsf::mf_map` uses `quantile` as default for choropleth maps
- A work-in-progress bookdown book is available at https://rcarto.github.io/cartographie_avec_r/.

The `tmap` package

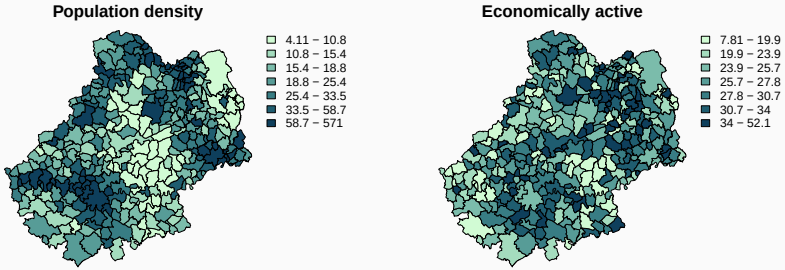
- The tmap package is evolving towards version 4, which is to be documented in a work-in-progress bookdown book, with class intervals presented in this section: https://r-tmap.github.io/tmap-book/visual-variables.html#color-scale-styles.
- Earlier versions of tmap are covered by Tennekes (2018), and a section in the second edition of the bookdown book https://r.geocompx.org/adv-map.html#color-settings (Lovelace, Nowosad, and Muenchow 2019); `tmap::tm_fill` uses `pretty` as default for choropleth maps

The `terra` package - continuous

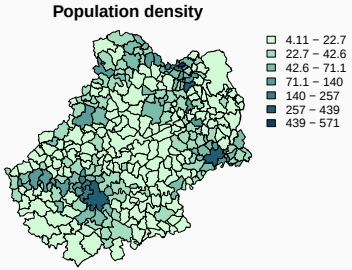


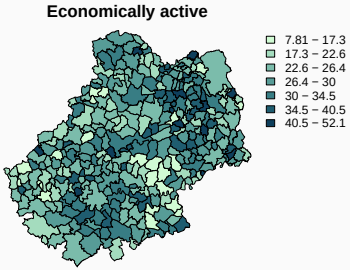


The terra package - quantile

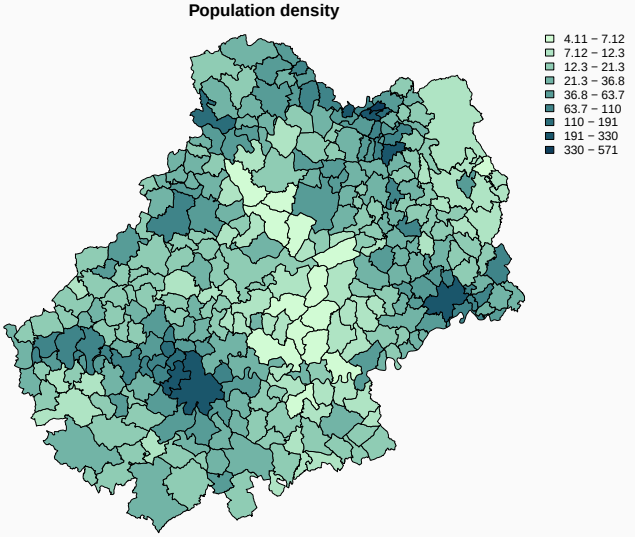


The terra package - Fisher





The `terra` package - geometric progression



Extensions

- Similarly, the Python module `mapclassify` is a helper in the background rather than being offered the attention it arguably deserves.
- While many users will be more familiar with graphical user interfaces for choosing how to construct class intervals, programmatic interfaces reveal a good deal of what is happening when choices are made.

sessionInfo i

```
sessionInfo()

## R version 4.3.1 (2023-06-16)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Fedora Linux 38 (Workstation Edition)
##
## Matrix products: default
## BLAS:   /home/rsb/topics/R/R431-share/lib64/R/lib/libRblas.so
## LAPACK: /home/rsb/topics/R/R431-share/lib64/R/lib/libRlapack.so;  LAPACK version 3.11.0
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Lisbon
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods
```

sessionInfo ii

```
## [7] base
##
## other attached packages:
## [1] modelsummary_1.4.2 classInt_0.4-10    mapsf_0.7.1
## [4] ggplot2_3.4.3
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.3          generics_0.1.3      xml2_1.3.5
##  [4] class_7.3-22        KernSmooth_2.23-22 stringi_1.7.12
##  [7] digest_0.6.33       magrittr_2.0.3      evaluate_0.21
## [10] grid_4.3.1          fastmap_1.1.1       backports_1.4.1
## [13] e1071_1.7-13        DBI_1.1.3           packcircles_0.3.6
## [16] httr_1.4.7          rvest_1.0.3         fansi_1.0.4
## [19] viridisLite_0.4.2   cartogram_0.3.0     scales_1.2.1
## [22] codetools_0.2-19    cli_3.6.1           rlang_1.1.1
## [25] units_0.8-4         munsell_0.5.0       withr_2.5.0
## [28] yaml_2.3.7          tools_4.3.1         checkmate_2.2.0
## [31] dplyr_1.1.3         colorspace_2.1-0    DT_0.29
## [34] webshot_0.5.5       kableExtra_1.3.4    vctrs_0.6.3
## [37] R6_2.5.1            proxy_0.4-27        lifecycle_1.0.3
## [40] stringr_1.5.0       htmlwidgets_1.6.2   insight_0.19.5
## [43] pkgconfig_2.0.3     terra_1.7-48        pillar_1.9.0
## [46] gtable_0.3.4        glue_1.6.2          moments_0.14.1
```

sessionInfo iii

```
## [49] Rcpp_1.0.11        systemfonts_1.0.4  sf_1.0-14
## [52] xfun_0.40          tibble_3.2.1       tidyselect_1.2.0
## [55] rstudioapi_0.15.0  knitr_1.44         farver_2.1.1
## [58] htmltools_0.5.6    tables_0.9.17      svglite_2.1.1
## [61] rmarkdown_2.24     labeling_0.4.3     compiler_4.3.1
```

Aftermatter

References i

Becker, R. A., and J. M. Chambers. 1984. *S: An Interactive Environment for Data Analysis and Graphics*. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole.

Carr, Daniel B., Linda Williams Pickle, and micromapST Author Team. 2010. *Visualizing Data Patterns with Micromaps*. Boca Raton, FL: CRC Press.

Cleveland, W. S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Dickinson, G. C. 1973. *Statistical Mapping and the Presentation of Statistics*. London: Edward Arnold.

Dolnicar, Sara, and Friedrich Leisch. 2003. “Winter Tourist Segments in Austria: Identifying Stable Vacation Styles Using Bagged Clustering Techniques.” *Journal of Travel Research* 41 (3): 281–92. https://doi.org/10.1177/0047287502239037.

———. 2004. “Segmenting Markets by Bagged Clustering.” *Australasian Marketing Journal* 12 (1): 51–65. https://doi.org/10.1016/S1441-3582(04)70088-9.

Fisher, Walter D. 1958. “On Grouping for Maximum Homogeneity.” *Journal of the American Statistical Association* 53 (284): 789–98. https://doi.org/10.1080/01621459.1958.10501479.

Giraud, Timothée, and Hugues Pecout. 2023. “Cartographie avec R.” https://doi.org/10.5281/zenodo.7528161.

Hyndman, Rob J., and Yanan Fan. 1996. “Sample Quantiles in Statistical Packages.” *The American Statistician* 50 (4): 361–65. https://doi.org/10.1080/00031305.1996.10473566.

Jenks, George F., and Fred C. Caspall. 1971. “Error on Choroplethic Maps: Definition, Measurement, Reduction.” *Annals of the Association of American Geographers* 61 (2): 217–44. https://doi.org/10.1111/j.1467-8306.1971.tb00779.x.

Jiang, Bin. 2013. “Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-Tailed Distribution.” *The Professional Geographer* 65 (3): 482–94. https://doi.org/10.1080/00330124.2012.700499.

References iii

Leisch, Friedrich. 1999. “Bagged Clustering.” 51. Working Papers SFB ”Adaptive Information Systems and Modelling in Economics and Management Science”.

Lovelace, Robin, Jakub Nowosad, and Jannes Muenchow. 2019. *Geocomputation with R*. Chapman & Hall/CRC. https://r.geocompx.org/.

Payton, Quinn, Tony Olsen, Marc Weber, Michael McManus, Tom Kincaid, and Marcus W. Beck. 2015. “micromap: A Package for Linked Micromaps.” *Journal of Statistical Software* 63 (2): 1–16. https://doi.org/10.18637/jss.v063.i02.

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science with Applications in R*. Chapman & Hall/CRC. https://www.routledge.com/Spatial-Data-Science-With-Applications-in-R/Pebesma-Bivand/p/book/9781138311183.

Perpiñán Lamigueiro, Oscar. 2018. *Displaying Time Series, Spatial, and Space-Time Data with R, Second Edition*. Boca Raton, FL: CRC Press.

Pickle, Linda Williams, James B. Pearson, and Daniel B. Carr. 2015. “micromapST: Exploring and Communicating Geospatial Patterns in US State Data.” *Journal of Statistical Software* 63 (3): 1–25. https://doi.org/10.18637/jss.v063.i03.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with r*. New York: Springer.

Slocum, Terry A., Robert B. McMaster, Fritz C. Kessler, and Hugh H. Howard. 2005. *Thematic Cartography and Geographic Visualization*. Upper Saddle River NJ: Prentice Hall.

Tennekes, Martijn. 2018. “tmap: Thematic Maps in R.” *Journal of Statistical Software* 84 (6): 1–39. https://doi.org/10.18637/jss.v084.i06.

Tobler, W. R. 1973. “Choropleth Maps Without Class Intervals?” *Geographical Analysis* 5 (3): 262–65. https://doi.org/10.1111/j.1538-4632.1973.tb01012.x.

Tyler, Judith A. 2010. *Principles of Map Design*. New York: Guildford.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wilkinson, Leland. 2006. *The Grammar of Graphics*. Springer Science & Business Media.