Reproducible Workflow for Cartography - Migrants Deaths in the Mediterranean

Timothée Giraud[a]*, Nicolas Lambert[b]

[a] *UMS RIATE, Paris Diderot University, CNRS, France, timothee.giraud@cnrs.fr*
[b] *UMS RIATE, Paris Diderot University, CNRS, France, nicolas.lambert@cnrs.fr*

* Corresponding Author

Abstract: As any scientific production, maps must be disputed and debated. The implementation of reproducible processes based on free software and open data is essential. In this paper, we demonstrate that this objective can be achieved in the R software ecosystem. In our demonstration, we propose a set of cartographic visualizations based on the example of dead and missing migrants in the Mediterranean over the period 2014-2018. Each representation focuses on one aspect of the phenomenon and the R code used is available. We argue that this multi-visualization contributes to bring new knowledge on the migration debate at European borders and aims at illustrating its geographical complexity.

Keywords: Cartography, GeoVizualisation, Workflow, Reproducible research, Open Science, Migrations, Europe, Mediterranean

1. Introduction

”Science is infallible; but scientists are always wrong.” (France, 1928). The validity of scientific studies can be assessed by their reproducibility. Maps, as part of scientific production , must be reproducible. In the academic context, data visualization allows both exploratory analysis to generate new ideas and graphical evidences to confirm hypotheses. Maps are an efficient way to communicate and explain research findings. They are prominent parts of studies and there is a strong need for tracing the different steps taken to design them.

Most of the time, several software packages are required to cover the entire processing chain from data gathering to graphical representation. The stacking of multiple software packages implies a variety of datafile formats, the creation of many intermediate files. Hence, the reproducibility of the produced maps is difficult, it requires to possess and master each software packages involved. The use of non open source solutions could aslo be a serious impediment to reproducibility: statistical and graphical methods are hiden inside ”black boxes” that are not available to study nor modification.

We propose in this paper a practical example addressing these different issues using open data, open-source statistics and graphics language (R), literate programing and version control system. The aim of this scheme is to be able to produce maps and geographical analyses in a collaborative approach that is both free and open-source, fully reproducible and ready for scientific interactions and discussions. In this contribution, we apply this technical framework to the example of death of migrants in the Mediterranean over the period 2014-2018.

2. Reproducible Workflow

2.1 The Cartographic Design Process

A map is designed according to its aims (exploratory vs explanatory), to the degree of knowledge of the data (known vs. unknown) and to the audience targeted (for me vs. for a public). A map can have several objectives depending the context (MacEachren, 1994) (DiBiase, 1990).

Nevertheless, different stages always punctuate the cartographic construction process: data retrieval, data handling and cleaning, data processing and analysis, data representation, layout and graphic design, documentation of methods. In this non-linear process, different steps are often distributed among different operators or tools.

In practice, each maps is the sum of a multitude of small choices determining the final rendering (colors, sizes, words, etc.). A map is the result of a set of technical actions that materializes a thinking process. The question of the objectivity is raised at each step. Our idea is to render these different stages of construction in an integrated, readable, shareable scripting language. The script describes the making-off of the map; the written trace of the steps involved in its construction.

2.2 R for Spatial Analysis

Among various scripting solutions we have decided to use the R software because of the high versatility it offers and its widespread use among our own scientific community. Most of findings we propose here are not language dependent and could apply to other technical solutions (Python for example is a classical alternative to R). R (R Core Team, 2018) is a free software environment for statistical computing and graphics. R is enriched by various packages, packages are user contributed plug-ins that add functionalities to the software. The huge increase in the number of packages in the recent years (Hornik et al., 2018) reflects the ever growing popularity of the software. The first purpose of the software is to compute statistics however a large number of thematics were rapidely introduced (econometrics, NLP, web technologies...). Spatial thematics have been specifically addressed since 2003 via `rgdal` (providing bindings to `GDAL` and `PROJ.4` libraries to manipulate geographical projections and data import/export). Quickly

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-38-2019 |

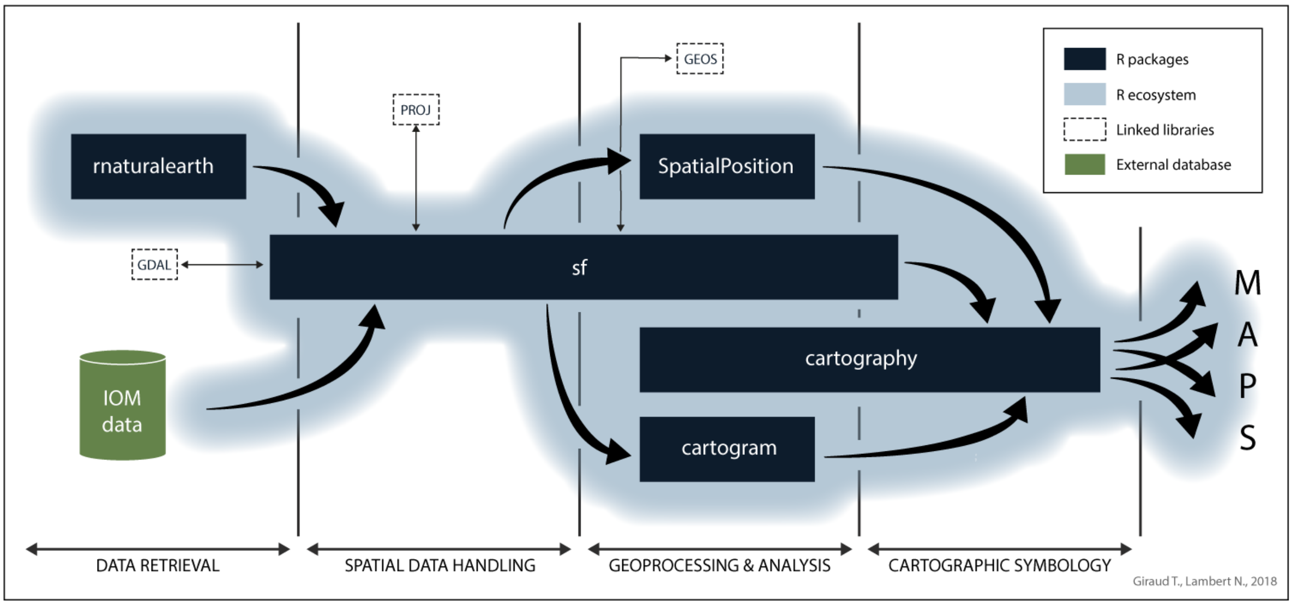

Figure 1. Reproducible mapping process in the R ecosystem

other packages were created to tackle specific thematics: `raster` for raster data handling, `sp` for vector data handling or `rgeos`, an interface to the `GEOS` library aiming at geo-processing spatial features. These packages have fostered the developpement of a dedicated spatial environement within R. In 2016, the `sf` package (Pebesma, 2018) was published in an effort to unite import, export, transformations and data handling functionnalities in a single package. On the geovizualisation side, the `cartography` package (Giraud and Lambert, 2017) first appears in 2015 and allows users to design maps following Bertin's semiology rules (Bertin, 1967) and producing high quality maps with standard layouts (barscale, north arrow...). In addition to this numerous other packages are dedicated to specific spatial data processing needs (statistical analysis, transformation...). The large R ecosystem of packages covers all the steps of the map design process: data and basemap gathering, data handling and transformation, geoprocessing, data analysis, statistics, geovizualisation and layout design.

2.3 Ensuring Traceability and Collaborative Work

The use of a mature and vast script language like R let us design comprehensive geographical analyses. The main output of these analyses is a set of raw text files (R programs) and figures. This material alone is not sufficient to ensure reproducibility. Raw text files do not show explicitly what tasks a program is performing. They do not explain the methodological choices that were made. They do not allow to trace the various versions of the program, nor the contributions of each authors. To sum up, the study of the bare material demand the reader a lot of effort to reproduce analyses. To ease the reproducibility, we suggest a toolset based on two concepts: literate programming and version control. Literate programming, introduced by Donald Knuth (Knuth, 1984), strongly links analyses, statistical outputs and their related code in a single document. We specifically suggest the use of Markdown (Leonard, 2016), a lightweight markup language particularly well integrated in the R environment. The Markdown language allows easy combining or R program pieces and explanatory texts. Litterate programing ensures a good readability of the programs and analyses and is well suited for version control systems (or VCS). VCS are made for and used in software development but they also meet the needs of reproducible research processes (Ram, 2013). These frameworks allow to manage and track changes and versions of datasets, statistical codes, figures and manuscripts. They also offer asynchronous collaborations among authors by creating "branches" or "forks". Among the various open source VCS we suggest the use of git which is widely used.

3. Mapping Deaths of Migrants in the Mediterranean Sea

3.1 Thematic Issues

In March 2011, an inflatable boat with 72 migrants on board left Libya for Europe. Running out of fuel, the boat eventually drifted until it ran aground on the Libyan coast 14 days later. Out of 72 passengers, 63 died, including 20 women and 3 children (Heller and Pezzani, 2014). More recently on September 25, 2018, a ship carrying 25 migrants from Mdiq-Fnideq on the Moroccan coast to Spain was stopped by the Marocan authorities. During the interception a young woman was shot dead and 3 other severly injured. On April 18 2015, the Mediterranean was facing the worst shipwreck in its recent history. During the night, an overloaded trawler with nearly 750 people on board capsized. Only 28 people were rescued.

Theses tragic events are not isolated, year after year the Mediterranean is the scene of multiple shipwrecks of migrants seeking to reach European coasts. The Mediterranean is one of the most frequented and best monitored seas in the world but also one of the most dangerous area for migrants.

According to the International Organization for Migration, more than 17,000 people died or disappeared between 1 January 2014 and 3 October 2018 (figure 2). This means that about ten people disappear into the sea every day.

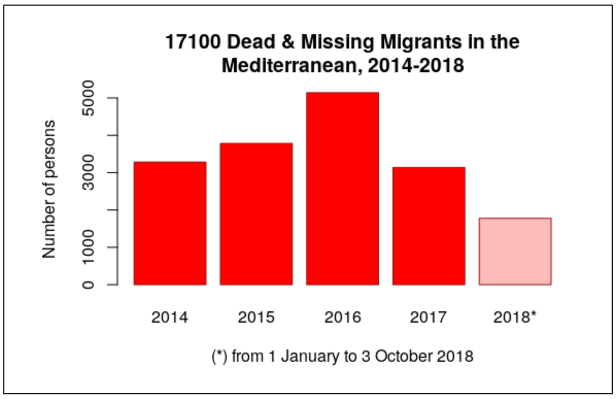

Figure 2. Dead and missing migrants in the Mediterranean, 2014-2018

3.2 Data

The first database listing the dead and missing migrants at European scale was designed in 1993 by UNITED for Intercultural Action, a Netherlands based NGO. A second database was created in 1998 by the italian journalist Gabriele Del Grande. Later, a collective of journalists compiled and cross validated these previous sources and additional ones to create a third database, the Migrant's File (mid 2013 - mid 2016). Since 2016 the only maintained database on the whole mediterranea is held by an UN organization: the International Office for Migration (IOM). All theses databases have in common their collecting process: the gathering of information through official records, media reports, NGO's surveys and interviews of survivors. Data provided by these organisations can be usefull to grasp the phenomena, even if they are not comprehensive, and not fully reliable at the event level. Migrations experts commonly agree on the fact that these databases tend to underestimate the number of dead and missings (Kobelinsky and Le Courant, 2017).

3.3 Implementation

In the following example we used a set of R packages (figure 1) to conduct analyses and produce maps. These packages cover all steps of the map production process. `rnaturaleath` (South, 2017), `sf` and `base` functions are used to import datasets (data and geometries). Data handling (basemap selection and modification) is done with `sf`. Some maps are based on spatial analysis and geoprocessing therefore we have used specialised packages to produce them: `cartogram` (Jeworutzki, 2018) for cartogram creation, `SpatialPosition` (Giraud and Commenges, 2017) for spatial smoothing, `cartography` for transforming irregular data to grids. We have mainly used `cartography` for cartographic symbologies, layout, legend or scale. All data and scripts are available online[1]. This material allows to reproduce our analyses and can be easily adapted to other time periods or geographical areas.

4. Geovizualisation

The first map triying to visualise the amount of dead and missing migrants in the mediterranean area was published in 2003 (Clochard, 2003). This initial map has been regularly updated, enriched and republished by differents authors (Rekacewicz and Clochard, 2006), (Migreurop, 2009) (Migreurop, 2012) (Migreurop, 2017).

At each update, we can visualize the shape and the geography of the European Union's migratory border. Cartographic visualizations help to understand the geographical logics, and to measure the dangerousness of the different routes to Europe over time (exploratory approach, scientific context). These maps are also effective tools to alert public opinion on the magnitude of this human tragedy taking place at the borders of the European continent (explanatory approach, general public). In particular, this work has shown that whenever a passing point is closed (Strait of Gibraltar, Canary Islands, Lampedusa, etc.), migration flows are redirected towards often more dangerous routes, causing an increase in the number of deaths. Maps can play an important function in interpelling governments and citizens. Considering the several cartographic representations already made in the past, we propose in this paper several reproducible, configurable and automated cartographic representations. Each of them provides a particular focus on the thematic.

4.1 Proportional Symbols

The first map (figure 3) represents the number of dead or missing migrants in the Mediterranean over the period 2014-2018. Each red circle corresponds to one event; the area of the circles is proportional to the number of dead or missing people during that event. Three main areas are clearly identifiable: the Alboran Sea in the east, near Gibraltar; the Aegean Sea in the west, between Greece and Turkey; and the central Mediterranean (area with the highest number of circles), off the Libyan coast. In addition, events near Cyprus, Crete and the Egyptian coast are also located. While this type of map (used, for example, on The Migrant's file website) makes it relatively easy to distinguish these different areas, much of the information is hidden because of the overlaps between circles.

4.2 Cartographic Aggregation (Clusters)

To solve the problem of overlapping information a first solution consists in aggregating the circles according to their locations. We used a hierarchical clustering of the events based on their euclidean distance matrix. Then we calculated the weighted average of the coordinates of the events in each classes to set the location of each circles. On this simplified map (figure 4), the message is clear: the most dangerous area is the central Mediterranean, with more than 13,000 dead and missing persons over the period.

4.3 Cartographic Aggregation (Gridded Map)

A representation in a regular grid (figure 5) can solve both the circles overlaps problem on the first map and provide precise information on the locations. Events are aggregated in the cells of a regular grid (one hundred square kilometers resolution). A density calculation is then used to determine the dangerousness of the different zones. With 24.5 dead or missing per square kilometer, the most dangerous area is located very close to the Libyan coast. This area should be a priority zone for planning sea rescues.

[1] https://riatelab.github.io/MDM/index.html

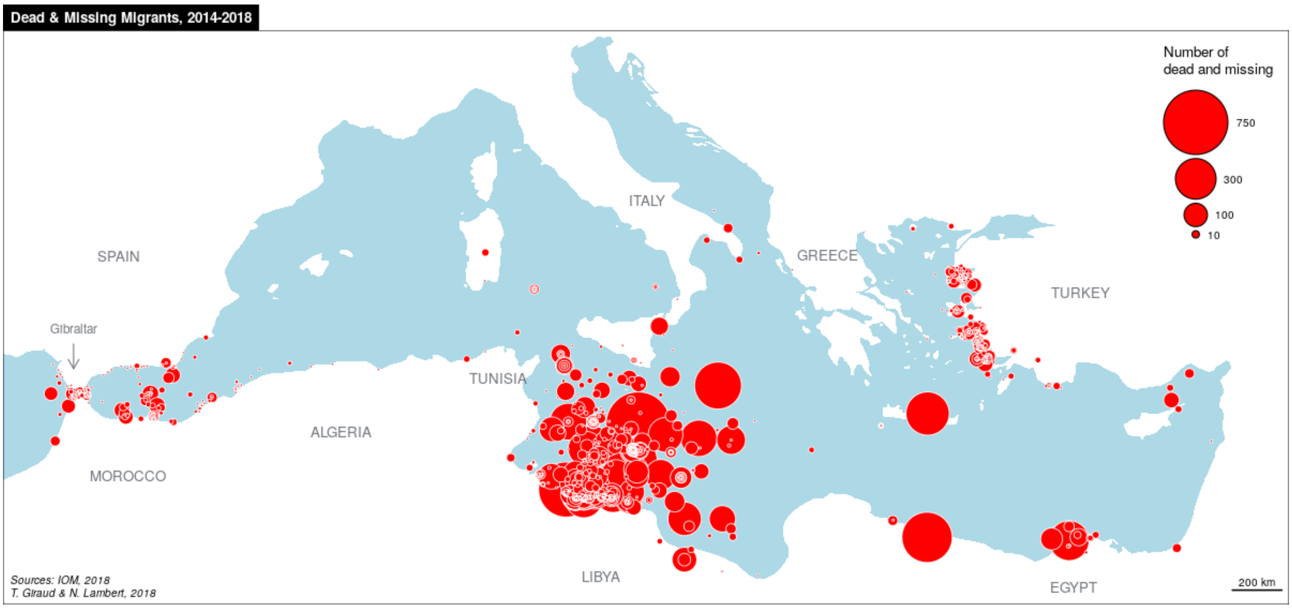

Figure 3. Number of dead or missing, simple representation

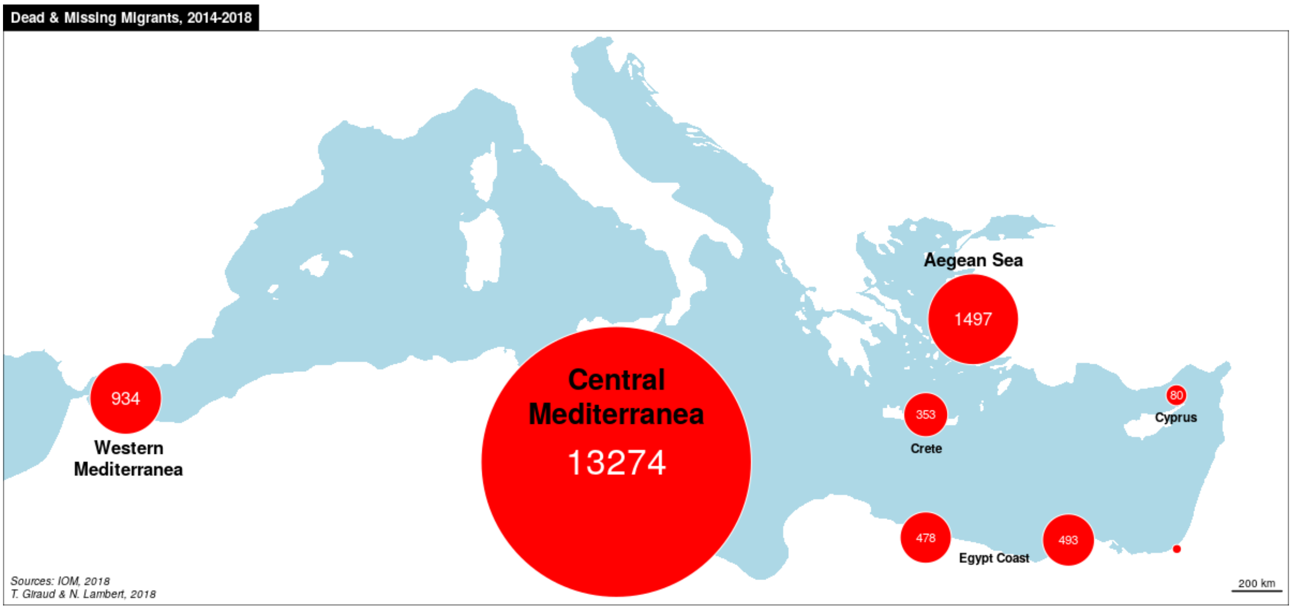

Figure 4. Distance based aggregation

4.4 Cartographic Aggregation (Isopleth Map)

The fourth map was made using a smoothing method. We specificaly used the Stewart potentials with a 75km span and an exponential interaction function, see (Stewart, 1942) and (Commenges et al., 2016) for detailed explainations. The classic output of this method is an isopleth map that does not adequately reflect the order of magnitude of the data points. That is why we used a pseudo-3D rendering using the relief contour method (Tanaka, 1950). Smoothed maps emphasizes the spatial structures by constructing a simplified image of a geographical phenomenon. On this map (figure 6) the geographical organization of the phenomenon is clear. The Central Mediterranean is the largest area of the region. It is also the area where the maximum values are observed, especially on the Lybian coast. As raw values of potentials are difficult to interpret for non initiated readers, we have chosen to translate them into a qualitative scale.

4.5 A Cartographic Metaphor

The next map uses the same smoothing method but potentials values are inverted to represent hollows instead of hills. This Mediterranean whirlpool is the cartographic metaphor of the enormous chasm in which thousands migrants have drowned, in particular off the coast of Libya. This map is a possible representation of the European "migratory border" (Lambert and Clochard, 2015).

4.6 Dorling Cartogram

Despite allowing a good understanding of spatial structures, the main flaws of the previous representations is that they hide each individual stories by aggregating them. Each shipwreck is gummed and melted in a colored area on the map. The Dorling Cartogram method (Dorling, 1996) (figure 8) is a good approach to solve the overlapping circles

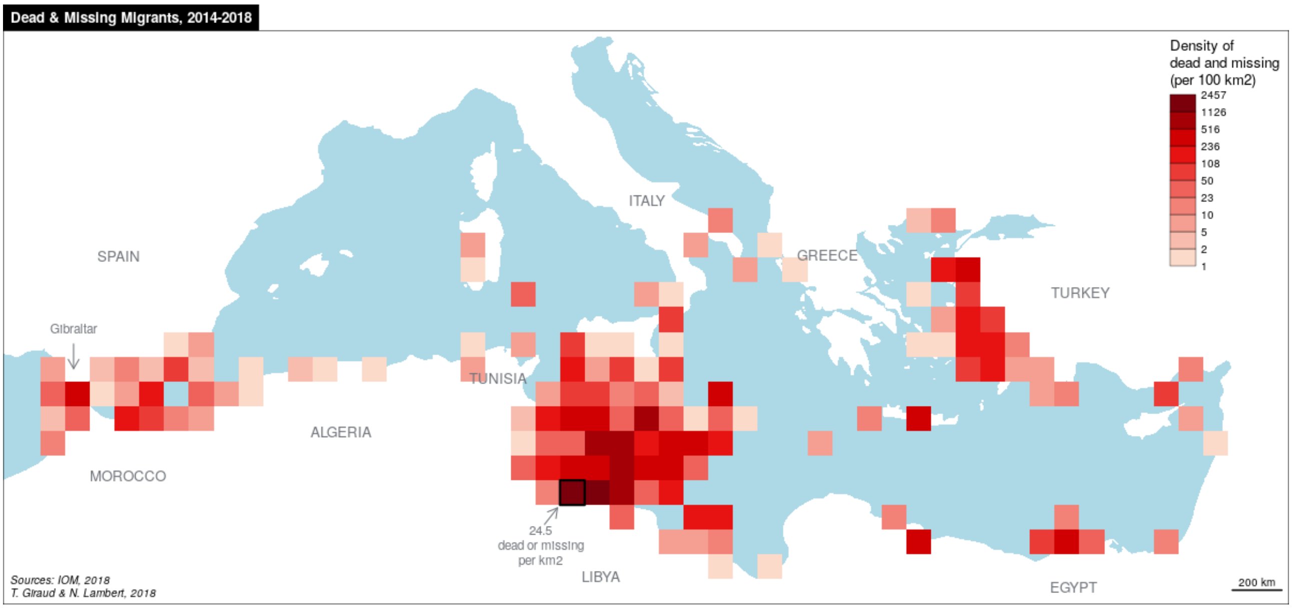
Dead & Missing Migrants, 2014-2018
Density of dead and missing (per 100 km2)
2457
1126
516
236
108
50
23
10
5
2
1
ITALY
SPAIN
GREECE
TURKEY
Gibraltar
TUNISIA
ALGERIA
MOROCCO
24.5 dead or missing per km2
LIBYA
EGYPT
Sources: IOM, 2018
T. Giraud & N. Lambert, 2018
200 km

Figure 5. Aggregation in a regular mesh

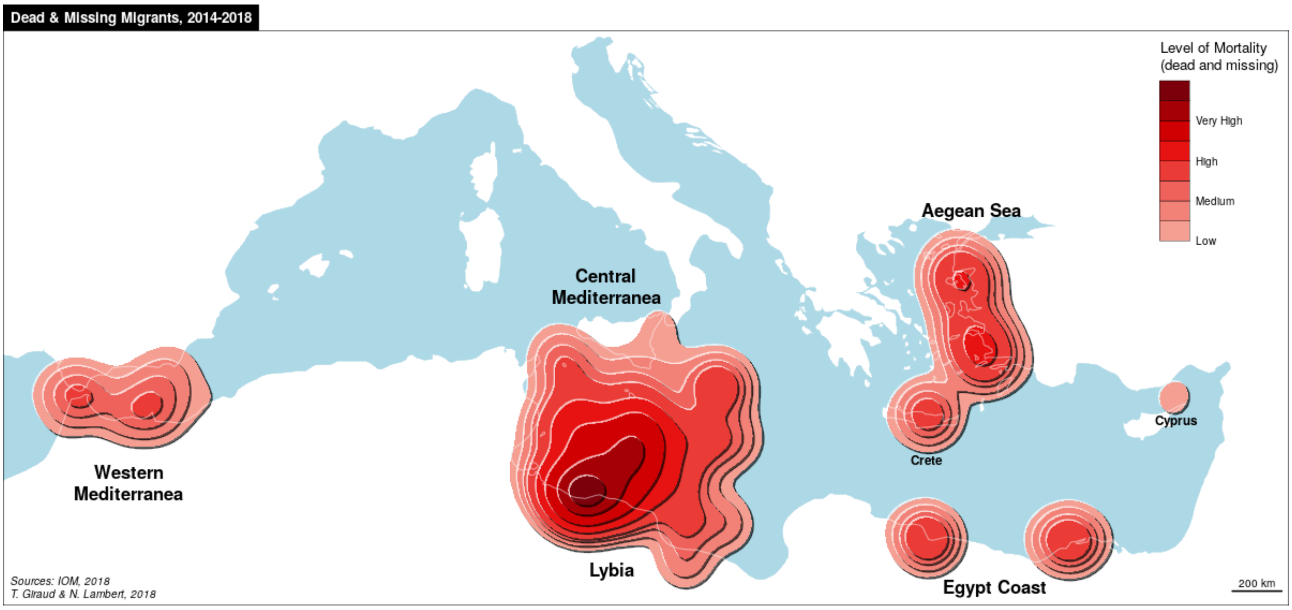

Figure 6. Smoothing using potentials

problems without aggregating information on the size and number of shipwrecks. The compromise is to override the exact locations of the events in order to display them all.

5. Conclusion

Maps are the result of creative acts, choices and intentions of authors. Maps do not merely represent space, they shape arguments (Wood, 2010). Moreover, no map is perfect and definitive. Proposing several cartographic representations is a way to avoid freezing a geographical discourse in a unique (and therefore incomplete) cartographic representation. Multi-representation allows us to reveal geographical complexity. Each map has its strengths and weaknesses, each map shows some facts and hides others. In a scientific context, these choices have to be discussed.

To allow scientific discussions, choices leading to the creation of the map should be meticulously recorded. This requires the implementation of a workflow that can be traced, documented and shared. The R ecosystem, combined with the Markdown language and the git version control system, makes it possible to fully cover these issues. This allows to share a map, the code and data used for its creation along with a literal description of the mapping intentions.

If the R language can be used to achieve reproducibility objectives, other technical choices are quite available. The stake is not at the choice of a specific technical solution but rather at anchoring cartographic process in the scientific production context.

References

Bertin, J., 1967. Sémiologie graphique: Les diagrammes - les réseaux - les cartes [semiology of graphics].

Clochard, O., 2003. La méditerranée: dernière frontière

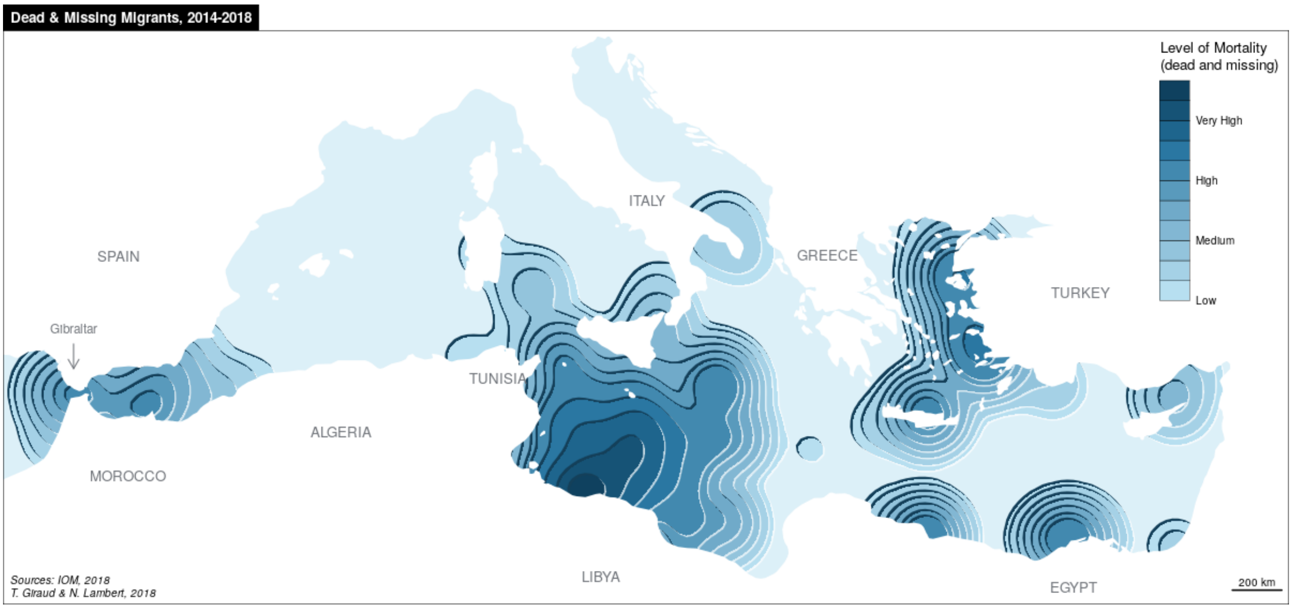

Figure 7. Smoothing using inversed potentials

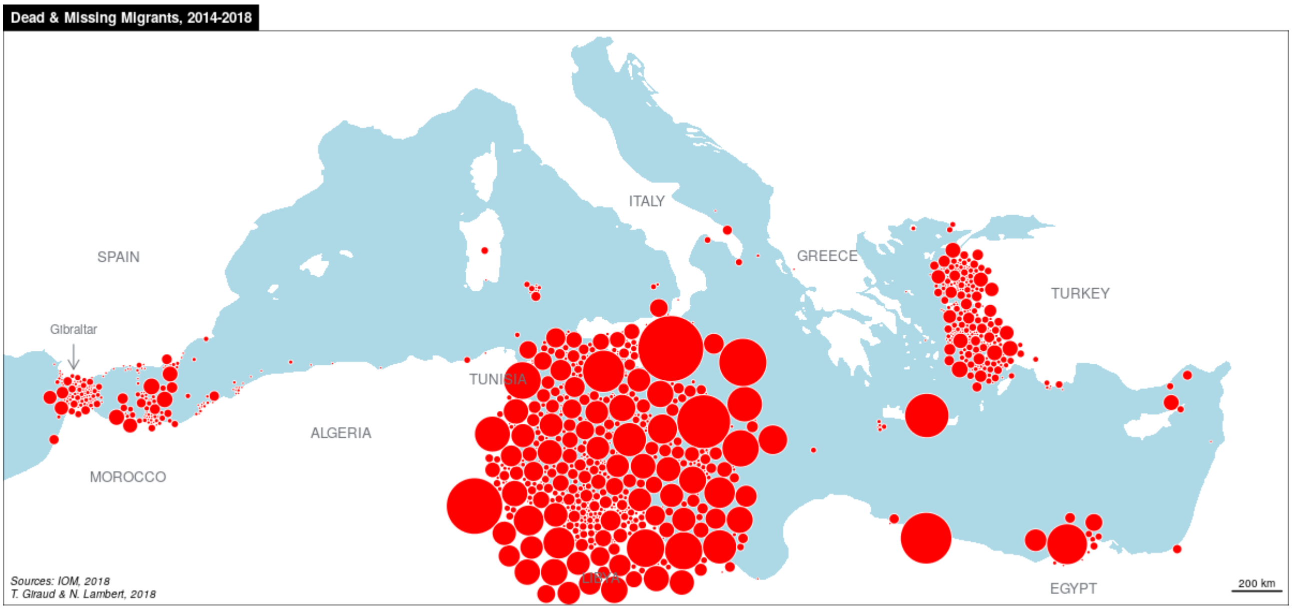

Figure 8. No overlap placement for circles

avant l'europe. *Les Cahiers d'Outre-Mer. Revue de géographie de Bordeaux* 56(222), pp. 159–180.

Commenges, H., Giraud, T. and Lambert, N., 2016. Espon fit: Functional indicators for spatial-aware policy-making. *Cartographica: The International Journal for Geographic Information and Geovisualization* 51(3), pp. 127–136.

DiBiase, D., 1990. Visualization in the earth sciences. *Earth and Mineral Sciences* 59(2), pp. 13–18.

Dorling, D., 1996. *Area cartograms: their use and creation, concepts and techniques in modern geography*. CATMOG: Concepts and Techniques in Modern Geography, Vol. 59, Institute of British Geographers.

France, A., 1928. *Balthasar*. Calmann-Lévy.

Giraud, T. and Commenges, H., 2017. SpatialPosition: Spatial Position Models. R package version 1.2.0.

Giraud, T. and Lambert, N., 2017. Reproducible cartography. In: M. Peterson (ed.), *Advances in Cartography and GIScience. ICACI 2017. Lecture Notes in Geoinformation and Cartography.*, Springer, Cham, Switzerland, pp. 173–183.

Heller, C. and Pezzani, L., 2014. Liquid traces: Investigating the deaths of migrants at the eu's maritime frontier. *Revue européenne des migrations internationales* 30(3), pp. 71–107.

Hornik, K., Ligges, U. and Zeileis, A., 2018. Changes on cran. *The R Journal* 10(1), pp. 556–559.

Jeworutzki, S., 2018. cartogram: Create Cartograms with R. R package version 0.1.0.

Knuth, D. E., 1984. Literate programming. *The Computer Journal* 27(2), pp. 97–111.

Kobelinsky, C. and Le Courant, S., 2017. Introduction. In: C. Kobelinsky and S. L. Courant (eds), *La mort aux frontières de l'Europe : retrouver, identifier, commémorer*, Bibliothèque des frontières.

Lambert, N. and Clochard, O., 2015. Mobile and Fatal: The EU Borders. In: *Borderities and the Politics of Contemporary Mobile Borders*.

Leonard, S., 2016. The text/markdown media type. Technical report.

MacEachren, A. M., 1994. Visualization in modern cartography: setting the agenda. *Visualization in modern cartography* 28(1), pp. 1–12.

Migreurop, C. O., 2009. *Atlas des migrants en Europe*.

Migreurop, C. O., 2012. *Atlas des migrants en Europe*.

Migreurop, C. O., 2017. *Atlas des migrants en Europe*.

Pebesma, E., 2018. Simple features for r: Standardized support for spatial vector data. *The R Journal* 10(1), pp. 439–446.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ram, K., 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine* 8(1), pp. 7.

Rekacewicz, P. and Clochard, O., 2006. Des morts par milliers aux portes de leurope. *Le Monde diplomatique*.

South, A., 2017. rnaturalearth: World Map Data from Natural Earth. R package version 0.1.0.

Stewart, J. Q., 1942. A measure of the influence of a population at a distance. *Sociometry* 5(1), pp. 63–71.

Tanaka, K., 1950. The relief contour method of representing topography on maps. *The Geographical Review*.

Wood, D., 2010. *Rethinking the power of maps*. Guilford Press.

https://doi.org/10.5194/ica-proc-2-38-2019 |