Using R to clean data from a map-based survey: challenges and lessons learned

Beatriz Milz, Gabriel Machado Araújo, Sandra Momm

Key-words: maptionnaire, data cleaning, tidyverse, mobility, urban planning.

In this project, we applied a map-based survey in 3 different regions: São Paulo (Brazil), Cape Town (South Africa) and Dortmund (Germany). With more than 300 responses, the survey aimed to understand how the COVID-19 pandemic affected the mobility of people from urban marginalized groups.

In recent years, map-based surveys have become more popular in a variety of research fields. In this research, we use a software called Maptionnaire (https://www.maptionnaire.com/). This type of software has simplified the collection of geospatial data by allowing respondents to interact with maps and provide location-based responses. Maptionaire is a software that provides a good user interface, facilitating data collection. The platform has a dashboard with simple plots of the data, that can be accessed as long as the license is active (Maptionnaire is not an open-source software).

However, in order to conduct data analysis, users need to download survey data from the platform (as Excel Files and Shapefiles) and use other statistical programs (Kyttä *et al.*, 2023). Kyttä *et al.* (2023) present some of the challenges faced in performing this step:

> "The data reading requires experience to understand how the survey contents translates to a database with all relevant meta data, including response times, zoom level and background map used for map markings etc. Noteworthy is also that data collected with online map-based surveys can also include qualitative descriptions in relation to mapped sites or in terms of non-spatial survey questions. Such data also requires analysis as a separate effort outside the Maptionnaire platform. Furthermore, the Maptionnaire analysis window does not assess spatial data quality, which all users should be cautious about. Before starting any analysis, collected PPGIS data should be cleaned by detecting, correcting, or removing inaccurate spatial records, and organized for the actual data analysis. Such data manipulation may include value (re)classification, data (re)ordering, data queries, and removal of outliers" (Kyttä *et al.*, 2023, p. 82).

The Excel file exported from Maptionnaire presents one sheet for general questions (non-spatial survey questions), and one sheet per each "Map question" (spatial survey questions). This results in several sheets that need to be joined together before the analysis can be performed, using the respondent ID as the key variable. Performing this (and other steps evolving in preparing the data for analysis) on Excel would be too

time-consuming, more prone to errors, and irreproducible. For that reason, the research group chose to conduct this step using a programming language called R (R Core Team, 2024).

This presentation will share how we used R to organize the raw data for the data analysis in a tidy format (WICKHAM, 2014), and the main challenges and lessons learned.

This presentation will also highlight how R, through the use of tidyverse (https://www.tidyverse.org/) (WICKHAM *et. al*, 2019) and package sf (https://r-spatial.github.io/sf/) (PEBESMA & BIVAND, 2023), enabled reproducible data cleaning processes, ultimately enhancing the quality and reliability of the subsequent data analysis.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2021/07554-8 and 2024/05779-0.

References

Kyttä, M., Fagerholm, N., Hausner, V.H., Broberg, A. (2023). Maptionnaire. In: Burnett, C.M. (eds) **Evaluating Participatory Mapping Software.** Springer, Cham. https://doi.org/10.1007/978-3-031-19594-5_4

Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

R Core Team (2024). **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Wickham H., *et al.* (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.