About Writing Dynamic Documents with R

Author Name[1]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, Zurich
name@geo.uzh.ch

Abstract

This is the abstract of the template document used to show how to write publications in R with R Markdown and the help of some packages. Based on a concrete use case, this document exemplifies some of the caveats that may occur when writing such a document and publishing it online on a Git repository. It also presents typical use cases in Markdown usage and presents some tricks.

Introduction

This example publication serves as a motivation on how to create reproducible documents in R and aims to promote reproducible research in general.

State of the Art

Various authors in qualitative and quantitive research argue that as many parts of the research workflow as possible should be reproducible. Brunsdon (2015) state "Reproducible quantitative research is research that has been documented sufficiently rigorously that a third party can replicate any quantitative results that arise".

To further motivate you, read (Buckheit and Donoho 1995, Healy 2011, 2016, Nüst *et al.* 2011, LeVeque *et al.* 2012, Pebesma *et al.* 2012, Vandewalle 2012, Baker 2016, Editorial 2016) or the short and to the point editorial from Editorial (2016).

Case Study: Parc Adula

This case study presents a small subset of data from a current study conducted at the Department of Geography at the University of Zurich. The study investigates social negotiations revolving around a national park project in Switzerland – *Parc Adula* – and aims for a better understanding of how people reason in a public environmental debate.

Exploratory topic analysis

For this case study, 16 interviews were carried out. Each of these semi-structured interviews was analyzed resorting to Mayring's qualitative content analysis (Mayring and Fenzl (2014)) – resulting in a code system, which was derived through mainly inductive category development. The following plots display sample output from the MXAQDA software for qualitative data analysis.

Overview of the interviews and representatives:

- *Cantonal Goverment* (n: 4): Representatives from four different departments
- *Environmental Organisation* (n: 1): Representative from a specific interest group with advisory function.
- *Federal Goverment* (n: 2): Must ensure the park follows the laws and decrees
- *Local* (n: 5): Local representatives of the park region
- *Parc Team* (n: 2): Team members involved in the park planning process
- *Tourism* (n: 2): Local tourism representatives

The following plot presents the frequency of occurence of a select list of topics that occured in the interviews. While there seems to be more focus on the *Pro Argument* against the *Contra Argument* during the interviews, topics on *Tourismus* seem to have far more weight than those on *Biodiversität*.

Table 1: Topic mentions.

Code	Mention
Biodiversitaet	5
Contra Argument	39
Pro Argument	68
Tourismus allgemein	48

The next figure presents the frequency matrix of the topic occurences across the different interviews. It provides an overview of where topics are mentioned and by whom.
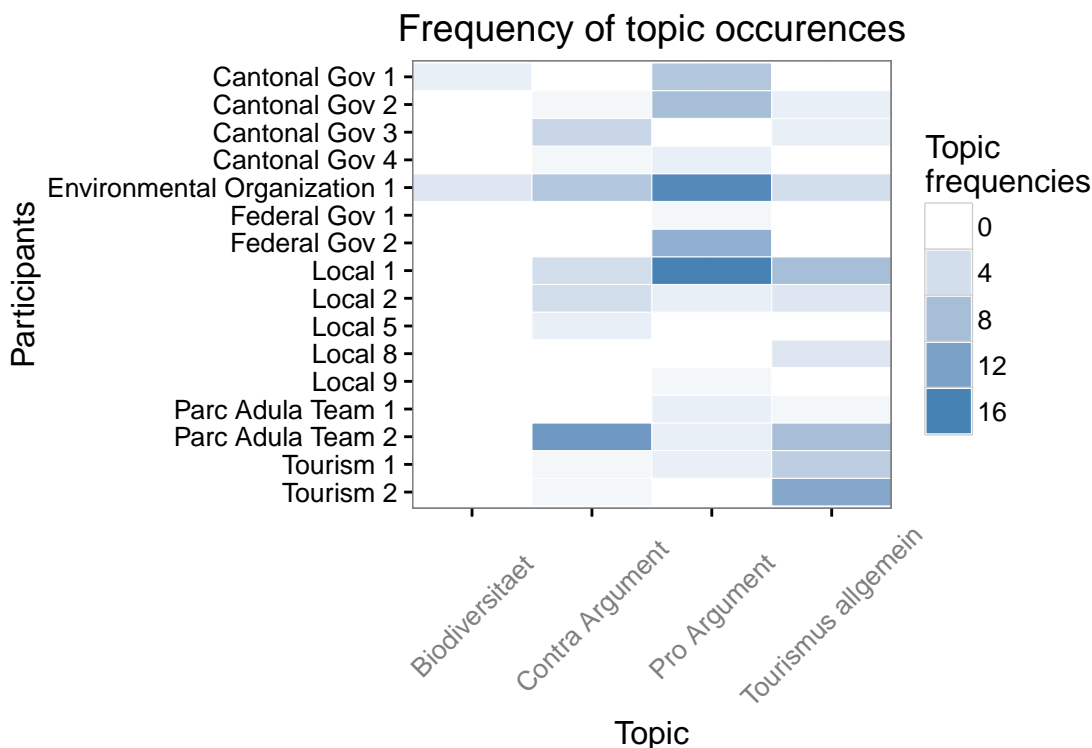

Figure 1: Frequency matrix of a selected list of topics across the various representants

Notes on reproducibility: Depending on the data to analyze, privacy may play a role. While for the analysis itself the data is being anonymised, storing the raw or preprocessed data on a public repository may pose privacy issues or even constitute a violation of contract.

Google query timeline

Overview of the Google trend evolution of the search query: *Parc Adula* (url, provides a CSV file). The timeline shows overall a small amount of queries for this word combination, with a spike on 2015-11-01. This was retrieved on August 11, 2016.

Notes on reproducibility: Web APIs are subject to license restrictions, can get altered by the service provider, or can simply cease to exist, so consider them carefully before using them in a scientific project. Consider instead using software which you can store locally and can better control the parameters and settings. If collecting data from an API, ensure to note down as much as possible about the data collection: the date range, all the query parameters, including the service limits at the time, any interruptions in service, and so on. It's also wise to back up the data thus obtained, if at all possible!
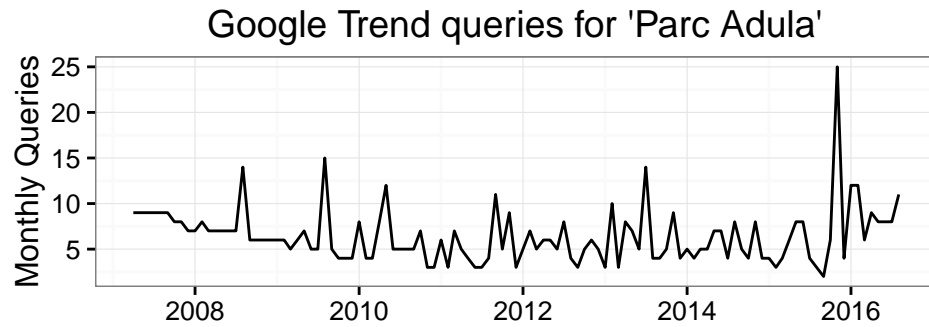

Figure 2: Timeline of queries for Parc Adula set in the Google search engine

Case study area

The proposed Parc Adula national park candidate is situated in Switzerland in the border region of the cantons Ticino and Grisons (Graubünden). The map below presents the current outer perimiter of the planned national park.

Notes on reproducibility: Due to license restrictions of the open geodata, it is not possible to store the data on a public Git repository. The included script `R/loadMapData.r` downloads the data directly from the link provided in the geodata catalog infobox of http://maps.geo.admin.ch

Concluding discussion

This template is based on data from an ongoing research project and presents some typical examples of material that could be used in a publication written in RMarkdown. It shows how to include data and analyses, features plots, tables, literature, and various markdown elements. The generated files in *PDF*, *Word* or *HTML* often still need some fine-tuning afterwards (particularly in Latex). It is nevertheless a great way of documenting the research process, generating initial drafts, and sharing workflows with collaborators or a wider audience.

Acknowledgements

References

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533 (7604), 452–454.

Brunsdon, C., 2015. Quantitative methods I: Reproducible research and quantitative geography. *Progress in*

Figure 3: Planned perimeter of Parc Adula, Switzerland, Data source: Swisstopo

Human Geography.

Buckheit, J. and Donoho, D., 1995. WaveLab and Reproducible Research. *Wavelets and Statistics*, 103, 55–81.

Editorial, 2016. Reality check on reproducibility. *Nature*, 533 (7604), 437–437.

Healy, K., 2011. Choosing Your Workflow Applications. *The Political Methodologist*, 18 (2), 9–18.

Healy, K., 2016. *The Plain Person's Guide to Plain Text Social Science.* Healy2016.

LeVeque, R.J., Mitchell, I.M., and Stodden, V., 2012. Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, 14 (4), 13–17.

Mayring, P. and Fenzl, T., 2014. Qualitative inhaltsanalyse, 601–613.

Nüst, D., Stasch, C., and Pebesma, E., 2011. Connecting R to the sensor Web. *In*: *Lecture notes in geoinformation and cartography.* 227–246.

Pebesma, E., Nüst, D., and Bivand, R., 2012. The R software environment in reproducible geoscientific research. *Eos, Transactions American Geophysical Union*, 93 (16), 163–163.

Vandewalle, P., 2012. Code Sharing Is Associated with Research Impact in Image Processing. *Computing in Science & Engineering*, 14 (4), 42–47.