Reproducibility review of: Unlocking social network analysis methods for studying human mobility

Jakub Krukar

2022-06-09

This report is part of the reproducibility review at the AGILE conference. For more information see https://reproducible-agile.github.io/. This document is published on OSF at OSF LINK HERE. To cite the report use

Krukar, J. (2022, May 23). Reproducibility review of: Unlocking social network analysis methods for studying human mobility. https://doi.org/10.17605/OSF.IO/MVQCW

Reviewed paper

Wiedemann, N., Martin, H., and Raubal, M.: Unlocking social network analysis methods for studying human mobility, AGILE GIScience Ser., 3, 19, https://doi.org/10.5194/agile-giss-3-19-2022

Summary

The paper provides a link to a GitHub repository that was initially difficult to use but was promptly improved by the authors after an email exchange. The repository contains only one out of two datasets presented in the paper but most results based on this dataset have been successfully reproduced with minor disparities due to automated scaling of graphs. In sum, the manuscript has been partially reproduced. The repository is well-documented, it includes the documentation of required software versions, and the authors' response to questions and bugs has been prompt and helpful.

Reproducibility reviewer notes

Running the code

The provided GitHub repository contains one out of two datasets used in the paper (the Foursquare dataset). Raw data are provided, together with python scripts for processing it. After executing them the user must run two provided R scripts in order to fit relevant models and, subsequently, another python script in order to generate the results that are reported in the manuscript (tables and figures).

In its initial form the repository was not described well and difficult to use. After contacting the authors they have promptly improved the documentation of the repository and responded by fixing minor bugs that initially prevented me from running the procedure. With the updated repository I found the procedure to be very clearly described and was able to follow all steps described in the README file. Some python modules required manual installation on my machine, despite running the command

```
pip install -r requirements.txt
```

Using the suggested `renv:restore()` function, there were no compatibility issues with my R environment and all scripts ran successfully. The runtime was similar or shorter to that indicated in README. While running the final step of the procedure:

```
cd python_scripts
python analyze.py
```

I have obtained multiple warnings that the authors suggested to ignore. It was not clear to me (and not mentioned in README) that results will be saved in a newly created 'results_foursquare' directory. This was clarified in the communication with the authors.

Results of the reproduction

The newly created 'results_foursquare' directory contains 4 files with figures and 2 files with numerical results. All figures presented below contain the versions obtained by running the code from the repository.

Figure 1

As indicated in the caption and further clarified in the communication with the authors, this is an exemplary figure that was not generated from the provided data.

Figure 2

File 'results_qap.pdf' contains results that seem to correspond to Figure 2 in the manuscript (i.e., to its part based on Foursquare data that in the manuscript is marked in orange).

The shape of the upper two histograms corresponds to the shape of histograms in the manuscript. The shape of the bottom two histograms does not correspond to those in the manuscript. The vertical axis (Number of users) demonstrates that the values underlying the histograms are different in the manuscript compared to those generated by the code. For example, the maximal values in the orange part of the upper-left histogram in the manuscript seem to be 4. In the reproduced figure (see below), the maximal value seems to be 14. ~~The figure from the manuscript has not been successfully reproduced.~~

In subsequent communication, the authors indicated that the manuscript contains a bug in the specification of the vertical scale (which should refer to the "Percentage" and not "Number" of users). This explains why the shape of the upper two reproduced histograms is identical to that in the manuscript, but the values are not.

As explained in the communication with the authors, the bottom two histograms in the manuscript have a different (automatically adjusted) bin width parameter compared to those reproduce in the code.

This explains why the shape of the bottom two histograms is different from that in the manuscript. The ranges of values on the x-axis seem to be similar to those in the manuscript, but since the scalling paramaters used in the manuscript are unknown, the figure has been only partially reproduced.

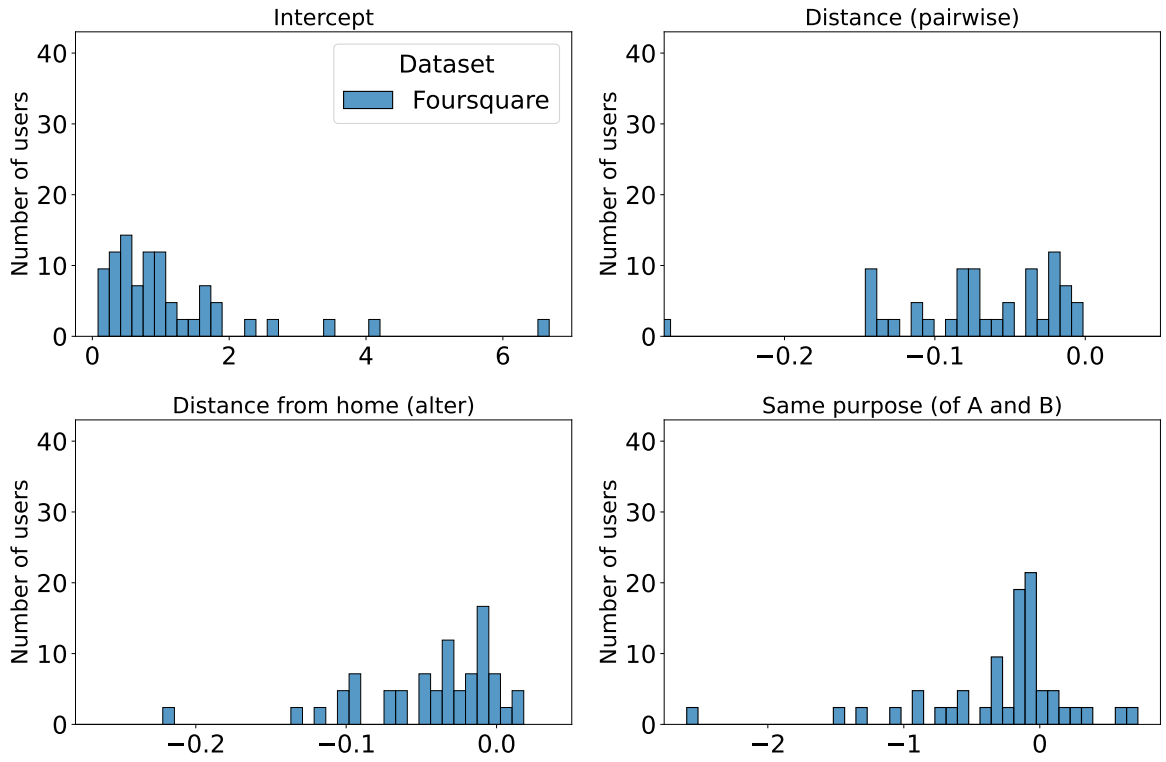


Figure 3

Figure 3 is a scatterplot. The reproduced scatterplot seems to correspond to that in the manuscript, with 2 minor exceptions: (1) the colour in the reproduced figure is blue, which corresponds to the other of two datasets shown in the manuscript (where Foursquare data is presented in orange and the unavailable dataset in blue); (2) the ranges on the x- and y- axes are different, since the figure in manuscript plots a wider range of values. Otherwise the figure from the manuscript has been successfully reproduced (for the part of the data available).

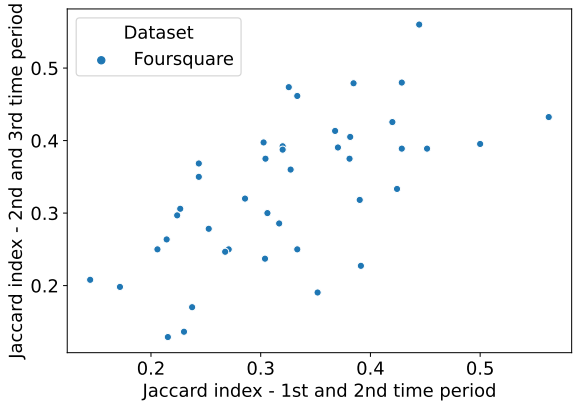
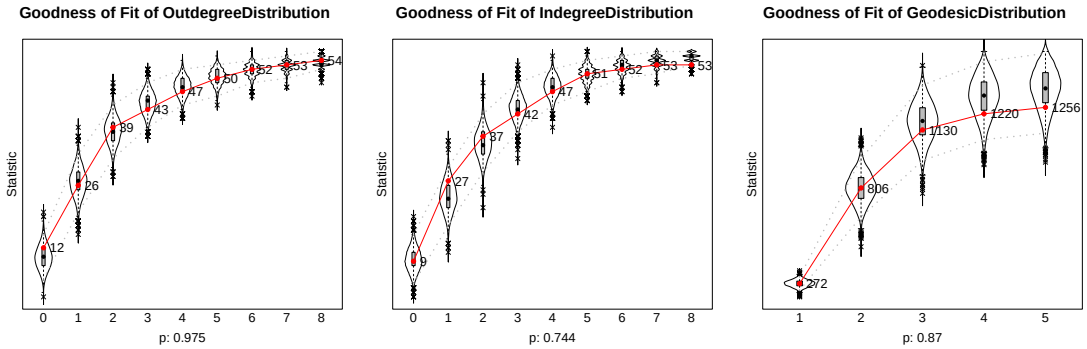


Figure 4

I was able to locate the graphs corresponding to user 327 in the directory 'data/foursquare_120/327' although the documentation does not explicitly mention that new files will be saved there. Three separate

files seem identical with those displayed in the figure. Figure 4 has been successfully reproduced.



Other figures

In addition, the 'results_foursquare' folder contains two files that are not clearly described but do not directly correspond to any figure shown in the manuscript: 'hist_jac1.pdf' and 'hist_jac2.pdf'.

Table 1

The 'results_foursquare' directory contains a file 'terminal.txt' with some numerical results. The first set of results correspond to the numerical results presented in Table 1, although are not described as such. Numerical results of Table 1 have been successfully reproduced.

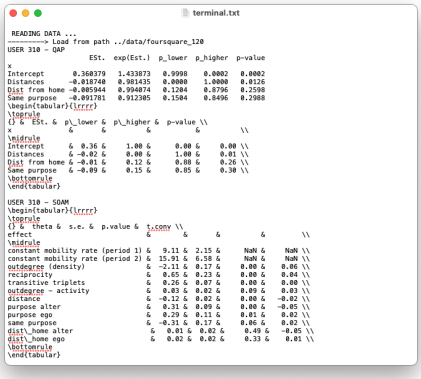


Table 2

~~I was not able to locate numerical results corresponding to Table 2.~~ In the subsequent communication, the authors indicated that the output of Table 2 is available in the file 'terminal.txt' under the line 'QAP: Comparison of intra user and inter user variance'. The numbers in the output differ from those in the manuscript as the table in the manuscript results from the joint computation of both datasets. In addition, the terminal output does not specify which value corresponds to which row in the table. Given that the order of rows in another table (Table 3) is *different* in the terminal and in the manuscript, it is unclear which numbers correspond to which parts of the table. The user would need to analyse the code in order to retrive this information.

Table 3

File 'terminal.txt' contains numerical data corresponding to Table 3 although they are not described as such. Table 3 has been successfully reproduced.

Table 4

An additional file titled ‘results_soam.csv’ contains numerical output corresponding to the last column of Table 4 (i.e., the column relevant to the Foursquare dataset). Table 4 has been successfully reproduced (for the available part of the dataset).

Recommendations

- Add information on where to find the results after the final step of the procedure.
- Describe which results generated by the code correspond to which results in the paper.
- In README, clarify the disparity of Figure 2 in the manuscript and the reproduction.
- In README, indicate how to localize the output of Table 2.