Reproducibility review of: Understanding COVID-19 Effects on Mobility: A Community-Engaged Approach

Philipp A. Friese

2022-06-11

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 https://osf.io/kf8sr/. To cite the report use

> Friese, Philipp A. (2022, May). Reproducibility review of: Understanding COVID-19 Effects on Mobility: A Community-Engaged Approach. https://doi.org/OSF.IO/KF8SR

Reviewed paper

> Sharma, A., Farhadloo, M., Li, Y., Kulkarni, A., Gupta, J., and Shekhar, S.: Understanding COVID-19 Effects on Mobility: A Community-Engaged Approach, AGILE GIScience Ser., 3, 14, https://doi.org/10.5194/agile-giss-3-14-2022

Summary

The software of the paper under reproduction is publicly available on GitHub. The majority of data sets are publicly available in a Google Drive folder. One data set is not available due to privacy policy concerns. Out of the 17 Figures in the paper, 10 are eligible for reproduction. Out of the 10 eligible Figures, 8 Figures where successfully reproduced, one Figure was partially reproduced and one Figure was not reproducible. Reproduction was *partially successful*.

Reproducibility reviewer notes

General remarks

The Data and Software Availability section in the paper links to a GitHub repository[1], which initially contained four Jupyter Notebooks and one data file for reproducing the relevant Figures 7 to 14, 16, and 17. All Jupyter Notebooks contain Python code. The software is published under a GPL-3 license. Initial reproduction was not possible due to missing data files and a reference to a PostgreSQL database, which was not reachable.

After contacting the authors, the repository was updated with four additional Jupyter Notebooks, suffixed with "AGILE". In addition, a link to a Google Drive was provided, which stores missing data files.

The README in the main repository contains information on which Notebooks reproduce which Figures, which is commendable.

During contact with the authors, it was noted that Figure 12 is not reproducible due to the privacy police of the data source SafeGraph. The authors stated that:

> *I was not able to share the whole data due to privacy policy and it was not open-source. Hence, I was not able to reproduce Figure 12.*

Figure 14b is not reproducible due to an issue with the submission code.

Figure 13, 14a and 17 diverge from the corresponding Figures in the paper. After contacting the authors, this difference is explained as follows:

> *For Figures 13(a,b), 14a, and 17(a,b) there were some minor changes in the values distance or time unit in the source data that impact the final values but the overall trend still remains the same. We no longer have the old version of the data so I was only able to share this new version which has such changes.*

During contact with the authors, it was stated that:

> *In addition, the code will take some time (especially Mobility Reports AGILE 2022.ipnyb and Outbreaks AGILE 2022.ipnyb).*

During reproduction, executing the mentioned notebooks took 23 minutes and 6 minutes respectively. The remaining two notebooks executed within seconds.

All reproduced Figures are included at the end of this report in section Reproduced Figures.

Technical remarks

All Jupyter Notebooks contain hard-coded paths to the required data files, which had to be adapted during reproduction.

Data files are not stored in the same repository and had to be fetched from a Google Drive folder provided by the authors.

During reproduction, several software requirements needed to be manually installed, as no `requirements.txt` file is given in the repository.

Reproduction of Figures 7 to 10 and Figure 16 required the aid of a Spreadsheet program for visualisation. For the sake of reproducibility, the open-source office suite LibreOffice was used. The spreadsheets and the visualisations generated during reproduction are available in the accompanying OSF repository in the zip archive `agile-reproreview-2022-023.zip`. Figures 13, 14, and 17, as well as the bar plots in Figure 7 and 16 are generated by the provided notebooks. Section Reproduced Figures lists the bar plots of Figures 7, 16 and tables of Figures 8 to 10 separately.

Across the provided notebooks, several smaller technical problems arose during reproduction. These comprise not-defined variables and invalid parameters to plot calls.

[1]The repository is located at: https://github.com/arunshar/COVID-19.

Comments to the authors

All notebook files contain hard-coded paths to the datasets, which had to be adjusted for reproduction. *I recommend using dynamic paths.* Python 3 provides the package `pathlib`, which offers the use of operating-system independent paths.

During reproduction, several smaller technical issues with the code, such as undefined variables or issues with invalid library calls, had to be resolved. These issues are likely to be leftovers from development. *I recommend testing code on an unrelated system* to catch and resolve such issues.

The authors provided information on the execution time only during communication. This information is not present in the publicly accessible repository. *I recommend adding information on the execution time to the repository.*

Several Python packages needed to be installed. *I recommend adding a `requirements.txt` file or similar* to provide an overview of the required software components.

Reproduction of Figures 7 to 10 and Figure 16 involved using a Spreadsheet program for visualisation. *I recommend automating all visualisation steps.*

Several data files are not available in the code repository and have to be fetched from a Google Drive provided by the authors. This may result in non-reproducible code if the Google Drive becomes inaccessible. One reason for using an alternative storage method is presumed to be the file size limitation of GitHub, currently at 50 MB. The largest file in the dataset is 6.12 GB in size. Using compression this can be reduced to 0.72 GB.[2] Data repositories such as Zenodo offer support for datasets up to 50GB in size. Alternatively, Git offers the feature Large File Storage (LFS), which is adopted by GitHub and currently supports storage of files up to 5 GB in size. *I recommend storing data and software in a stable and archive-friendly repository.*

Reproduced Figures

Using Notebook "Mobility Reports AGILE 2022" and LibreOffice Spreadsheets, Figures 7 to 10 and Figure 16 where reproducible. The numbers of tables of Figure 8, 9, and 10 are provided by the notebook.

Using Notebook "Outbreaks AGILE 2022", Figure 11 has been reproduced.

Using Notebook "Traffic Network AGILE 2022", Figure 13 has been reproduced.

Using Notebook "Mobility Assessment AGILE 2022", Figure 14a and 17 has been reproduced.

Table 1: Analysis of full-service restaurants visits based on the bucket dwell times - corresponds to table of Figure 8 in reproduced paper

dwelltime	visits
<5	416949
5-20	5446795
21-60	4416678
61-240	3867499
>240	1467323

Table 2: Figure 9. Analysis of Limited Service Restaurants based on the bucket dwell times - corresponds to table of Figure 9 in reproduced paper

dwelltime	visits
<5	592995
5-20	7262497
21-60	1522801
61-240	1000987

[2]The Python package `pandas`, which is used throughout the project, exposes the `compression` parameter in its `read_csv` and `to_csv` calls. This allows for transparent loading of compressed data files.

dwelltime	visits
>240	590133

Table 3: Analysis of bar visits based on bucket dwell times - corresponds to table of Figure 10 in reproduced paper

dwelltime	visits
<5	5646
5-20	73095
21-60	74254
61-240	81226
>240	25724

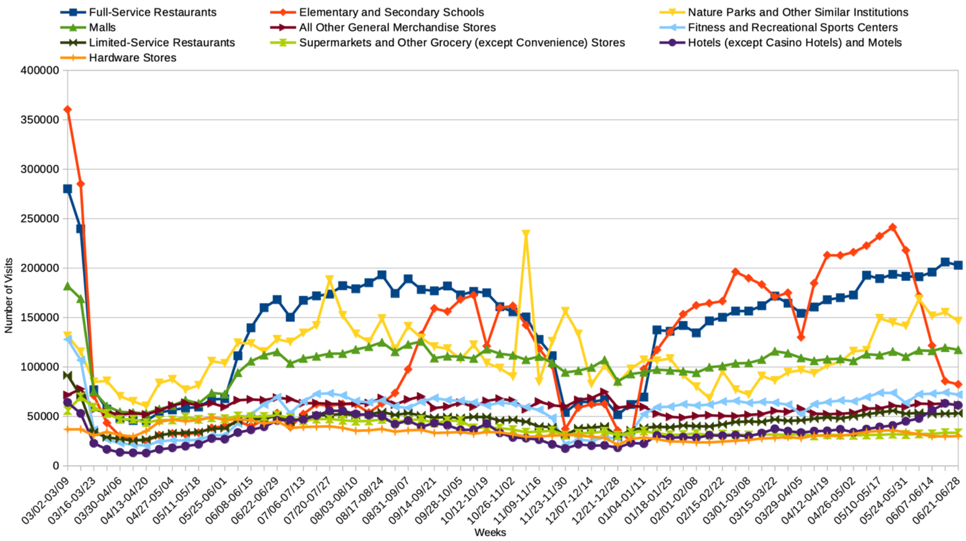

Figure 1: Most frequent long-duration visited business categories in Minnesota (March 2 ,2020 - June 28, 2021) - corresponds to Figure 7 in reproduced paper

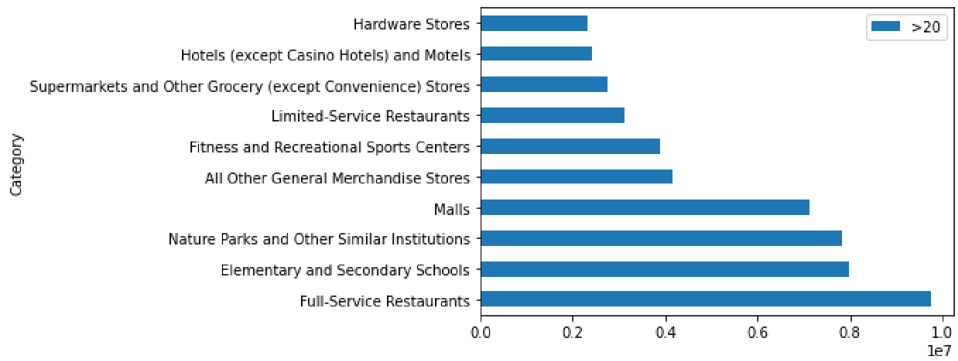

Figure 2: Most frequent long-duration visited business categories in Minnesota (March 2 ,2020 - June 28, 2021) - corresponds to the bar plot of Figure 7 in reproduced paper

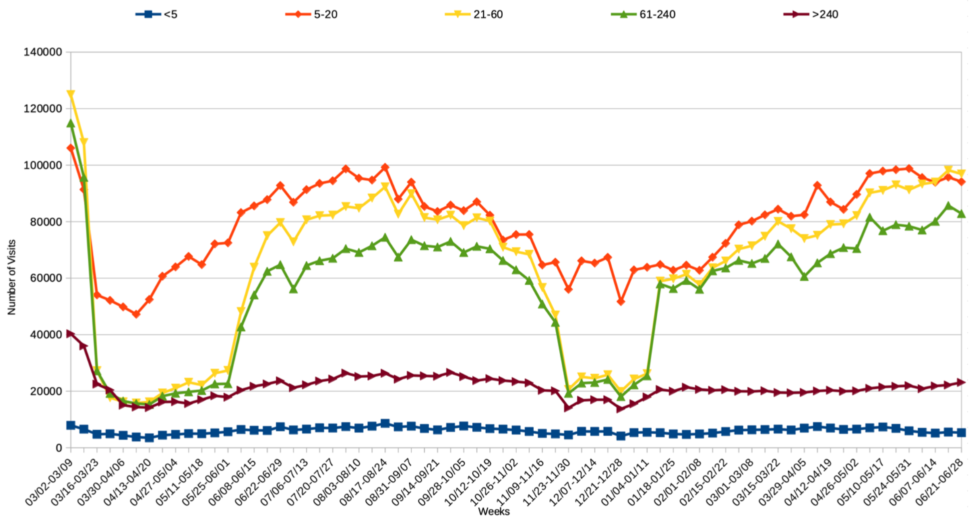

Figure 3: Analysis of full-service restaurants visits based on the bucket dwell times - corresponds to Figure 8 in reproduced paper

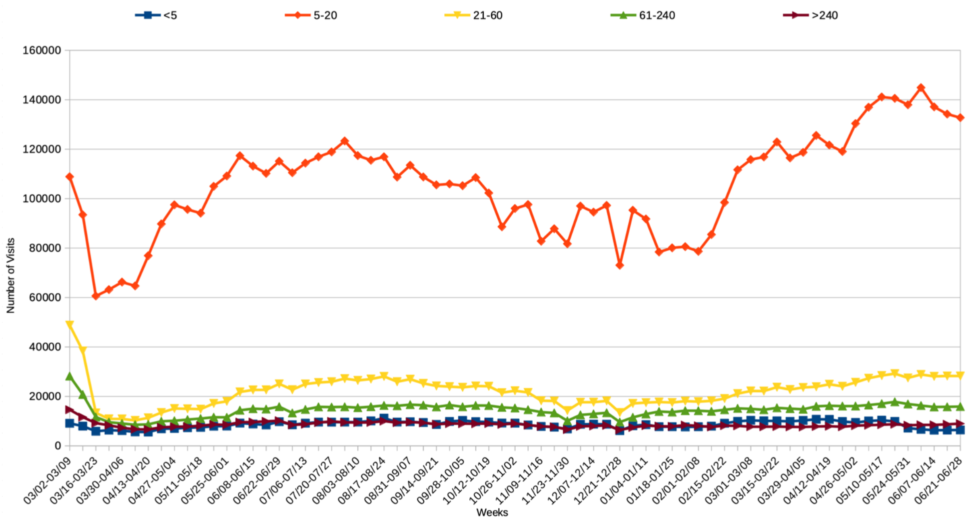

Figure 4: Figure 9. Analysis of Limited Service Restaurants based on the bucket dwell times - corresponds to Figure 9 in reproduced paper

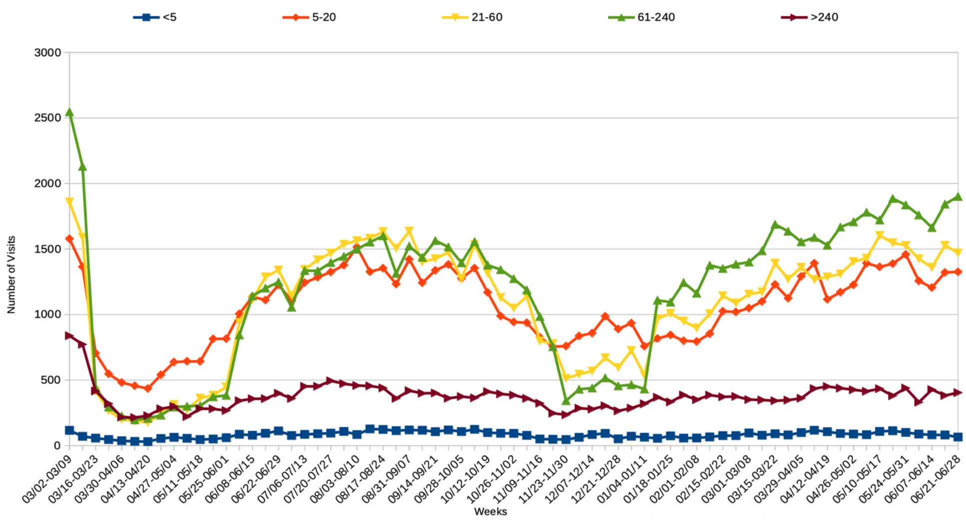

Figure 5: Analysis of bar visits based on bucket dwell times - corresponds to Figure 10 in reproduced paper

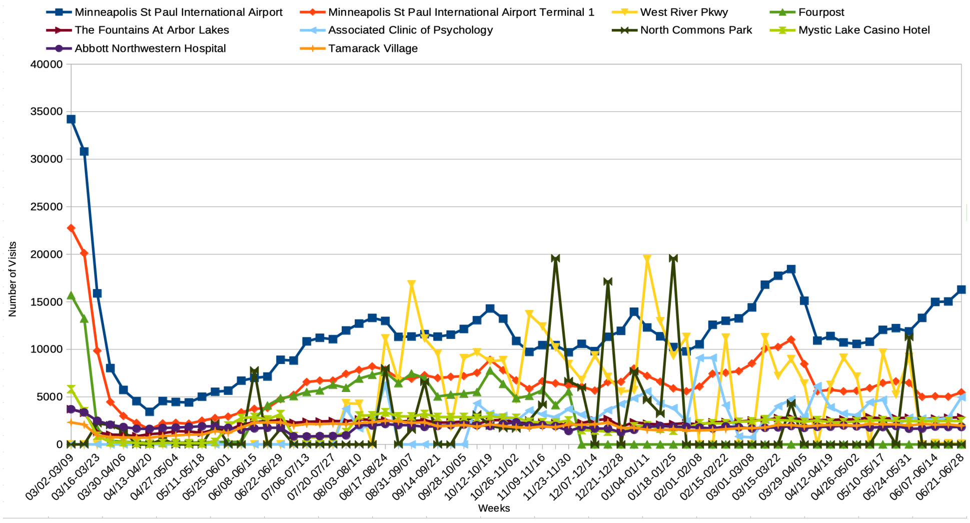

Figure 6: Most frequent long-duration visited Points of Interest in Minnesota (Since March 2, 2020 - June 28, 2021) - corresponds to Figure 16 in reproduced paper

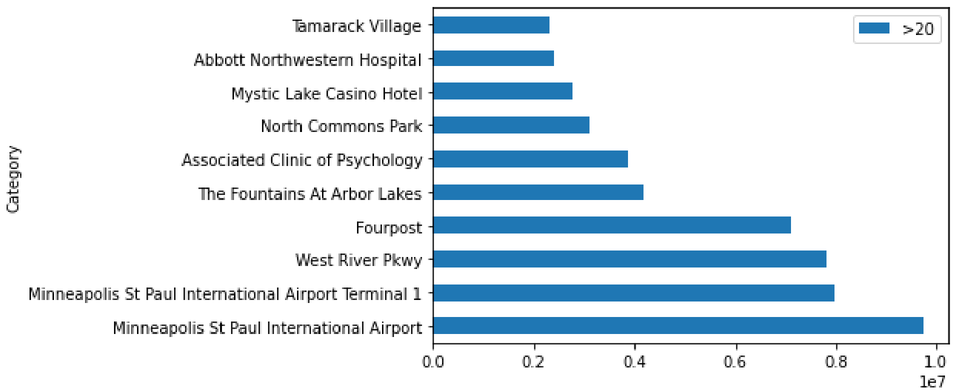

Figure 7: Most frequent long-duration visited Points of Interest in Minnesota (Since March 2, 2020 - June 28, 2021) - corresponds to the bar plot of Figure 16 in reproduced paper

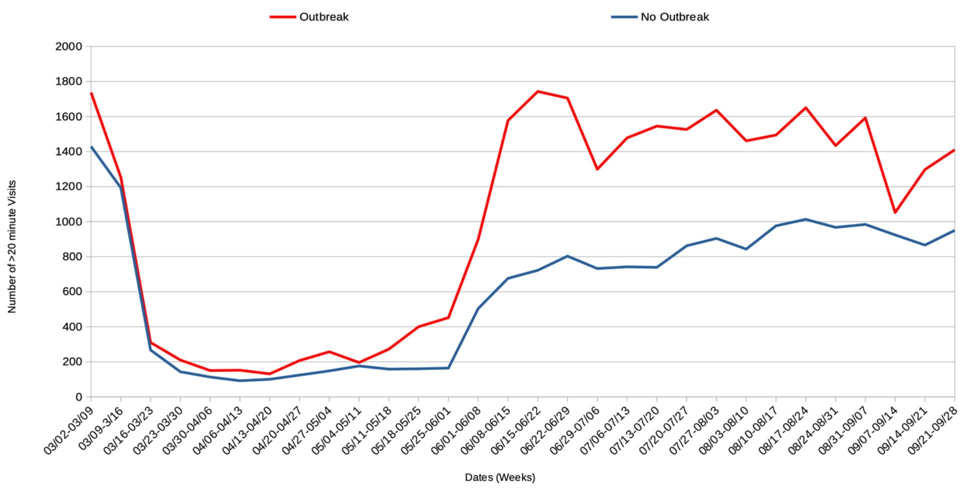

Figure 8: Long-duration visits for outbreak and non-outbreak groups - corresponds to Figure 11 in reproduced paper

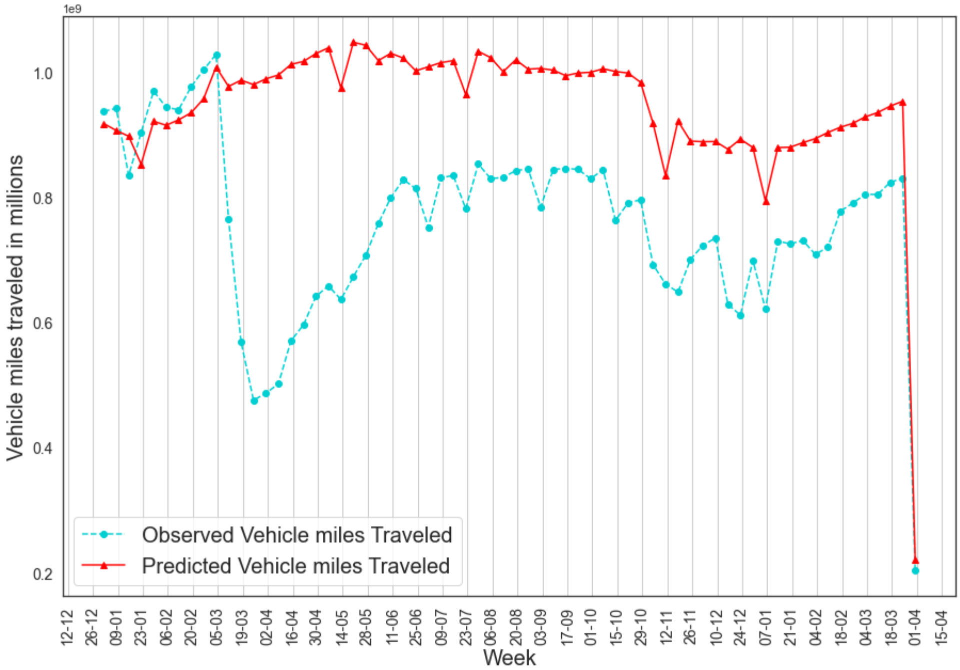

Figure 9: Decrease in traffic network across the COVID-19 timeline - corresponds to Figure 13a in reproduced paper

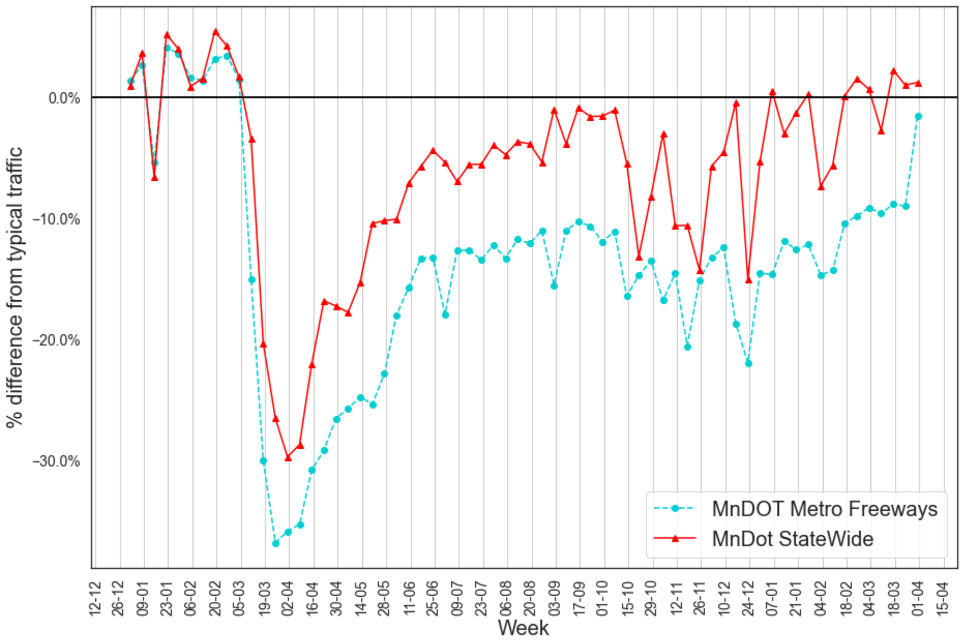

Figure 10: Decrease in traffic network across the COVID-19 timeline - corresponds to Figure 13b in reproduced paper

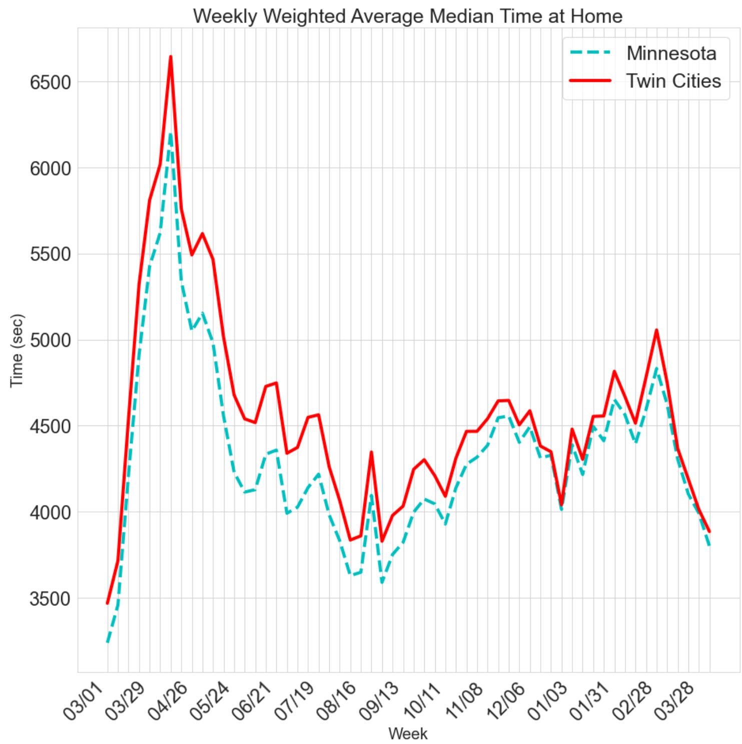

Figure 11: (a) Weekly average time spent at home, (b) Weekly average distance travel from home - corresponds to Figure 14a in reproduced paper

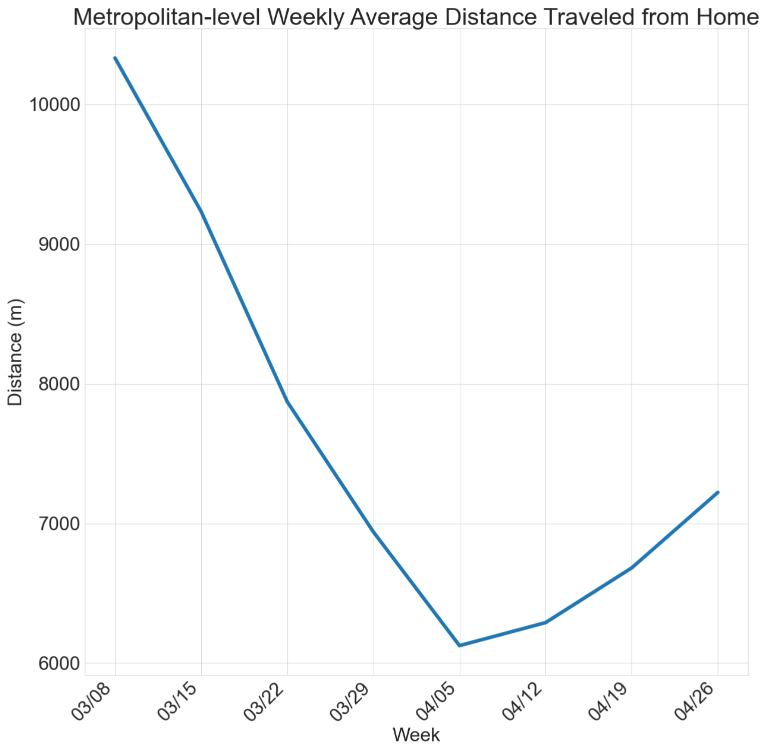

Figure 12: Mobility assessment in MN State between 03/02/2020 - 04/30/2020 - corresponds to Figure 17a in reproduced paper

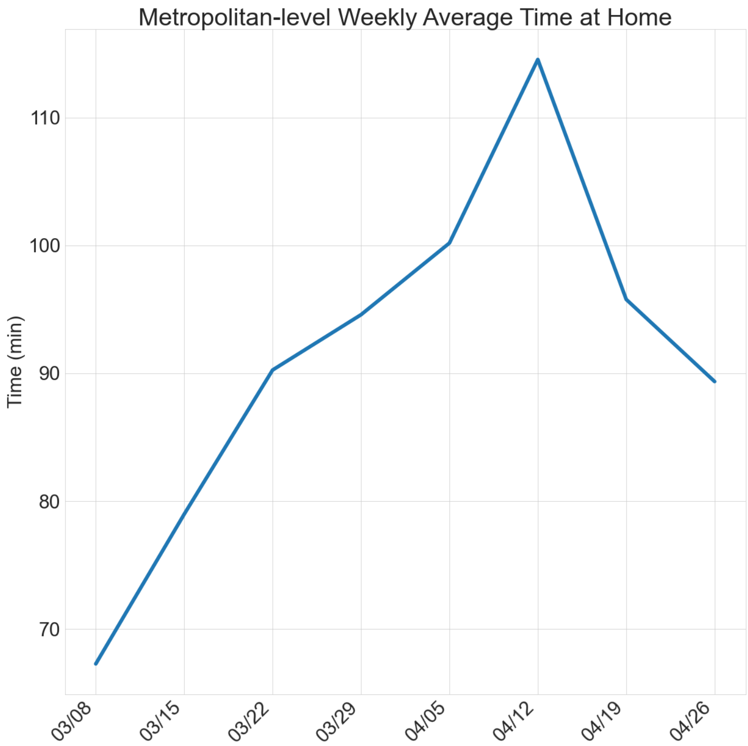

Figure 13: Mobility assessment in MN State between 03/02/2020 - 04/30/2020 - corresponds to Figure 17b in reproduced paper