Reproducibility review of: A Socially Aware Huff Model for Destination Choice in Nature-based Tourism

Jakub Krukar

2021-06-07

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. (2021, May 7). Reproducibility review of: A Socially Aware Huff Model for Destination Choice in Nature-based Tourism. https://doi.org/10.17605/OSF.IO/4CPM3

Reviewed paper

Shi, M., Janowicz, K., Cai, L., Mai, G., and Zhu, R.: A Socially Aware Huff Model for Destination Choice in Nature-based Tourism, AGILE GIScience Ser., 2, 14, https://doi.org/10.5194/agile-giss-2-14-2021, 2021.

Summary

The code, sample API query, and downloaded data were published in a public GitHub repository with a working Binder link. All files containing the code could be executed and all tables presented in the paper could be reproduced with only minor changes to the code. However, the code does not create figures contained in the paper and an attempt to change the results of the model evaluation by changing its numerical assumption was unsuccessful. The authors demonstrate concern for the reproducibility of their work and actively improved the reproducibility workflow throughout the reproducibility review process.

Reproducibility reviewer notes

The paper contains a Data and Software Availability Section with a link to the Flickr API and to a Github repository which I forked (https://github.com/reproducible-agile/Socially-aware-Huff-model). The Github repository contains a Binder link which I ran.

The code is divided into three files: `Trip Construction.ipynb`, `SA-Huff_model_Acadia.ipynb` and `SA-Huff_model_Yosemite.ipynb`. From the README file, it is not immediately clear: (a) that all files must be run in order to reproduce the results of the paper, (b) whether the sequence of running the files is important, and (c) which specific files reproduce specific parts of the paper. Issue (c) is clarified inside each file.

I have attempted the reproduction of the paper's results by following the Binder link (reviewed version: https://mybinder.org/v2/gh/meilinshi/Socially-aware-Huff-model/4dcdacf8e31fee16fe1c844935754f2531c7d231) and running all three Jupyter notebook files, as reported below. The result is a successful reproduction of partial (but most important) results of the paper, with minimal changes to the code necessary, such as (un-)commenting marked lines of code. After this review, the authors made significant improvement to the code. Issues that were improved are crossed out.

File Trip Construction.ipynb

File `Trip Construction.ipynb` can be used to reproduce Tables 1, 8, and 9.

~~Table 1 in the paper corresponds to two separate output commands in the code and the file does not describe the fact that a line in chunk [3] needs to be un-commented in order to obtain these values~~:

```
table_stats(acadia_ttl)
#table_stats(yosemite_ttl)
```

Nevertheless, after changing the comment hash sign, all values in Table 1 are reproduced.

For Table 8, there are more columns in the table generated by the code compared to the table in the paper, however, I was able to reproduce and identify values in Table 8.

~~Obtaining corresponding values for Table 9 (Yosemite National Park) required un-commenting two lines of code and commenting two existing lines of code in chunk [4]. This can lead to a user error~~, e.g.:

```
#input_url = acadia_url
input_url = yosemite_url

position_url = acadia_position
#position_url = yosemite_position
```

which in turn breaks the execution of the code with the following error:

`ValueError: Length of values (13) does not match length of index (21).`

Nevertheless, with correct un-commenting, I was able to reproduce values of Table 9.

File SA-Huff_model_Acadia.ipynb

This file can be used to reproduce the Acadia related part of Tables 2, 3, and 7, as well as the entire Table 4. Headings make it clear which code chunks reproduce which table. I was able to obtain identical results for Tables 2, 3, 4, and 7, with identical number, order and naming of table columns, compared to those in the paper.

File SA-Huff_model_Yosemite.ipynb

This file can be used to reproduce the Yosemite related part of Tables 2, 3, and 7, as well as the entire Table 5 and 6. Headings make it clear which code chunks reproduce which table. I was able to obtain identical results for Tables 2, 3, 5, and 7, with identical number, order and naming of table columns, compared to those in the paper. Table 6 returned identical numerical results, but with a different ordering of table rows, compared to the paper.

Changing assumptions of the model

The paper describes an assumption that 2 photos contributed by the same user more than 4 days apart from each other are coming from 2 separate visits. One scientific reviewer commented on this assumption as reasonable but debatable. I have attempted changing the threshold value for this assumption from 4 days to 10 days. This should result in a different number of trips generated by the file `Trip Construction.ipynb` and can be verified by comparing the values from the output of chunk [12] to Table 8 in the paper. I expected that the values in the `Number of photos` column would stay the same (because the number of photos does not change), but the values in `outgoing/incoming trips` columns would change.

The location of the threshold variable is marked with an in-code comment in the file `Trip Construction.ipynb`:

```
if length.days > 4: #time threshold: average length of stay in both NPs
```

Table 8 Summary of attractions in Acadia National Park

Attraction	Number of photos	Outgoing trips	Incoming trips
Schoodic Institute	1119	53	64
Bass Harbor	2298	260	288
Southwest Harbor	723	109	111
Northeast Harbor	605	67	76
Bar Harbor	6259	433	357
Wild Gardens of Acadia	550	60	66
Cadillac Mountain	3285	349	345
Penobscot Peak	776	16	15
Bubble Rock	703	83	89
Jordan Pond	1250	227	250
Boulder Beach	536	85	102
Thunder Hole	977	167	185
Sand Beach	1253	216	177

Figure 1: Original Table 8 from the paper.

Places	a	b	c	d	e	f	g	h	i	j	k	l	m	total_out	total_in	cross_boundary	photos
Schoodic Institute	0	13	7	1	12	1	8	0	0	4	3	2	6	57	66	123	1119
Bass Harbor	12	0	34	9	64	13	53	4	6	25	12	15	21	268	295	563	2298
Southwest Harbor	3	44	0	6	30	3	15	4	1	4	1	2	2	115	117	232	723
Northeast Harbor	5	16	8	0	13	1	7	0	2	10	1	2	3	68	78	146	605
Bar Harbor	20	60	25	21	0	17	118	3	12	50	15	40	56	437	367	804	6259
Wild Gardens of Acadia	1	3	1	2	10	0	6	1	1	6	4	11	15	61	67	128	550
Cadillac Mountain	8	57	12	13	102	16	0	0	14	51	12	24	45	354	350	704	3285
Penobscot Peak	2	3	3	2	2	0	0	0	0	2	0	1	1	16	15	31	776
Bubble Rock	1	16	5	0	19	1	13	2	0	17	3	3	5	85	89	174	703
Jordan Pond	6	36	4	10	50	5	53	1	44	0	4	8	12	233	255	488	1250
Boulder Beach	3	11	7	4	10	3	18	0	0	19	0	4	9	88	104	192	536
Thunder Hole	1	17	5	3	26	3	28	0	5	42	32	0	7	169	187	356	977
Sand Beach	4	19	6	7	29	4	31	0	4	25	17	75	0	221	182	403	1253

Figure 2: Reproduced table corresponding to Table 8, after changing the threshold value for length.days from 4 to 10.

The result (Fig. 1 and 2) demonstrates that the number of outgoing/incoming trips from/to all but one (Penobscot Peak) attractions in the Acadia National Park *increased* after changing the temporal threshold.

I assume this is due to more sequences being classified as belonging to a single visit, and therefore a higher potential number of trips between any two attractions within a single visit of a single user.

In the next step I attempted re-calculating the output of the model evaluation after changing the number of days threshold. ~~The file `Trip Construction.ipynb` does not clearly describe how to do this~~ but the last chunk of the code contains a commented line that I un-commented and re-ran the script:

```
# split the trips and generate probability matrix for each month
# the output is provided in the data folder --> acadia_pmatrix_example

for i in range(1,13):
    df_sub = pd.DataFrame()
    df_sub = split_fmatrix(NP_trips,i)
    pmatrix_sub = prob_matrix(df_sub)
    #pmatrix_sub.to_csv('acadia_NP_cluster_prob_matrix_'+str(i)+'.csv')
```

This resulted in generating 12 files named `acadia_NP_cluster_prob_matrix_(...).csv`. ~~However, the files were saved to the `./Code` directory, not to `./Data/acadia_pmatrix` directory (which is the correct location of the files, despite the in-code comment pointing to 'acadia_pmatrix_example'). I moved the files manually and re-ran the file `SA-Huff_model_Acadia.ipynb`. There was no change to the output compared to the files originally stored in `./Data/acadia_pmatrix` and reported in the paper. I repeated the procedure making a different change to the number of days threshold (from 4 to 1) that resulted in greater differences to the number of trips. Re-running `SA-Huff_model_Acadia.ipynb` did not generate a different output.~~

Figures

The paper contains four Figures, two of which (Paper Figure 1 and 2) display the output of data processing.

Paper Figure 1 displays 'photo clusters detected by HDBSCAN in the two national parks'. The repository does not contain code to generate the figure and the README file does not describe which data file was used to generate it.

Paper Figure 2 was generated using the open source flowmap.blue (https://flowmap.blue/) service. The README file contains links that lead to the flowmap.blue website and re-create the figure from the paper. The reproducibility of this output is however dependent on the future changes to the flowmap.blue service. The README file does not describe how to re-generate data files that are used as input in the flowmap.blue link.

Runtime

The repository does not contain information about expected runtime. For the steps described above, the runtime on a standard laptop was the longest for re-running the `SA-Huff_model_Acadia.ipynb` (under 4 minutes).

Communication with the authors

Communication with the authors throughout writing this report has been exemplary. The corresponding author responded promptly to concerns by directly improving the code on the GitHub repository.

Conclusion

This is a very good example of a partially reproducible paper. Most of results presented in the paper could be reproduced with very little changes to the provided code, directly in the Binder environment running in the web browser. The repository is potentially a very useful supplement to the paper for researchers interested in the topic. I provide suggestions for improvement which in great majority concern better documentation of the workflow.

Comments to the Authors

- ~~For someone unfamiliar with Binder it would be very useful to include a few explanatory sentences around the Binder link, mentioning, e.g., that the code can be reproduced in a web browser, without downloading the code or software. This is the feature that can be used the easiest by less competent programmers, but has the least documention in the README file.~~
- The README could better describe the steps to reproduce the workflow and explain what the person can expect (i.e., which Figures/Tables can be reproduced and which cannot). Some additional information is included within Jupyter notebooks but the lack of overview in README makes it difficult to understand the relation of the repository to the paper.
- ~~The format of comments particularly inside `Trip Construction.ipynb` is confusing, as some of them are provided at the beginning of the file, some as headers, and some as comments within code chunks. Following a consistent convention would be helpful (e.g., clear headers like in the remaining files).~~
- The repository contains data files that cannot be generated using the attached code. These could be better described in README (i.e., a list of data files that are in the repository but are not generated by the code, some description on how they were generated, and whether there is a relation between files generated by `Trip Construction.ipynb` and those loaded by `SA-Huff_model_Acadia.ipynb` and `SA-Huff_model_Yosemite.ipynb`).
- ~~In order to make it possible to manually modify model's assumptions, the README could describe how to use the re-generated trip sequences to re-calculate the model evaluation statistics.~~