Reproducibility review of: Process Analysis in Humanitarian Voluntary Geographic Information: the case of the HOT Tasking Manager

Rémy Decoupes

2024-05-27

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 https://osf.io/fmgb4/). To cite the report use

> Decoupes R, (2024, April): Reproducibility review of: Process Analysis in Humanitarian Voluntary Geographic Information: the case of the HOT Tasking Manager. https://doi.org/10.17605/OSF.IO/fmgb4

Reviewed paper

Herrera-Murillo, D.J., Ochoa-Ortiz, H., Ahmed, U., López-Pellicer, F.J., Re, B., Polini, A., and Nogueras-Iso, J.: Process Analysis in Humanitarian Voluntary Geo-graphic Information: the case of the HOT Tasking Manager, AGILE GIScience Ser., 5, 5, https://doi.org/10.5194/agile-giss-5-5-2024, 2024

Summary

The authors made their code available through a GitHub repository. The Readme.md file provides clear instructions for executing the entire code. It is worth noting that the input data is available provided that accounts are created with HOTOSM and Bunting Labs. The data retrieval process takes a long time. I initiated the retrieval process for 24 hours but only managed to receive one-third of the data. Therefore, the authors provided me with all the input data for my review.

I did not encounter any difficulties in executing the Python notebooks. As a result, I produced intermediate data (in 4 CSV files). Unfortunately, I encountered some blocking errors when executing the R Markdown (which is the notebook responsible for producing tables and figures of the manuscript).

Assisted by the authors, we discovered that the errors were caused by the content of the intermediate files I had generated. Strictly speaking, since the authors shared all their input data with me, I should not have encountered these errors.

To continue my review, the authors sent me their version of the intermediate files. Thanks to them, I was able to execute the entire R Markdown and generate all the tables and figures. At the time of finalizing this report, I do not know why the generated intermediate files were different from those of the authors. I cannot affirm that the Python notebooks produce intermediate data of good quality.

Reproducibility reviewer notes

The authors shared their code through a GitHub repository. The Readme.md file clearly explains the prerequisites and the steps to follow to reproduce the figures and tables. Four notebooks are provided, the 3 firsts in Python and the last one in R Markdown:

1. Download_data.ipynb
2. Build_density.ipynb
3. Data_preprocessing.ipynb
4. Process_Analysis.Rmd

What's more, to download the data, two accounts need to be created to activate API keys: OpenStreetMap and BuntingLags.

I ran the reproducibility review on Linux (Debian 10) with 32GB RAM.

Python environment installation

First, I created a new Conda environment with dependencies as shared in the Readme.md file.

```
conda create -n agile-21 python=3.10.2 ipython pip ipykernel
conda activate agile-21

python ipykernel install --user --name=agile-21

pip install pandas==1.5.3
pip install geopandas==0.13.2
pip install numpy==1.25.2
pip install requests==2.31.0
pip install json==2.0.9
pip install ipywidgets==7.7.1
pip install tqdm==4.66.1
pip install utm==0.7.0
```

Then I ran the notebooks sequentially.

1. Download_data.ipynb

I followed the instructions to create the two tokens API from: - HOTOSM Tasking Manager API. I had to create first an OpenStreetMap account and authorize HOTOSM to access to my OpenStreetMap account. I faced an issue with Firefox, the redirection from OpenStreetMap to HOTOSM was not working but it worked with Chromium. - Buntinglabs: I just provided my e-mail address.

I create manually a directory `data` in which the notebook will save the collected data.

I had to stop my computer during the cell "Get project activities" after 20 hours of run. But when I re-start the notebook, the cell only downloaded the missing file (it was not trying to download already downloaded files). Unfortunately, even after 24 hours of execution, I had barely exceeded the 1/3 of the necessary data. So, to continue my review within the allotted time, I asked the authors to provide me with their input data.

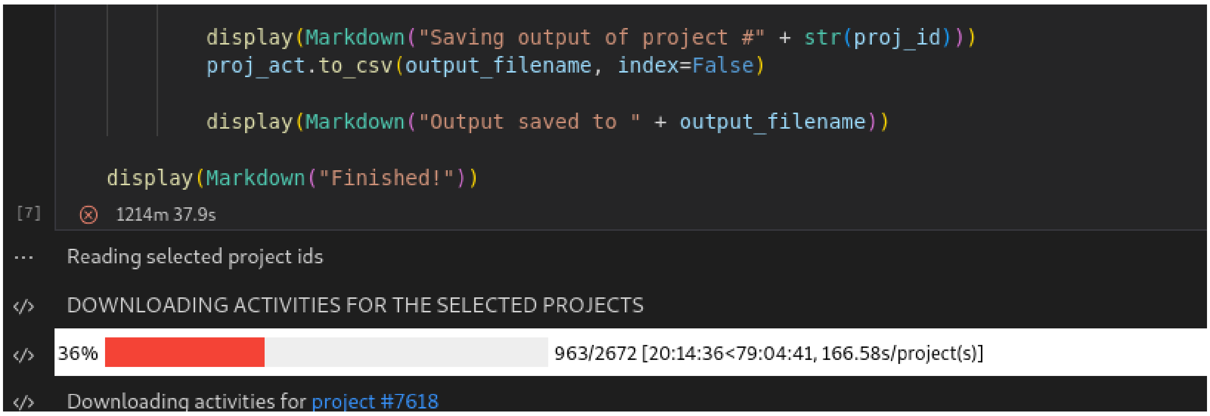
The authors granted me access to all the datasets that this notebook retrieves, as I have an account and API key from HOT and Bunting Labs. This should be compliant with their terms of use.

2. Building Density

I had to setting a path to the output folder at the beginning of the notebook. At the end, it creates a csv file

3. Data preprocessing

- Missing installation of matplotlib: `pip install matplotlib`



```
            display(Markdown("Saving output of project #" + str(proj_id)))
            proj_act.to_csv(output_filename, index=False)

            display(Markdown("Output saved to " + output_filename))

    display(Markdown("Finished!"))
```

```
[7]  1214m 37.9s
...  Reading selected project ids
     DOWNLOADING ACTIVITIES FOR THE SELECTED PROJECTS
     36% 963/2672 [20:14:36<79:04:41, 166.58s/project(s)]
     Downloading activities for project #7618
```

Figure 1: agile-21-get-activity_too-long

- Error in the path to the file output_densities.csv in section `3. Create file for regression analysis with task density ("regression.csv.csv)`, l1 `densities=pd.read_csv("output_densities.csv"` should be `densities=pd.read_csv(data_folder + "output_densities.csv")` as data_folder containing output_densities.csv has been set in notebook `building_densitiy.ipynb`

At this end, this notebook created 4 csv files: - project.csv - initial_tasks.csv - contributors.csv - regression.csv

4. Process Analysis

As the notebook is a Rmarkdown, and in order to have the same R kernel as recommended by the authors, I used a docker image built by the Rocker Project. The image `rocker/rstudio:4.3.2` contains an RStudio server with R version 4.3.2. I mount my local directory (with all the code and data for reproducing this paper) into `/agile` inside the container.

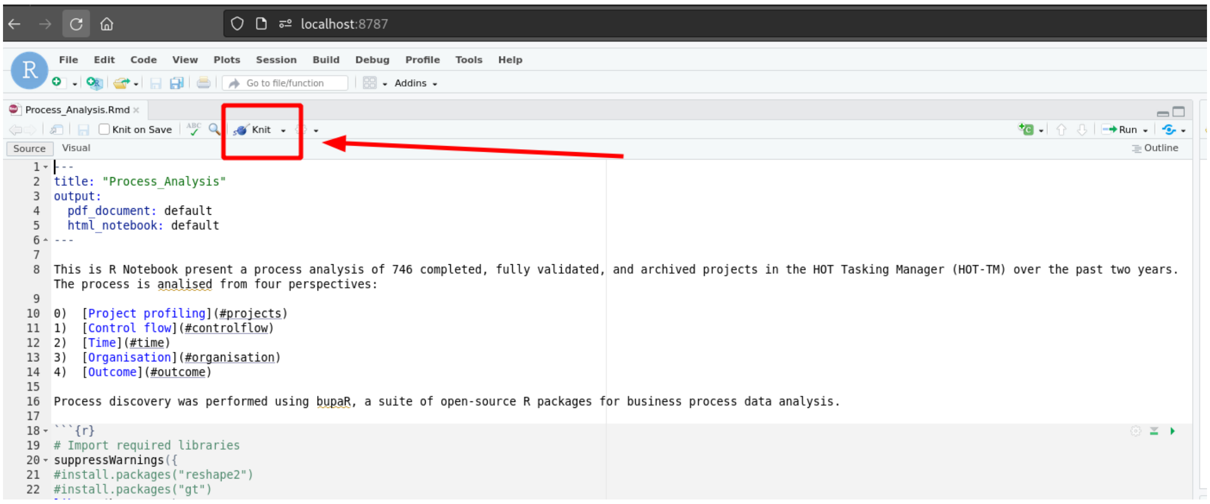
```
sudo docker run -e ROOT=true -e PASSWORD=agile --rm -p 8787:8787 -v ./:/agile rocker/rstudio:4.3.2
```

When the Docker container is started, I connect to RStudio with a web browser: http://localhost:8787, with credentials as follows:

- login: rstudio
- password: agile

Then I installed the dependencies as listed in the Readme.md:

```
install.packages("bupaverse", version = "0.1.0")
install.packages("reshape2", version = "1.4.4")
install.packages("gt", version = "0.10.1")
install.packages("scales", version = "1.3.0")
install.packages("readr", version = "2.1.4")
install.packages("dplyr", version = "1.1.4")
install.packages("magrittr", version = "2.0.3")
install.packages("ggplot2", version = "3.4.4")
install.packages("Hmisc", version = "5.1-1")
install.packages("gamlss", version = "5.4-20")
```



~~~
---
title: "Process_Analysis"
output:
  pdf_document: default
  html_notebook: default
---

This is R Notebook present a process analysis of 746 completed, fully validated, and archived projects in the HOT Tasking Manager (HOT-TM) over the past two years.
The process is analised from four perspectives:

0)  [Project profiling](#projects)
1)  [Control flow](#controlflow)
2)  [Time](#time)
3)  [Organisation](#organisation)
4)  [Outcome](#outcome)

Process discovery was performed using bupaR, a suite of open-source R packages for business process data analysis.

```{r}
# Import required libraries
suppressWarnings({
#install.packages("reshape2")
#install.packages("gt")
~~~

Unfortunately I encountered several errors some of them were blocking:

- Missing installation of library:

```
Execution halted

No LaTeX installation detected (LaTeX is required to create PDF output). You should install a LaTeX distribution for your platform: https://www.latex-project.org/get/

  If you are not sure, you may install TinyTeX in R: tinytex::install_tinytex()

  Otherwise consider MiKTeX on Windows - http://miktex.org

  MacTeX on macOS - https://tug.org/mactex/
  (NOTE: Download with Safari rather than Chrome _strongly_ recommended)

  Linux: Use system package manager
```

So I installed tinytex:

```
tinytex::install_tinytex()
```

- In line 177, I got this error

```
Error in `group_by()`:
! Must group by variables found in `.data`.
 Column `mappingLevel` is not found.
Backtrace:
 1. event_log_df %>% group_by(mappingLevel) %>% ...
 4. dplyr:::group_by.data.frame(., mappingLevel)
```

It seems that in `event_log_df`, the correct name of the column is `mappingLevel.x` instead of `mappingLevel`.

I replaced by

```
mappingLevel <- event_log_df %>%  group_by(mappingLevel.x) %>% summarise(count = n_distinct(actionBy))
```

- In line 185, I got this error

```
Error in xtfrm.data.frame(x) : cannot xtfrm data frames
```

It seems that the dcast is missing a dimension (`data_pivot <- dcast(event_log_df, action ~ mappingLevel,value.var = "taskId", length)`). I could not find a fix but since `data_pivot` is not further used, I ignore this error

- In line 209, I got this error

```
Error in gamlss(percentage_area_covered_by_building ~ splits1 + invalidations1 +  :
  could not find function "gamlss"
```

I installed gamlss: `install.packages("gamlss")` and I had import it the cell in l.209: `library(gamlss)`.

- But then, I encountered this error that I could not bypass

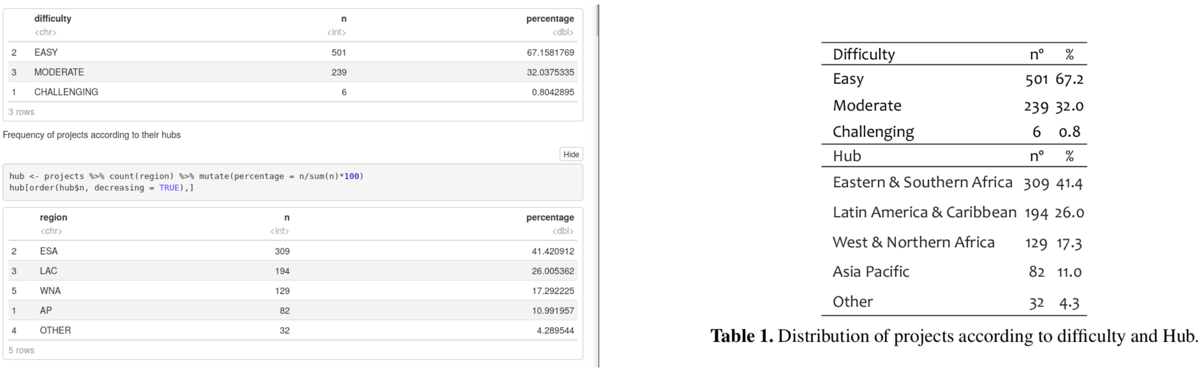
```
Error in `contrasts<-`(`*tmp*`, value = contr.funs[1 + isOF[nn]]) :
  contrasts can be applied only to factors with 2 or more levels
```

As I was blocked by this issue, the Authors helped me to debug. They compared my intermediate CSV files generated by the three python notebooks. It appeared that they had significant differences affecting two file: regression.csv and initial_tasks that they were generated by the 3rd python notebook. In writing this report, I'm unsure where these differences might originate. Is it due to the input data, or perhaps there's an issue with the Python notebooks?

In order to test the generation of the figures and tables, I asked the authors to also send me the intermediate files. I was then able to execute the entire notebook without any errors.

Here in details the tables and figures I was able to reproduce

- Table 1:



	difficulty <chr>	n <int>	percentage <dbl>
2	EASY	501	67.1581769
3	MODERATE	239	32.0375335
1	CHALLENGING	6	0.8042895

3 rows

Frequency of projects according to their hubs

Hide

```r
hub <- projects %>% count(region) %>% mutate(percentage = n/sum(n)*100)
hub[order(hub$n, decreasing = TRUE),]
```

	region <chr>	n <int>	percentage <dbl>
2	ESA	309	41.420912
3	LAC	194	26.005362
5	WNA	129	17.292225
1	AP	82	10.991957
4	OTHER	32	4.289544

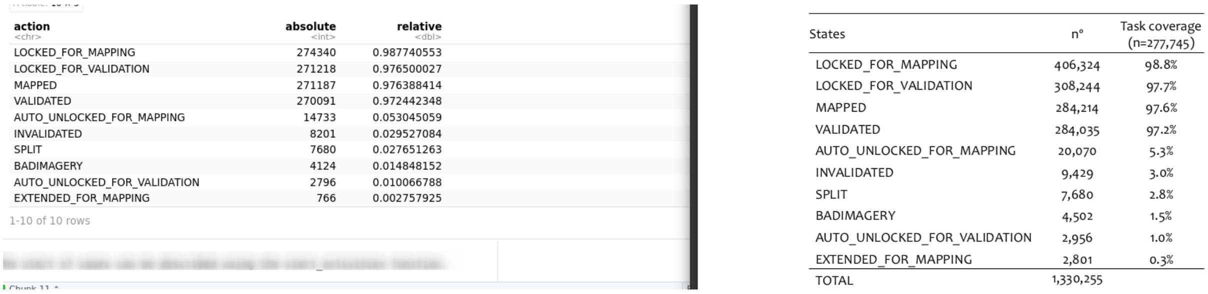
5 rows

Difficulty	n°	%
Easy	501	67.2
Moderate	239	32.0
Challenging	6	0.8
Hub	n°	%
Eastern & Southern Africa	309	41.4
Latin America & Caribbean	194	26.0
West & Northern Africa	129	17.3
Asia Pacific	82	11.0
Other	32	4.3

Table 1. Distribution of projects according to difficulty and Hub.

Figure 2: Reproduction of table 1

- Table 3:



action <chr>	absolute <int>	relative <dbl>
LOCKED_FOR_MAPPING	274340	0.987740553
LOCKED_FOR_VALIDATION	271218	0.976500027
MAPPED	271187	0.976388414
VALIDATED	270091	0.972442348
AUTO_UNLOCKED_FOR_MAPPING	14733	0.053045059
INVALIDATED	8201	0.029527084
SPLIT	7680	0.027651263
BADIMAGERY	4124	0.014848152
AUTO_UNLOCKED_FOR_VALIDATION	2796	0.010066788
EXTENDED_FOR_MAPPING	766	0.002757925

1-10 of 10 rows

States	n°	Task coverage (n=277,745)
LOCKED_FOR_MAPPING	406,324	98.8%
LOCKED_FOR_VALIDATION	308,244	97.7%
MAPPED	284,214	97.6%
VALIDATED	284,035	97.2%
AUTO_UNLOCKED_FOR_MAPPING	20,070	5.3%
INVALIDATED	9,429	3.0%
SPLIT	7,680	2.8%
BADIMAGERY	4,502	1.5%
AUTO_UNLOCKED_FOR_VALIDATION	2,956	1.0%
EXTENDED_FOR_MAPPING	2,801	0.3%
TOTAL	1,330,255	

Figure 3: Reproduction of table 3

- Figure 7: Note that there are differences between those two graph, it's because I generated mine on a subset of data.

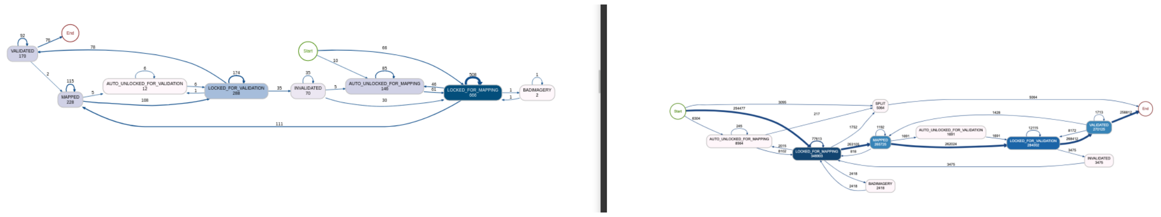

End
VALIDATED 170
Start
MAPPED 228
AUTO_UNLOCKED_FOR_VALIDATION 12
LOCKED_FOR_VALIDATION 288
INVALIDATED 70
AUTO_UNLOCKED_FOR_MAPPING 146
LOCKED_FOR_MAPPING 606
BADIMAGERY 2
Start
SPLIT 5064
End
AUTO_UNLOCKED_FOR_MAPPING 8564
LOCKED_FOR_MAPPING 346903
MAPPED 265725
AUTO_UNLOCKED_FOR_VALIDATION 1691
LOCKED_FOR_VALIDATION 284002
VALIDATED 270125
INVALIDATED 3475
BADIMAGERY 2418

Figure 4: Reproduction of fig 7

- figure 8:
- Table 4: It seems that there is a little difference between the tables. This may be due to a difference in rounding or truncation.
- Table 5:
- Figure 9:
- Figure 10:

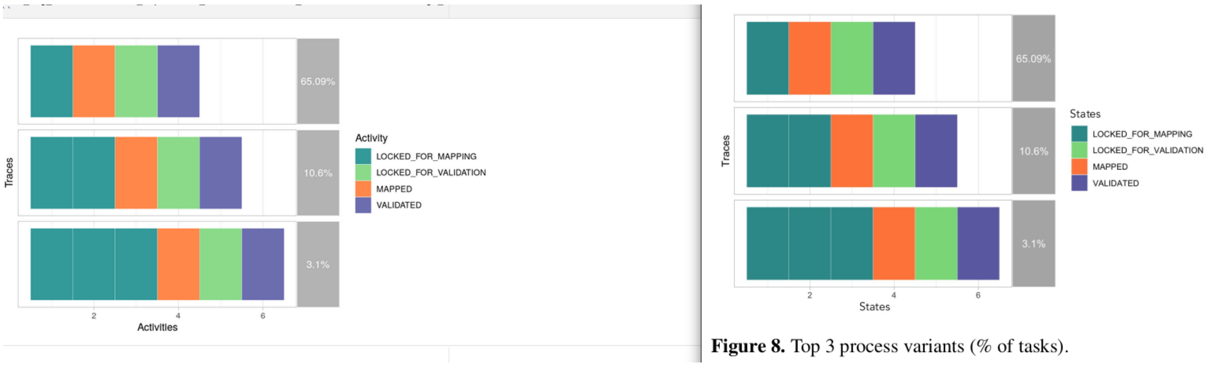
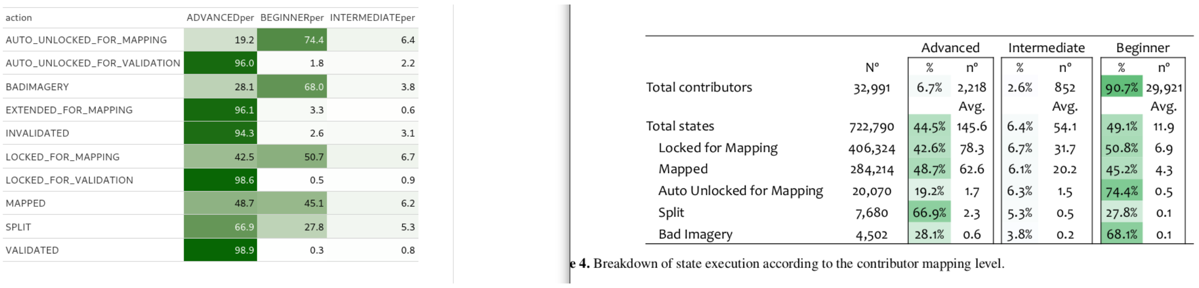


Figure 8. Top 3 process variants (% of tasks).

Figure 5: Reproduction of fig 8

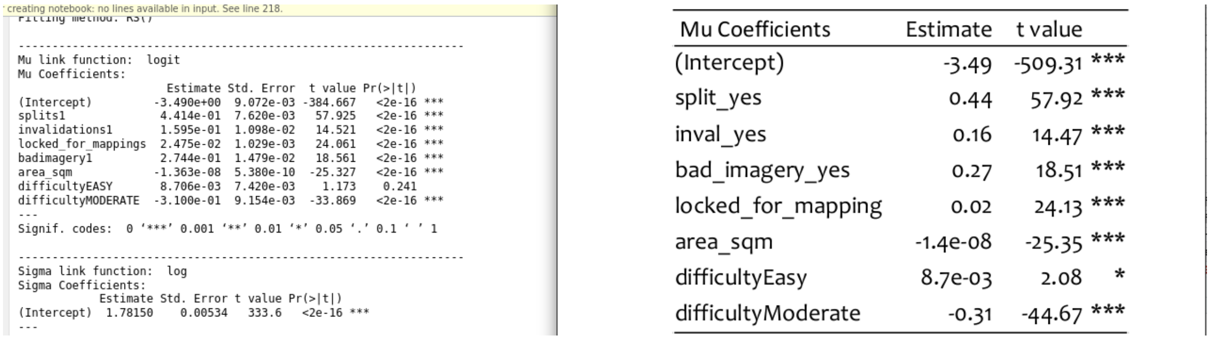


action	ADVANCEDper	BEGINNERper	INTERMEDIATEper
AUTO_UNLOCKED_FOR_MAPPING	19.2	74.4	6.4
AUTO_UNLOCKED_FOR_VALIDATION	96.0	1.8	2.2
BADIMAGERY	28.1	68.0	3.8
EXTENDED_FOR_MAPPING	96.1	3.3	0.6
INVALIDATED	94.3	2.6	3.1
LOCKED_FOR_MAPPING	42.5	50.7	6.7
LOCKED_FOR_VALIDATION	98.6	0.5	0.9
MAPPED	48.7	45.1	6.2
SPLIT	66.9	27.8	5.3
VALIDATED	98.9	0.3	0.8

		Advanced		Intermediate		Beginner	
	N°	%	n°	%	n°	%	n°
Total contributors	32,991	6.7%	2,218	2.6%	852	90.7%	29,921
			Avg.		Avg.		Avg.
Total states	722,790	44.5%	145.6	6.4%	54.1	49.1%	11.9
Locked for Mapping	406,324	42.6%	78.3	6.7%	31.7	50.8%	6.9
Mapped	284,214	48.7%	62.6	6.1%	20.2	45.2%	4.3
Auto Unlocked for Mapping	20,070	19.2%	1.7	6.3%	1.5	74.4%	0.5
Split	7,680	66.9%	2.3	5.3%	0.5	27.8%	0.1
Bad Imagery	4,502	28.1%	0.6	3.8%	0.2	68.1%	0.1

e 4. Breakdown of state execution according to the contributor mapping level.

Figure 6: Reproduction of Table 4



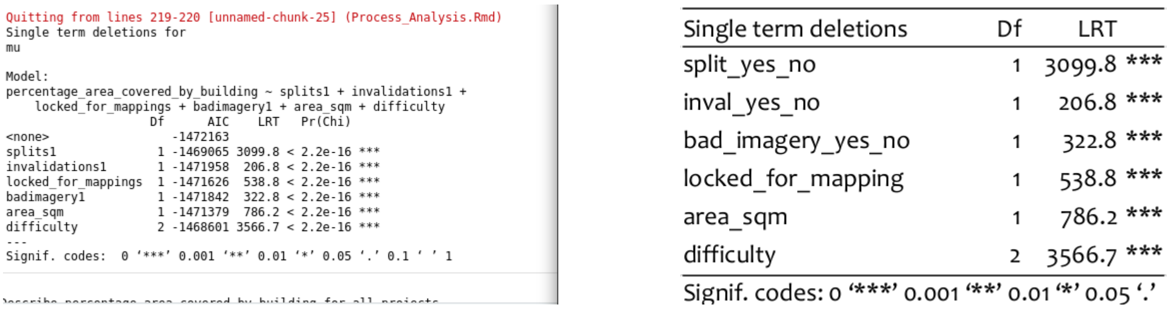
```
creating notebook: no lines available in input. See line 218.
Fitting method: RS()

-------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)         -3.490e+00  9.072e-03 -384.667   <2e-16 ***
splits1              4.414e-01  7.620e-03   57.925   <2e-16 ***
invalidations1       1.595e-01  1.098e-02   14.521   <2e-16 ***
locked_for_mappings  2.475e-02  1.029e-03   24.061   <2e-16 ***
badimagery1          2.744e-01  1.479e-02   18.561   <2e-16 ***
area_sqm            -1.363e-08  5.380e-10  -25.327   <2e-16 ***
difficultyEASY       8.706e-03  7.420e-03    1.173    0.241
difficultyMODERATE  -3.100e-01  9.154e-03  -33.869   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.78150    0.00534   333.6   <2e-16 ***
---
```

Mu Coefficients	Estimate	t value
(Intercept)	-3.49	-509.31 ***
split_yes	0.44	57.92 ***
inval_yes	0.16	14.47 ***
bad_imagery_yes	0.27	18.51 ***
locked_for_mapping	0.02	24.13 ***
area_sqm	-1.4e-08	-25.35 ***
difficultyEasy	8.7e-03	2.08 *
difficultyModerate	-0.31	-44.67 ***

Figure 7: Reproduction of the 1rst part of Table 5



```
Quitting from lines 219-220 [unnamed-chunk-25] (Process_Analysis.Rmd)
Single term deletions for
mu

Model:
percentage_area_covered_by_building ~ splits1 + invalidations1 +
    locked_for_mappings + badimagery1 + area_sqm + difficulty
                    Df      AIC    LRT   Pr(Chi)
<none>                 -1472163
splits1              1 -1469065 3099.8 < 2.2e-16 ***
invalidations1       1 -1471958  206.8 < 2.2e-16 ***
locked_for_mappings  1 -1471626  538.8 < 2.2e-16 ***
badimagery1          1 -1471842  322.8 < 2.2e-16 ***
area_sqm             1 -1471379  786.2 < 2.2e-16 ***
difficulty           2 -1468601 3566.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Describe percentage area covered by building for all projects
```

Single term deletions	Df	LRT
split_yes_no	1	3099.8 ***
inval_yes_no	1	206.8 ***
bad_imagery_yes_no	1	322.8 ***
locked_for_mapping	1	538.8 ***
area_sqm	1	786.2 ***
difficulty	2	3566.7 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Figure 8: Reproduction of the 2nd part of Table 5

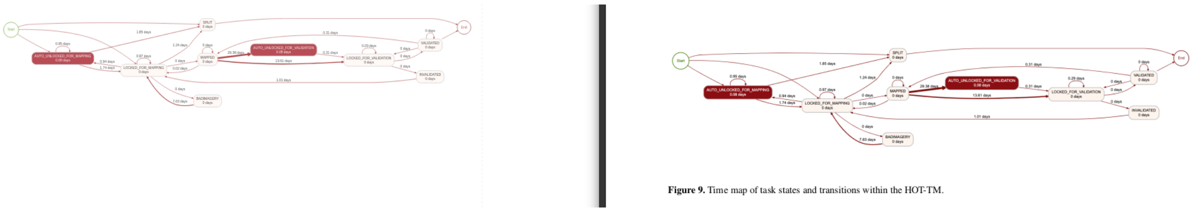


Figure 9. Time map of task states and transitions within the HOT-TM.

Figure 9: Reproduction of fig 9

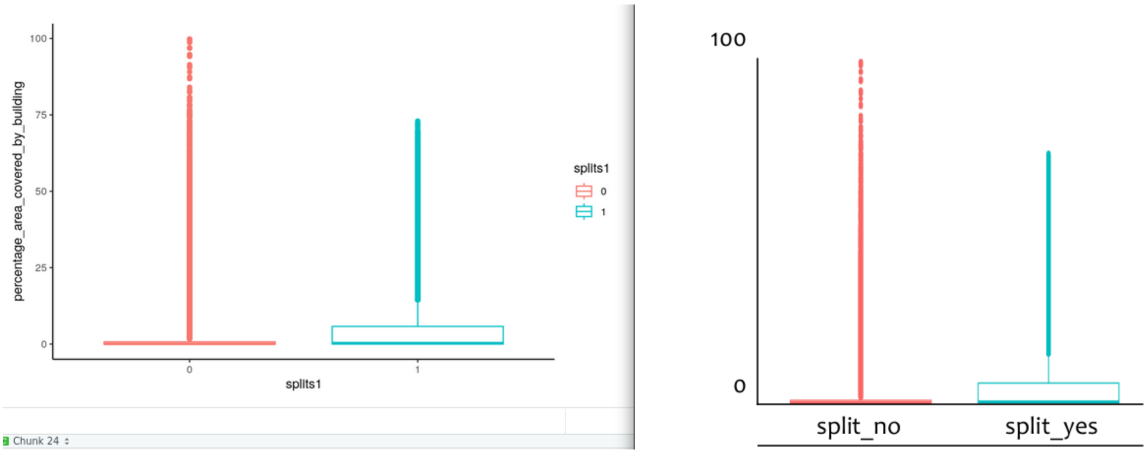


Figure 10: Reproduction of fig 10