The Legacy of the European Post-Master in Urbanism at TU Delft: A Text Mining Approach

Jane Doe1,✉, and John Q. Doe1

January 13, 2023

Text of abstract

1 Delft University of Technology

✉ Correspondence: [Jane Doe <janedoe@fosg.org>](mailto:janedoe@fosg.org)

Keywords: urbanism; education; post-master; text mining

Highlights: These are the highlights.

# Introduction

The research presented in this paper was prompted by the closure of the European post-Master of Urbanism (EMU) of the Department of Urbanism at the Faculty of Architecture and the Built Environment, TU Delft. The EMU was an advanced master ran jointly by TU Delft, KU Leuven, UPC Barcelona and Università IUAV di Venezia.

In order to describe the legacy of the EMU program, including the distinctive features of its didactic approach, this paper aims to reveal the main topics taught in it and how those topics had evolved through the years of the program. To that end, we employed a text mining approach in which we analysed its output: 96 theses with an average of 1.6506518^{5} words produced over the years for the duration of the program between 2007-2021. The first year of the program was not represented, as PDF files were only available from 2008 onward.

# Methods

## Data collection

The theses are available in PDF format with complex layouts typical to an urbanism project in which text and various types of graphics are combined.

A total number of 81 were analysed.

|  |
| --- |
| Figure 1: Location of EMU theses included in the analysis |

Figure [Figure 1](#fig-map-theses) shows the geographic spread of the theses included in the analysis.

## Data analysis

(Silge and Robinson, 2017) (Hvitfeldt and Silge, 2021)

### Tokenisation

Word-level tokenisation was applied with the unnest\_tokens() function of the tidytext package. This was preferred over other tokenisers as it produced results that work seamlessly with other tidyverse tools.

### Stop words

* Three types of stop words: global, subject-specific and document stop words.
* Global stop words:
  + All theses were written in English, so for common English stop words the “snowball” lexicon was used
  + Optionally, try creating a custom stop word list on the entire corpus, by removing high-frequency words, starting with 20 words and increasing by 10 until reaching words that are not appropriate as stop words
* Subject-specific stop words:
  + Manually constructed list, needs domain knowledge
  + Words that are commonly used to structure a theses, such as “preface”, “contents”, and “introduction”, were added as custom stop words
* Document stop words:
  + Stop words from titles and author names should only be removed from corresponding theses

stop\_words\_custom <- emu\_theses %>%   
 unnest\_tokens(output = word, input = title) %>% ### add words from the title  
 select(word) %>%   
 rbind(., data.frame(word = c("preface", "foreword", "introduction", "conclusion", "thesis",   
 "source", "author"))) %>%   
 # rbind(., data.frame(word = c("city", "urban", "urbanism"))) %>%   
 rbind(., data.frame(word = c("hab", "km")))  
  
emu\_theses\_words <- emu\_theses %>%  
 mutate(author = paste(first\_name, last\_name)) %>%   
 select(author, title, text, -c(graduation\_year, first\_name, last\_name)) %>%   
 unnest\_tokens(output = word, input = text) %>% # remove punctuation, convert to lowercase, seperate all words  
 anti\_join(stop\_words, by = "word") %>% # remove stop words  
 anti\_join(stop\_words\_custom, by = "word")

Stemming is a pre-processing step that needs to be carefully thought through, as it might either degrade topic modeling or produce no meaningful results (Hvitfeldt and Silge, 2021).

emu\_theses\_words %>%   
 mutate(stem = wordStem(words = word,   
 language = "porter")) %>% # Implement the Porter stemming algorithm for English provided by the SnowballC package  
 count(stem, sort = TRUE)

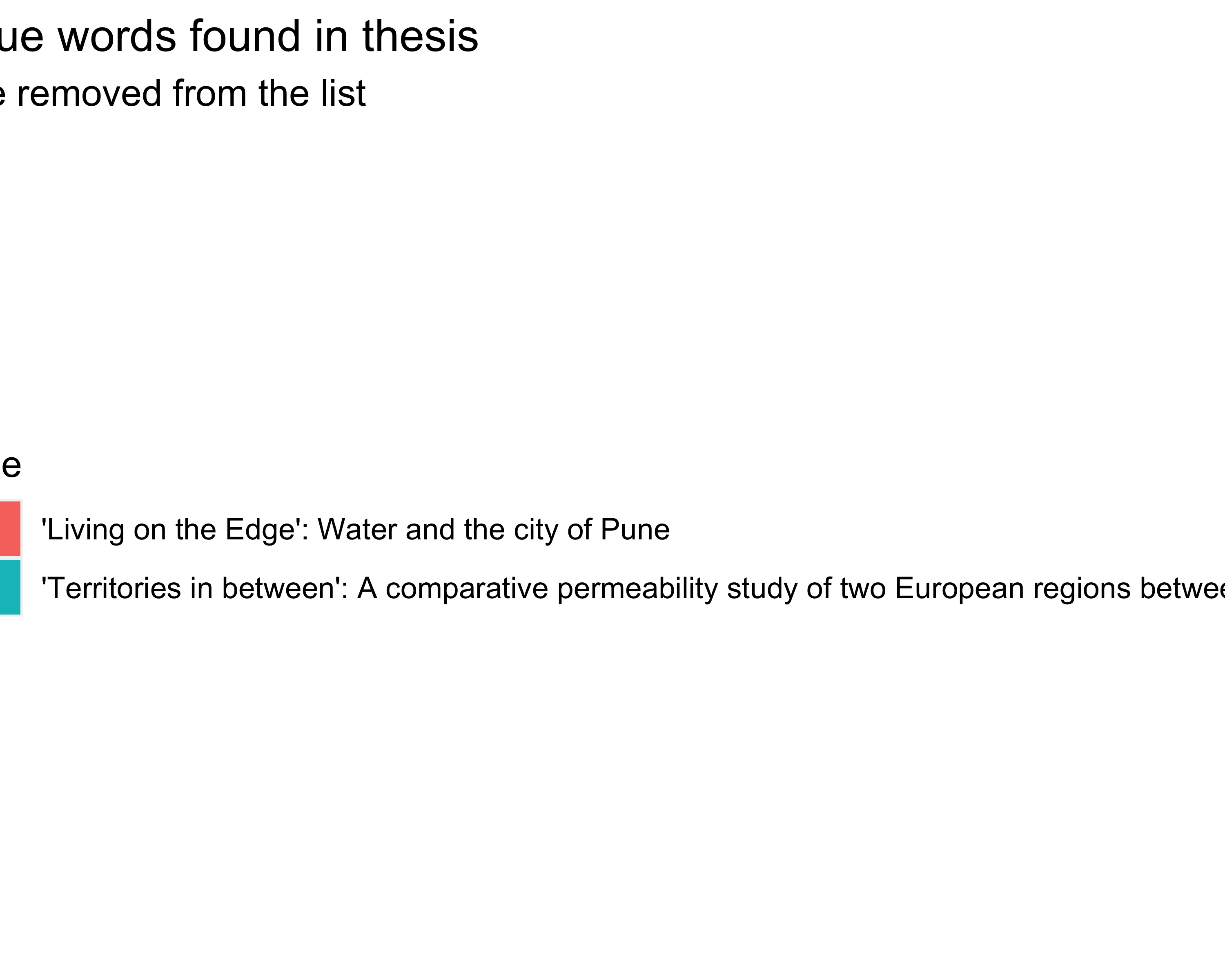
# A tibble: 47,027 × 2  
 stem n  
 <chr> <int>  
 1 fig 4220  
 2 1 3974  
 3 2 3787  
 4 plan 3391  
 5 structur 3290  
 6 natur 3238  
 7 3 3132  
 8 develop 3112  
 9 main 3026  
10 network 2978  
# … with 47,017 more rows

# Results

# Which are the top 20 most frequently used words in each thesis?  
emu\_theses\_top\_20 <- emu\_theses\_words %>%  
 group\_by(title) %>%  
 count(word, sort = TRUE) %>%  
 slice\_max(n, n = 20) %>% ### top\_n is superseded; better use slice\_max instead  
 # filter(n > 50) %>% ### filter words with a minimum count  
 ungroup() %>%   
 mutate(word = reorder(word, n))  
  
# Which are first 5 most frequently used words in each thesis?  
emu\_theses\_first\_5 <- emu\_theses\_words %>%  
 count(title, word) %>%   
 group\_by(title) %>%   
 arrange(desc(n)) %>%   
 slice(1:5)  
  
# Which theses contain "fabric" in the top 20 most frequently used words?  
emu\_theses\_top\_20 %>%   
 filter(word == "fabric")

# A tibble: 3 × 3  
 title word n  
 <chr> <fct> <int>  
1 'Living on the Edge': Water and the city of Pune fabr… 84  
2 Building Back Better: The redevelopment of Bam in Post-disaster C… fabr… 15  
3 Productive landscapes of Moscow: Binding modernities fabr… 43

emu\_theses\_top\_20 %>%  
 slice(1:40) %>%   
 ggplot(aes(x = reorder(word, n), y = n, fill = title)) +  
 geom\_col() +  
 xlab(NULL) +  
 coord\_flip() +  
 labs(y = "Count",  
 x = "Unique words",  
 title = "Count of unique words found in thesis",  
 subtitle = "Stop words were removed from the list") +  
 facet\_wrap( ~ title, scales = "free", ncol = 2)



# # wip function ngrams  
# get\_ngrams <- function(df, n) {  
# if (!is.numeric(n)) {  
# stop("n is not numeric")  
# }  
# emu\_theses %>%  
# mutate(author = paste(first\_name, last\_name)) %>%   
# select(author, title, text, -c(year, first\_name, last\_name, text\_start, text\_end)) %>%   
# unnest\_tokens(bigram, text, token = "ngrams", n = 2) %>%   
# separate(bigram, into = c("first","second"), sep = " ", remove = FALSE) %>%  
# anti\_join(stop\_words, by = c("first" = "word")) %>%  
# anti\_join(stop\_words, by = c("second" = "word")) %>%  
# anti\_join(stop\_words\_custom, by = c("first" = "word")) %>%  
# anti\_join(stop\_words\_custom, by = c("second" = "word")) %>%  
# filter(str\_detect(first, "[a-z]") & str\_detect(second, "[a-z]")) %>%   
# filter(first != second) %>%   
# filter(first != "source")  
# }  
# emu\_theses\_ngrams <- get\_ngrams(emu\_theses, 2)  
  
# bigrams  
emu\_theses\_bigrams <- emu\_theses %>%  
 mutate(author = paste(first\_name, last\_name)) %>%   
 select(author, title, text, -c(graduation\_year, first\_name, last\_name)) %>%   
 unnest\_tokens(bigram, text, token = "ngrams", n = 2) %>%   
 separate(bigram, into = c("first","second"), sep = " ", remove = FALSE) %>%  
 anti\_join(stop\_words, by = c("first" = "word")) %>%  
 anti\_join(stop\_words, by = c("second" = "word")) %>%  
 anti\_join(stop\_words\_custom, by = c("first" = "word")) %>%  
 anti\_join(stop\_words\_custom, by = c("second" = "word")) %>%  
 filter(str\_detect(first, "[a-z]") & str\_detect(second, "[a-z]")) %>%   
 filter(first != second) %>%   
 filter(first != "source")  
  
emu\_theses\_bigrams\_top <- emu\_theses\_bigrams %>%  
 group\_by(title) %>%  
 count(bigram, sort = TRUE) %>%  
 slice\_max(n, n = 10) %>% ### top\_n is superseded; better use slice\_max instead  
 # filter(n > 50) ### filter words with a minimum count  
 ungroup() %>%   
 mutate(bigram = reorder(bigram, n))  
  
emu\_theses\_bigrams\_top %>%  
 ggplot(aes(x = bigram, y = n, fill = title)) +  
 geom\_col() +  
 xlab(NULL) +  
 coord\_flip() +  
 labs(y = "Count",  
 x = "Unique words",  
 title = "Count of unique words found in thesis",  
 subtitle = "Stop words were removed from the list") +  
 facet\_wrap( ~ title, scales = "free", ncol = 2)



# find the most frequently used bigrams  
emu\_thesis\_bigrams\_top <- emu\_theses\_bigrams %>%  
 group\_by(title) %>%   
 count(bigram, sort = TRUE) %>%  
 slice\_max(n, n = 10) %>% ### change the number of top words to show  
 mutate(bigram = reorder(bigram, n)) %>%   
 print()

# A tibble: 963 × 3  
# Groups: title [80]  
 title bigram n  
 <chr> <fct> <int>  
 1 'Living on the Edge': Water and the city of Pune aditya deshmukh 71  
 2 'Living on the Edge': Water and the city of Pune informal settlements 15  
 3 'Living on the Edge': Water and the city of Pune ambil stream 10  
 4 'Living on the Edge': Water and the city of Pune wa ter 10  
 5 'Living on the Edge': Water and the city of Pune majoor adda 6  
 6 'Living on the Edge': Water and the city of Pune close proximity 5  
 7 'Living on the Edge': Water and the city of Pune natural resources 5  
 8 'Living on the Edge': Water and the city of Pune pimpri chinchwad 5  
 9 'Living on the Edge': Water and the city of Pune riv er 5  
10 'Living on the Edge': Water and the city of Pune 12th july 4  
# … with 953 more rows

# Discussion

# Conclusion

# Acknowledgements

# References

Hvitfeldt, E., Silge, J., 2021. Supervised Machine Learning for Text Analysis in R, 1st edition. ed. Chapman and Hall/CRC, Boca Raton.

Silge, J., Robinson, D., 2017. Text Mining with R: A Tidy Approach, 1st edition. ed. O’Reilly Media, Beijing ; Boston.

### Colophon

This report was generated on 2023-01-13 21:55:15 using the following computational environment and dependencies:

# which R packages and versions?  
if ("devtools" %in% installed.packages()) devtools::session\_info()

─ Session info ───────────────────────────────────────────────────────────────  
 setting value  
 version R version 4.2.2 (2022-10-31)  
 os macOS Monterey 12.1  
 system aarch64, darwin20  
 ui X11  
 language (EN)  
 collate en\_US.UTF-8  
 ctype en\_US.UTF-8  
 tz Europe/Amsterdam  
 date 2023-01-13  
 pandoc 2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (via rmarkdown)  
  
─ Packages ───────────────────────────────────────────────────────────────────  
 ! package \* version date (UTC) lib source  
 P assertthat 0.2.1 2019-03-21 [?] CRAN (R 4.2.0)  
 P backports 1.4.1 2021-12-13 [?] CRAN (R 4.2.0)  
 P bit 4.0.5 2022-11-15 [?] CRAN (R 4.2.0)  
 P bit64 4.0.5 2020-08-30 [?] CRAN (R 4.2.0)  
 P broom 1.0.2 2022-12-15 [?] CRAN (R 4.2.0)  
 P cachem 1.0.6 2021-08-19 [?] CRAN (R 4.2.0)  
 P callr 3.7.3 2022-11-02 [?] CRAN (R 4.2.0)  
 P cellranger 1.1.0 2016-07-27 [?] CRAN (R 4.2.0)  
 P cli 3.5.0 2022-12-20 [?] CRAN (R 4.2.0)  
 P colorspace 2.0-3 2022-02-21 [?] CRAN (R 4.2.0)  
 P crayon 1.5.2 2022-09-29 [?] CRAN (R 4.2.0)  
 P curl 4.3.3 2022-10-06 [?] CRAN (R 4.2.0)  
 P DBI 1.1.3 2022-06-18 [?] CRAN (R 4.2.0)  
 P dbplyr 2.2.1 2022-06-27 [?] CRAN (R 4.2.0)  
 P desc 1.4.2 2022-09-08 [?] CRAN (R 4.2.0)  
 P devtools 2.4.5 2022-10-11 [?] CRAN (R 4.2.0)  
 P digest 0.6.31 2022-12-11 [?] CRAN (R 4.2.0)  
 P dplyr \* 1.0.10 2022-09-01 [?] CRAN (R 4.2.0)  
 P ellipsis 0.3.2 2021-04-29 [?] CRAN (R 4.2.0)  
 P evaluate 0.19 2022-12-13 [?] CRAN (R 4.2.0)  
 P fansi 1.0.3 2022-03-24 [?] CRAN (R 4.2.0)  
 P farver 2.1.1 2022-07-06 [?] CRAN (R 4.2.0)  
 P fastmap 1.1.0 2021-01-25 [?] CRAN (R 4.2.0)  
 P forcats \* 0.5.2 2022-08-19 [?] CRAN (R 4.2.0)  
 P fs 1.5.2 2021-12-08 [?] CRAN (R 4.2.0)  
 P gargle 1.2.1 2022-09-08 [?] CRAN (R 4.2.0)  
 P generics 0.1.3 2022-07-05 [?] CRAN (R 4.2.0)  
 P ggplot2 \* 3.4.0 2022-11-04 [?] CRAN (R 4.2.0)  
 P glue 1.6.2 2022-02-24 [?] CRAN (R 4.2.0)  
 P googledrive 2.0.0 2021-07-08 [?] CRAN (R 4.2.0)  
 P googlesheets4 1.0.1 2022-08-13 [?] CRAN (R 4.2.0)  
 P gtable 0.3.1 2022-09-01 [?] CRAN (R 4.2.0)  
 P haven 2.5.1 2022-08-22 [?] CRAN (R 4.2.0)  
 P here \* 1.0.1 2020-12-13 [?] CRAN (R 4.2.0)  
 P hms 1.1.2 2022-08-19 [?] CRAN (R 4.2.0)  
 P htmltools 0.5.4 2022-12-07 [?] CRAN (R 4.2.0)  
 P htmlwidgets 1.6.1 2023-01-07 [?] CRAN (R 4.2.2)  
 P httpuv 1.6.7 2022-12-14 [?] CRAN (R 4.2.0)  
 P httr 1.4.4 2022-08-17 [?] CRAN (R 4.2.0)  
 P janeaustenr 1.0.0 2022-08-26 [?] CRAN (R 4.2.0)  
 P jsonlite 1.8.4 2022-12-06 [?] CRAN (R 4.2.0)  
 P knitr 1.41 2022-11-18 [?] CRAN (R 4.2.0)  
 P labeling 0.4.2 2020-10-20 [?] CRAN (R 4.2.0)  
 P later 1.3.0 2021-08-18 [?] CRAN (R 4.2.0)  
 P lattice 0.20-45 2021-09-22 [3] CRAN (R 4.2.2)  
 P lifecycle 1.0.3 2022-10-07 [?] CRAN (R 4.2.0)  
 P lubridate 1.9.0 2022-11-06 [?] CRAN (R 4.2.0)  
 P magrittr 2.0.3 2022-03-30 [?] CRAN (R 4.2.0)  
 maps 3.4.1 2022-10-30 [1] CRAN (R 4.2.0)  
 P Matrix 1.5-1 2022-09-13 [3] CRAN (R 4.2.2)  
 P memoise 2.0.1 2021-11-26 [?] CRAN (R 4.2.0)  
 P mime 0.12 2021-09-28 [?] CRAN (R 4.2.0)  
 P miniUI 0.1.1.1 2018-05-18 [?] CRAN (R 4.2.0)  
 R minTEMU \* 0.0.0.9000 <NA> [?] <NA>  
 P modelr 0.1.10 2022-11-11 [?] CRAN (R 4.2.0)  
 P munsell 0.5.0 2018-06-12 [?] CRAN (R 4.2.0)  
 P NLP \* 0.2-1 2020-10-14 [?] CRAN (R 4.2.0)  
 P pillar 1.8.1 2022-08-19 [?] CRAN (R 4.2.0)  
 P pkgbuild 1.4.0 2022-11-27 [?] CRAN (R 4.2.0)  
 P pkgconfig 2.0.3 2019-09-22 [?] CRAN (R 4.2.0)  
 P pkgload 1.3.2 2022-11-16 [?] CRAN (R 4.2.0)  
 png 0.1-8 2022-11-29 [1] CRAN (R 4.2.0)  
 P prettyunits 1.1.1 2020-01-24 [?] CRAN (R 4.2.0)  
 P processx 3.8.0 2022-10-26 [?] CRAN (R 4.2.0)  
 P profvis 0.3.7 2020-11-02 [?] CRAN (R 4.2.0)  
 P promises 1.2.0.1 2021-02-11 [?] CRAN (R 4.2.0)  
 P ps 1.7.2 2022-10-26 [?] CRAN (R 4.2.0)  
 P purrr \* 1.0.0 2022-12-20 [?] CRAN (R 4.2.0)  
 P R6 2.5.1 2021-08-19 [?] CRAN (R 4.2.0)  
 P Rcpp 1.0.9 2022-07-08 [?] CRAN (R 4.2.0)  
 P readr \* 2.1.3 2022-10-01 [?] CRAN (R 4.2.0)  
 P readxl 1.4.1 2022-08-17 [?] CRAN (R 4.2.0)  
 P remotes 2.4.2 2021-11-30 [?] CRAN (R 4.2.0)  
 P reprex 2.0.2 2022-08-17 [?] CRAN (R 4.2.0)  
 reticulate 1.27 2023-01-07 [1] CRAN (R 4.2.2)  
 P rlang 1.0.6 2022-09-24 [?] CRAN (R 4.2.0)  
 P rmarkdown 2.19 2022-12-15 [?] CRAN (R 4.2.0)  
 P rprojroot 2.0.3 2022-04-02 [?] CRAN (R 4.2.0)  
 P rstudioapi 0.14 2022-08-22 [?] CRAN (R 4.2.0)  
 P rvest 1.0.3 2022-08-19 [?] CRAN (R 4.2.0)  
 P scales 1.2.1 2022-08-20 [?] CRAN (R 4.2.0)  
 P sessioninfo 1.2.2 2021-12-06 [?] CRAN (R 4.2.0)  
 P shiny 1.7.4 2022-12-15 [?] CRAN (R 4.2.0)  
 P slam 0.1-50 2022-01-08 [?] CRAN (R 4.2.0)  
 P SnowballC \* 0.7.0 2020-04-01 [?] CRAN (R 4.2.0)  
 P stringi 1.7.8 2022-07-11 [?] CRAN (R 4.2.0)  
 P stringr \* 1.5.0 2022-12-02 [?] CRAN (R 4.2.0)  
 P tibble \* 3.1.8 2022-07-22 [?] CRAN (R 4.2.0)  
 P tidygeocoder \* 1.0.5 2021-11-02 [?] CRAN (R 4.2.0)  
 P tidyr \* 1.2.1 2022-09-08 [?] CRAN (R 4.2.0)  
 P tidyselect 1.2.0 2022-10-10 [?] CRAN (R 4.2.0)  
 P tidytext \* 0.4.1 2023-01-07 [?] CRAN (R 4.2.0)  
 P tidyverse \* 1.3.2 2022-07-18 [?] CRAN (R 4.2.0)  
 P timechange 0.1.1 2022-11-04 [?] CRAN (R 4.2.0)  
 P tm \* 0.7-10 2022-12-14 [?] CRAN (R 4.2.0)  
 P tokenizers 0.3.0 2022-12-22 [?] CRAN (R 4.2.0)  
 P tzdb 0.3.0 2022-03-28 [?] CRAN (R 4.2.0)  
 P urlchecker 1.0.1 2021-11-30 [?] CRAN (R 4.2.0)  
 P usethis 2.1.6 2022-05-25 [?] CRAN (R 4.2.0)  
 P utf8 1.2.2 2021-07-24 [?] CRAN (R 4.2.0)  
 P vctrs 0.5.1 2022-11-16 [?] CRAN (R 4.2.0)  
 P vroom 1.6.0 2022-09-30 [?] CRAN (R 4.2.0)  
 P withr 2.5.0 2022-03-03 [?] CRAN (R 4.2.0)  
 P xfun 0.36 2022-12-21 [?] CRAN (R 4.2.0)  
 P xml2 1.3.3 2021-11-30 [?] CRAN (R 4.2.0)  
 P xtable 1.8-4 2019-04-21 [?] CRAN (R 4.2.0)  
 P yaml 2.3.6 2022-10-18 [?] CRAN (R 4.2.0)  
  
 [1] /Users/claudiuforgaci/Library/Caches/org.R-project.R/R/renv/library/minTEMU-a20e3900/R-4.2/aarch64-apple-darwin20  
 [2] /Users/claudiuforgaci/Projects/minTEMU/renv/sandbox/R-4.2/aarch64-apple-darwin20/84ba8b13  
 [3] /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/library  
  
 P ── Loaded and on-disk path mismatch.  
 R ── Package was removed from disk.  
  
─ Python configuration ───────────────────────────────────────────────────────  
 python: /Users/claudiuforgaci/Projects/minTEMU/R/pyvenv/bin/python  
 libpython: /Users/claudiuforgaci/.pyenv/versions/3.9.7/lib/libpython3.9.dylib  
 pythonhome: /Users/claudiuforgaci/Projects/minTEMU/R/pyvenv:/Users/claudiuforgaci/Projects/minTEMU/R/pyvenv  
 version: 3.9.7 (default, Jan 9 2023, 13:50:30) [Clang 13.1.6 (clang-1316.0.21.2.5)]  
 numpy: /Users/claudiuforgaci/Projects/minTEMU/R/pyvenv/lib/python3.9/site-packages/numpy  
 numpy\_version: 1.24.1  
   
 NOTE: Python version was forced by use\_python function  
  
──────────────────────────────────────────────────────────────────────────────

The current Git commit details are:

# what commit is this file at?   
if ("git2r" %in% installed.packages() & git2r::in\_repository(path = ".")) git2r::repository(here::here())

Local: analyse-cf /Users/claudiuforgaci/Projects/minTEMU  
Remote: analyse-cf @ origin (https://github.com/UD3-Lab/minTEMU.git)  
Head: [4aec000] 2023-01-12: Update setup code in paper