Indoor dust bacterial and fungal microbiota composition and allergic diseases: a scoping review

Data cleaning and preparation

Javier Mancilla Galindo, MSc student
Supervisors: Inge Wouters and Alex Bossers

2024-05-17

Packages and session information

```r
if (!require("pacman", quietly = TRUE)) {
  install.packages("pacman")
}

pacman::p_load(
  tidyverse, # Basic data handling.
  readxl, # Import data in .xlsx format.
  table1, # Used for column labeling.
  maps, # Used to retrieve ISO3 codes for countries.
  haven, # Export data into different formats.
  report, # Used to generate package citations in markdown format.
  gt # Print html tables.
)
```

```
R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

Matrix products: default


locale:
[1] LC_COLLATE=Spanish_Mexico.utf8  LC_CTYPE=Spanish_Mexico.utf8
[3] LC_MONETARY=Spanish_Mexico.utf8 LC_NUMERIC=C
[5] LC_TIME=Spanish_Mexico.utf8

time zone: Europe/Berlin
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gt_0.10.1       report_0.5.8    haven_2.5.4     maps_3.4.2
 [5] table1_1.4.3    readxl_1.4.3    lubridate_1.9.3 forcats_1.0.0
 [9] stringr_1.5.1   dplyr_1.1.4     purrr_1.0.2     readr_2.1.5
[13] tidyr_1.3.1     tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
[17] pacman_0.5.1
```

Main dataframe

I will load dataset and remove redundant columns or those for self use. The original data charting dataset is in the wide format. I will separate some variables into new tables later and convert to long formate to facilitate analyses later on.

The original dataset has 144 rows and 43 columns.

```r
columns_to_remove <- c(
  "Dupl", # Duplicate records in search, only for own records
  "Type", # No varying data since all were journal articles
  "Download available", # Only for own records, all were available
  "Abstract","Citation","Link", # Info also in references-dust-microbiome.csv
  "Pathway_internal", # Internal pathway to access PDFs in my personal laptop
  "Indoor_dust_microbiome", # No varying data, as all are "yes"
  "ISO3", # Not needed for to do the join.
  "Study_unit", # Redundant with 'Building' variable.
  "Comments", # Annotations used for my own use.
  "Confounding/causality_comments", # Annotations used for my own use.
  "Study_size" # Only registered this for few studies of my interest.
)

data <- data %>% select(!all_of(columns_to_remove))
```

The dataset now has 144 rows and 30 columns.

I will now import the attributes for the dataset from the sourced script *variable_names.R*

I will now import the bibliography dataset.

References dataframe

I will now add the citation key column to `data`.

`data` now has 31 columns.

Countries dataframe

I will now process country data and link it to their corresponding regions and income classification by using the data from the world bank (The World Bank, 2024).

Note that there are studies for which sampled occured in more than 1 country, reason why the count of countries can exceed the initial total count. Additionally, there were studies for which sampled ocurred in the international space station (ISS).

More than 1 country	n	Percentage
No	125	86.8
Yes	14	9.7
ISS	5	3.5

After excluding studies in the ISS, `countries` now has 175 rows.

`countries` now has 5 columns.

Dust collectors dataframe

`collectors` has 144 rows.

Environmental determinants dataframe

`environmental_determinants` has 144 rows.

Writing and saving into different data formats for greater reusability

S4 object

I will store individual dataframes in an S4 object:

```
Formal class 'DataFrameCollection' [package ".GlobalEnv"] with 5 slots
  ..@ data                      : tibble [144 x 31] (S3: tbl_df/tbl/data.frame)
  ..@ countries                 : tibble [175 x 5] (S3: tbl_df/tbl/data.frame)
  ..@ collectors                : tibble [184 x 2] (S3: tbl_df/tbl/data.frame)
  ..@ environmental_determinants: tibble [600 x 2] (S3: tbl_df/tbl/data.frame)
  ..@ references                :'data.frame':  144 obs. of  90 variables:
```

R Data

```r
save(Data_Dust_Microbiome_Review,
     file = paste0(psfolder,"/Data_Dust_Microbiome_Review.RData"))
```

SPSS

```r
sav_folder <- "../data/processed/sav"
dir.create(sav_folder, showWarnings = FALSE, recursive = TRUE)

# Save each data frame to a .sav file
write_sav(data, file.path(sav_folder, "main_data.sav"))
write_sav(countries, file.path(sav_folder, "countries.sav"))
write_sav(collectors, file.path(sav_folder, "collectors.sav"))
write_sav(environmental_determinants,
          file.path(sav_folder, "environmental_determinants.sav"))
```

SAS

```r
xpt_folder <- "../data/processed/xpt"
dir.create(xpt_folder, showWarnings = FALSE, recursive = TRUE)

# Save each data frame to a .xpt file
write_xpt(data, file.path(xpt_folder, "main_data.xpt"))
write_xpt(countries, file.path(xpt_folder, "countries.xpt"))
write_xpt(collectors, file.path(xpt_folder, "collectors.xpt"))
write_xpt(environmental_determinants,
          file.path(xpt_folder, "environmental_determinants.xpt"))
```

STATA

```
dta_folder <- "../data/processed/dta"
dir.create(dta_folder, showWarnings = FALSE, recursive = TRUE)

# Save each data frame to a .dta file
write_dta(data, file.path(dta_folder, "main_data.dta"))
write_dta(countries, file.path(dta_folder, "countries.dta"))
write_dta(collectors, file.path(dta_folder, "collectors.dta"))
write_dta(environmental_determinants,
          file.path(dta_folder, "environmental_determinants.dta"))
```

CSV

```
csv_folder <- "../data/processed/csv"
dir.create(csv_folder, showWarnings = FALSE, recursive = TRUE)

# Save each data frame to a CSV file
write.csv(data, file.path(csv_folder, "main_data.csv"), row.names = FALSE)
write.csv(countries, file.path(csv_folder, "countries.csv"), row.names = FALSE)
write.csv(collectors, file.path(csv_folder, "collectors.csv"), row.names = FALSE)
write.csv(environmental_determinants, file.path(csv_folder, "environmental_determinants.csv")
write.csv(references, file.path(csv_folder, "references.csv"), row.names = FALSE)
```

References

Package references

- Becker, Minka, Deckmyn. A (2023). *maps: Draw Geographical Maps.* R package version 3.4.2, https://CRAN.R-project.org/package=maps.
- Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." *Journal of Statistical Software*, *40*(3), 1-25. https://www.jstatsoft.org/v40/i03/.

- Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2024). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.10.1, https://CRAN.R-project.org/package=gt.
- Makowski D, Lüdecke D, Patil I, Thériault R, Ben-Shachar M, Wiernik B (2023). "Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption." *CRAN*. https://easystats.github.io/report/.
- Müller K, Wickham H (2023). *tibble: Simple Data Frames*. R package version 3.2.1, https://CRAN.R-project.org/package=tibble.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Rich B (2023). *table1: Tables of Descriptive Statistics in HTML*. R package version 1.4.3, https://CRAN.R-project.org/package=table1.
- Rinker TW, Kurkiewicz D (2018). *pacman: Package Management for R*. version 0.5.0, http://github.com/trinker/pacman.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- Wickham H (2023). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 1.0.0, https://CRAN.R-project.org/package=forcats.
- Wickham H (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1, https://CRAN.R-project.org/package=stringr.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.
- Wickham H, Bryan J (2023). *readxl: Read Excel Files*. R package version 1.4.3, https://CRAN.R-project.org/package=readxl.
- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, https://CRAN.R-project.org/package=dplyr.
- Wickham H, Henry L (2023). *purrr: Functional Programming Tools*. R package version 1.0.2, https://CRAN.R-project.org/package=purrr.
- Wickham H, Hester J, Bryan J (2024). *readr: Read Rectangular Text Data*. R package version 2.1.5, https://CRAN.R-project.org/package=readr.
- Wickham H, Miller E, Smith D (2023). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.5.4, https://CRAN.R-project.org/package=haven.
- Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, https://CRAN.R-project.org/package=tidyr.

Other references

The World Bank, 2024. World bank country and lending groups. Data help desk.