Data cleaning and preparation

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

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

2024-05-26

Packages and session information

```
if (!require("pacman", quietly = TRUE)) {
  install.packages("pacman")
}
pacman::p_load(
  tidyverse,  # Basic data wrangling.
  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.
  officer,    # Export tables
  gto,        # Add gt table to a word document.
  gt          # Print and save 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       gto_0.1.1       officer_0.6.5   report_0.5.8
 [5] haven_2.5.4     maps_3.4.2      table1_1.4.3    readxl_1.4.3
 [9] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
[13] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
[17] ggplot2_3.5.1   tidyverse_2.0.0 pacman_0.5.1
```

Main dataframe

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

The original dataset has 144 rows and 44 columns.

```
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"
  "Environmental_category", # Will be recreated with later code.
  "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 sampling occurred in more than 1 country, reason why the count of countries can exceed the initial total count. Additionally, there were studies for which samples were obtained 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.

The final `countries` has 175 rows and 5 columns.

Dust collectors dataframe

`collectors` has 184 rows and and 2 columns.

Buildings dataframe

`buildings` has 159 rows and and 2 columns.

Environmental determinants dataframe

I will filter only studies that reported environmental characteristics and exclude study number 69 since this study reports 668 environmental determinants (Pakpour et al., 2016), which would be very challenging to summarize in a way that is comparable to other studies included in this review.

`environmental_determinants` has 595 rows and 2 columns.

Some processing of environmental determinants is needed to analyze:

`environmental_determinants` now has 595 rows and 3 columns.

This is the record of which environmental categories were mapped to each environmental determinant extracted from the studies in the review:

Category	Environmental determinants assessed
air pollutants	black carbon, indoor CO2, indoor NO2, indoor PM2.5, indoor PM10, outdoor coarse particles, outdoor fine particles, outdoor NO, outdoor NO2, outdoor PM10, outdoor PM2.5, outdoor SO2, traffic air pollution
allergen	alternaria allergen, aspergillus allergen, cat allergen, cockroach allergen, dog allergen, mite allergen, mouse allergen
building characteristics	age of building, building architecture, building condition, building function, building material, building organization of space, building orientation, building structure, building type, curtains, curtains size, textile curtain factor, distance from bed, floor level, floor material, floor type, gas cooker, housing type, human use patterns, location in building, number of rooms, privacy index, open kitchen connected to the living room, ratio of window to floor area, recent renovation, roof type, room type, size of indoor environment, wall surface type, wing, woodstove
chemicals	ambient chemical compounds, DEHP, endotoxin, ergosterol, formaldehyde, microbial toxins, microplastics, muramic acid, pesticides, polybrominated diphenyl ethers (PBDEs)
cleaning habits	cleaning, cleaning habits, cleaning frequency, cleaning method, cleaning status, net weight of vacuumed dust as indicator of cleaning habits
farming	farm, farmer, farming, living on farm, type of farming
furniture	electronics, furniture surfaces, furniture
geography	altitude, climate, density of buildings, density of roads, distance between buildings, distance from the Equator, distance to city center, distance to coast, elevation, geographical location, geographical distance, hog density, land use, living near expressway, meteorological conditions, other geographical data, population density, precipitation, wind speed
green environment	biodiversity of forests nearby, flowering plants in vicinity, green spaces, green-renovated building, indoor plants, main vascular plant species outdoors, number of indoor plants, plant diversity, plants, plants in building, plants in room, percentage of woody vegetation cover, proximity to green areas, residential green space, species of indoor plants, vascular plant diversity
heating	heating, heating systems, type of heating
humidity/dampness	dampness, degree of flood-related damage, flooded building, humidity, humidity variance, indoor relative humidity, moisture, moisture damage, relative humidity, water leaks, water damage
infestation	bug infestation, cockroaches, infestations, insecticide use, rodents, mites, pests

light	light, light in microenvironment
mold	visible mold, mold
building	adult inhabitants, children, household members, number of inhabitants,
occupants	number of occupants, occupants, occupant density, person visits per day, time that people spend in room
outdoor	arid wasteland soil, farm dust microbiome, lakeshore soil, outdoor
microbiome	microbiome, outdoor haze microbiome, soil microbiome, woods soil
pets	birds, cat, dog, guinea pig, hamster, pets, rabbit
season	season, month of sampling
smoking	smoking, tobacco exposure
temperature	temperature, temperature outdoor, temperature variance
urbanicity	urbanicity
ventilation	aeration time, airflow rate, air conditioning, air exchange rate, natural ventilation, number of windows, outdoor air delivery rate, proportion of apertures, type of ventilation, ventilation
water sources	distance to water, water sources
other	composting, carbon in dust, dust pH, dust salinity, dust redox potential, dust conductivity, grass seeds, height of sampling, human oral microbiome, nitrogen in dust, occupational exposure, soil pH, use of antimicrobials

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 6 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)
  ..@ buildings                 : tibble [159 x 2] (S3: tbl_df/tbl/data.frame)
  ..@ environmental_determinants: tibble [595 x 3] (S3: tbl_df/tbl/data.frame)
  ..@ references                :'data.frame':  144 obs. of  90 variables:
```

R Data

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

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(buildings, file.path(csv_folder, "buildings.csv"),
          row.names = FALSE)
write.csv(environmental_determinants,
          file.path(csv_folder, "environmental_determinants.csv"),
          row.names = FALSE)
write.csv(references, file.path(csv_folder, "references.csv"),
          row.names = FALSE)
```

SPSS

```
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(buildings, file.path(sav_folder, "buildings.sav"))
write_sav(environmental_determinants,
          file.path(sav_folder, "environmental_determinants.sav"))
```

SAS

```
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(buildings, file.path(xpt_folder, "buildings.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(buildings, file.path(dta_folder, "buildings.dta"))
write_dta(environmental_determinants,
          file.path(dta_folder, "environmental_determinants.dta"))
```

References

Package references

- Becker RA, Minka TP, Deckmyn. A (2023). *maps: Draw Geographical Maps*. R package version 3.4.2, https://CRAN.R-project.org/package=maps.
- Gohel D, Moog S (2024). *officer: Manipulation of Microsoft Word and PowerPoint Documents*. R package version 0.6.5, https://CRAN.R-project.org/package=officer.
- 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/.
- Hughes E (2023). *gto: Insert 'gt' Tables into Word Documents*. R package version 0.1.1, https://CRAN.R-project.org/package=gto.
- 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

Pakpour, S., Scott, J.A., Turvey, S.E., Brook, J.R., Takaro, T.K., Sears, M.R., Klironomos, J., 2016. Presence of Archaea in the Indoor Environment and Their Relationships with Housing Characteristics. Microbial Ecology 72, 305–312. https://doi.org/10.1007/s00248-016-0767-z

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