Health metrics and the spread of infectious diseases

with machine learning applications and spatial model analysis

Federica Gazzelloni

12/16/22

Table of contents

Preface

Health metrics and the spread of infectious diseases, with machine learning applications and spatial model analysis are the topics of this book.

Here you will find everything you need to analyze the state of health of a country and compare it with that of other countries. You will also be able to evaluate the best model for predicting future trends.

The author of this book is **Federica Gazzelloni** who is an actuary and a statistician graduated from the Sapienza of Rome, in Italy. She is also a collaborator of the **Institute for Health Metrics and Evaluation (IHME)** , which inspired this work to serve as a guideline for making health metrics and spatial model analysis for evaluating the state of health of a population and spread of infectious diseases.

All data used in this book are from the **Institute for Health Metrics and Evaluation (IHME)**. **GBD Results**. Seattle, WA: IHME, University of Washington, 2020. Available from https://vizhub.healthdata.org/gbd-results/. (Accessed [January 2023]) and the **World Health Organization (WHO)**. **Global Health Observatory data repository**. Available from https://apps.who.int/gho/data/.

1 Introduction

Health metrics and the spread of infectious diseases, with machine learning applications and spatial model analysis, is a manual and a textbook for an introductory health data analysis course. It can also turn out to be a useful source code for both practitioners and data scientists.

Public health metrics such as **DALYs**, **YLL**, and **YLD** are expressed in numbers of years of life lost or lived with disabilities whose sum expresses a key value generally used for ranking the health status of a population.

A focus on the impact of recent infectious disease outbreaks, such as Covid19, on the state of health of the population, will be provided along with the most affected locations. The book is structured with an alternation of text and chunks of code in the R language to let the reader be a practitioner of real-world case studies on the topic.

To be more specific, the metrics used to summarize the state of health of a population will be compared across other locations and a prediction level tested on a few key models will be provided. The idea is to use {tidymodels}, and {INLA} as modeling tools. Finally, the material contains some interesting spatial visualization, made using {ggplot2}, {leaflet}, {sf}, {rgdal} R packages, plus other main packages for allowing the user for a wider understanding of the potentiality of the R language for both spatial and health metrics.

The book is foreseen for practitioners at early stages and graduated students in STEM.

Part I

Health metrics

Health metrics are key variables to understand more about the state of health of a population. In this book we'll talk about how to calculate the **numbers of years of life lost (YLLs)** and the **numbers of years lived with disabilities (YLDs)**, to finally obtain the key metric of the **DALYs** which identify the numbers of years of life lost due to death or a disability status, namely **disability adjusted life years**.

The numbers of years of life lost by a population in comparison to other countries or to the Global mean trend, is based on the latest study results relative to how the well being of a country is in terms of the definition of a healthy life. To give an example, let's think about a population whose individuals are living a good life, so defined *healthy life* measured on **life expectancy** established to be 80 years on average, as most of the World population meets this as a deadline.

The part of the population who do not meet this age, but dies earlier, contributes as a building block of the numbers of years of life lost (YLLs), as well as for all that is related with a healthy living, the numbers of years spent dealing with a disability contribute to increasing the numbers of years lived with disabilities for a country's population.

To establish a *healthy life status for a country*, meant as the healthy life defying the state of health of a population, the sum of the two values YLLs and YLDs releases the key metric of DALYs. This metric value is used to quickly identify the level of health of a population compared to a Global review based on the latest findings of the most updated studies.

In addition, this level is used to improve the proportion of countries who are in need of a better health status recognition. To be more specific, the focus on the numbers of years can help identify the areas where most of the years are lost and need for improvement, whether in facilities, research or investments.

An improvement of health at a Global level is reached when the definition of a **healthy file** is met in most of the countries where it wasn't before.

Furthermore, what we want to analyze are a series of data that are produced taking into account the tables of mortality and future life expectancy these are defined as means that these metrics have been considered important for assessing the state of health of one point population

What we refer to are the health metrics and which refer to the number of years lost due to an increase in mortality or in any case to a mortality trend that is above a certain general level which we can consider as the level optimal health globally.

2 YLLs, YLDs and DALYs

- A closer look at the metrics, their usage and potentiality for improvements

In this section are shown the methods used for building three key metrics: YLLs, YLDs and DALYs. These will be used throughout the book for making comparisons among the state of health of different countries.

3 Metrics components

- Life tables and Life expectancy used in the book
- How to use them and where to find them

This section is dedicated to a closer look at what are the components of the health metrics, how to build them and finally how to use them for making countries comparison. YLLs,YLDs and DALYs can be used for different cause of deaths and disabilities, at different age levels.

3.1 Components

Two fundamental components are used for calculating the DALYs:

- life tables
- life expectancy

Both of these elements are key for achieving the highest value of prediction of the state of health of a population.

3.1.1 Life tables

The life tables are selected among the most frequently used, more information about how to build a life table can be found in Appendix A of this book.

```
library(tidyverse)
xmart <- read_csv("data-raw/xmart.csv",skip = 1)

xmart %>% head
```

```
# A tibble: 6 x 17
  Indicator      Age G~1 Both ~2 Male.~3 Femal~4 Both ~5 Male.~6 Femal~7 Both ~8
  <chr>          <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 nMx - age-spe~ &lt;1 ~ 4.78e-2 5.13e-2 4.41e-2 5.66e-2 6.02e-2 5.27e-2 6.75e-2
2 nMx - age-spe~ 1-4 ye~ 3.61e-3 3.67e-3 3.55e-3 4.63e-3 4.67e-3 4.58e-3 6.43e-3
3 nMx - age-spe~ 5-9 ye~ 4.61e-4 4.69e-4 4.53e-4 6.68e-4 6.75e-4 6.62e-4 1.15e-3
```

```
4 nMx - age-spe~ 10-14 ~ 3.70e-4 3.87e-4 3.52e-4 4.98e-4 5.16e-4 4.79e-4 8.27e-4
5 nMx - age-spe~ 15-19 ~ 1.32e-3 1.46e-3 1.18e-3 2.22e-3 2.74e-3 1.67e-3 1.88e-3
6 nMx - age-spe~ 20-24 ~ 2.01e-3 2.22e-3 1.79e-3 3.41e-3 4.47e-3 2.27e-3 2.87e-3
# ... with 8 more variables: Male...10 <dbl>, Female...11 <dbl>,
#   `Both sexes...12` <dbl>, Male...13 <dbl>, Female...14 <dbl>,
#   `Both sexes...15` <dbl>, Male...16 <dbl>, Female...17 <dbl>, and
#   abbreviated variable names 1: `Age Group`, 2: `Both sexes...3`,
#   3: Male...4, 4: Female...5, 5: `Both sexes...6`, 6: Male...7,
#   7: Female...8, 8: `Both sexes...9`
```

```
xmart_yrs <- read_csv("data-raw/xmart.csv")
```

```
New names:
Rows: 134 Columns: 17
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(17): ...1, ...2, 2019...3, 2019...4, 2019...5, 2015...6, 2015...7, 2015...
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
* `` -> `...2`
* `2019` -> `2019...3`
* `2019` -> `2019...4`
* `2019` -> `2019...5`
* `2015` -> `2015...6`
* `2015` -> `2015...7`
* `2015` -> `2015...8`
* `2010` -> `2010...9`
* `2010` -> `2010...10`
* `2010` -> `2010...11`
* `2005` -> `2005...12`
* `2005` -> `2005...13`
* `2005` -> `2005...14`
* `2000` -> `2000...15`
* `2000` -> `2000...16`
* `2000` -> `2000...17`
```

```
xmart_yrs <- xmart_yrs[-1,]%>%
  janitor::clean_names()%>%
  pivot_longer(cols=3:17,names_to="years",values_to="values")%>%
  mutate(values=as.numeric(values))
```

```
xmart_yrs %>% names
```

```
[1] "x1"     "x2"     "years"  "values"
```

```
xmart_yrs
```

```
# A tibble: 1,995 x 4
   x1                                                    x2         years  values
   <chr>                                                 <chr>      <chr>   <dbl>
 1 nMx - age-specific death rate between ages x and x+n &lt;1 year x2019~ 0.0478
 2 nMx - age-specific death rate between ages x and x+n &lt;1 year x2019~ 0.0513
 3 nMx - age-specific death rate between ages x and x+n &lt;1 year x2019~ 0.0441
 4 nMx - age-specific death rate between ages x and x+n &lt;1 year x2015~ 0.0566
 5 nMx - age-specific death rate between ages x and x+n &lt;1 year x2015~ 0.0602
 6 nMx - age-specific death rate between ages x and x+n &lt;1 year x2015~ 0.0527
 7 nMx - age-specific death rate between ages x and x+n &lt;1 year x2010~ 0.0675
 8 nMx - age-specific death rate between ages x and x+n &lt;1 year x2010~ 0.0723
 9 nMx - age-specific death rate between ages x and x+n &lt;1 year x2010~ 0.0625
10 nMx - age-specific death rate between ages x and x+n &lt;1 year x2005~ 0.0822
# ... with 1,985 more rows
```

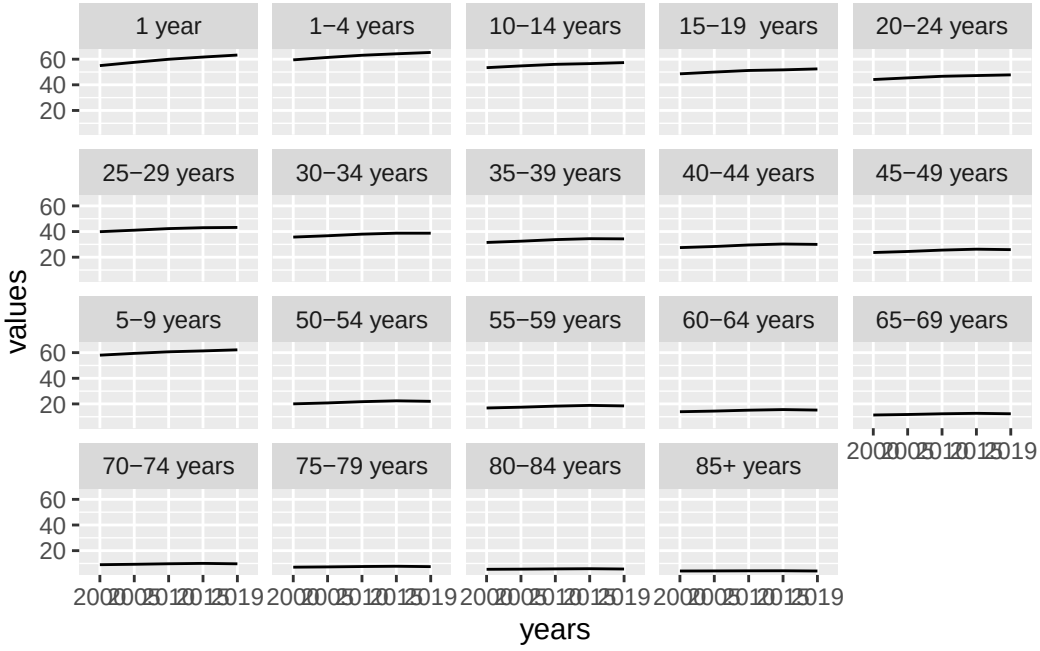
```
xmart_tidy <- xmart %>%
  janitor::clean_names()%>%
  pivot_longer(cols = 3:17,names_to="sex",values_to="values") %>%
  full_join(xmart_yrs,by=c("indicator"="x1","age_group"="x2","values")) %>%
  mutate(age_group=sub("&lt;","",age_group),
         sex=gsub("_\\d+","",sex),
         sex=ifelse(sex=="both_sexes","both",sex),
         years=sub("x","",years),
         years=gsub("_\\d+","",years))


xmart_tidy%>%count(indicator)
```

```
# A tibble: 7 x 2
  indicator                                             n
  <chr>                                             <int>
1 ex - expectation of life at age x                   285
```

```
2 lx - number of people left alive at age x              495
3 ndx - number of people dying between ages x and x+n    285
4 nLx - person-years lived between ages x and x+n        285
5 nMx - age-specific death rate between ages x and x+n   285
6 nqx - probability of dying between ages x and x+n      495
7 Tx - person-years lived above age x                    285
```

```
xmart_tidy %>%
  filter(sex=="both",
         indicator=="ex - expectation of life at age x")%>%
         #age_group=="1 year")%>%
  ggplot(aes(years,values,group=indicator))+
  geom_line()+
  facet_wrap(vars(age_group))
```



3.1.2 Life expectancy

The life expectancy rates are calculated with consideration of the probability of survival based on key parameter such as age, and deaths probabilities for that age. More info about how to calculate the life expectancy can be found #sec-tools of this book.

3.2 How to build the metrics

In this section a practical calculation of the health metrics is done for the practitioner to be able to replicate this calculation for further analysis based on these key elements.

3.2.1 YLLs

The number of years of life lost YLLs is the first of the three metrics that is calculated, and is important for releasing a first look at the status of a population. It is calculated for identifying the area where improvement is required for reducing the loss in health status and clearly reducing the probability of death.

Reiner and Hay [1]

3.2.2 YLDs

3.2.3 DALYs

3.3 How to use the metrics

4 Causes and risks

- How to use the metrics
- Overview of the causes and risks

5 Healthy life expectancy (HALE)

- Description and calculation

It adjusts overall life expectancy by the amount of time lived in less than perfect health. This is calculated by subtracting from the life expectancy a figure which is the number of years lived with disability multiplied by a weighting to represent the effect of the disability.[1]

More info [2]

[1] www.healthknowledge.org.uk

[2] about the methodology https://www.un.org/esa/sustdev/natlinfo/indicators/methodology_sheets/health/health_life_expectanc

Part II

Modeling

- Feature engineering
- Model selection
- Packages and functions to use for the analysis
- Attempt at predicting the future

6 Techniques

7 Packages and functions

8 Predicting the future

Part III

Data Visualizations

9 Application of the model results

10 Spatial data modeling

11 Examples of data visualizations

::: {.quarto-book-part}

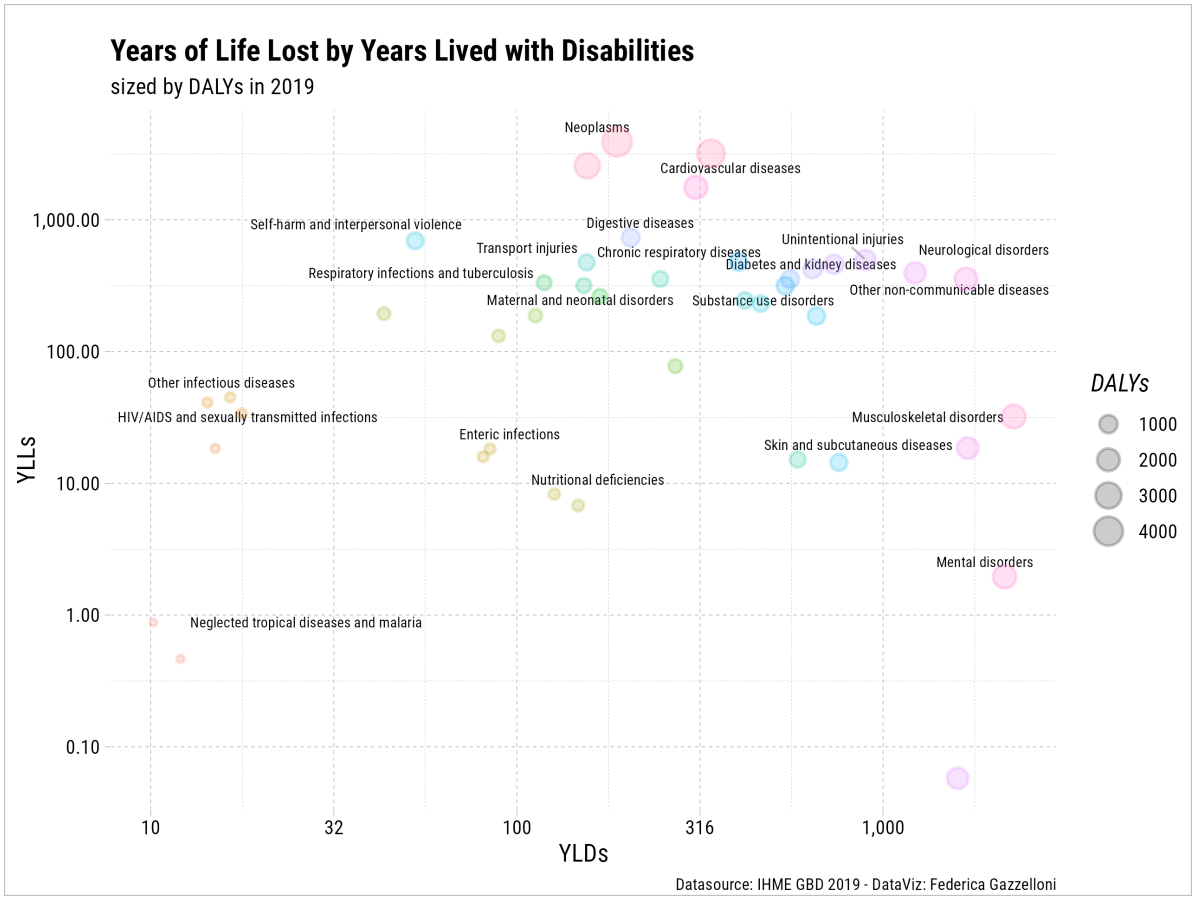
Part IV

Case Studies

In this section we will be looking at two case studies:

- Covid19 spread of infection as a cause for increased DALY for selected countries.
- The state of health of selected countries.

A systematic review study namely **The state of health in the European Union in 2019**, **João Vasco Santos et. all** summarized the results of all studies containing DALYs, YLLs and YLDs metrics for the European countries from 2010 to 2019. The result of the review shown a steady improvement of the DALY metrics in almost all countries of Europe. Cardiovascular and neoplasms account for most of deaths and increase in numbers of years lost in EU 2019. Important improvements in some areas (e.g. transport injuries) are opposite ar the increasing level of diabetes spread in the population.



:::

12 Covid19

13 The state of health

14 Summary

Conclusions

References

https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2019_cod_metho

life tables:

- https://apps.who.int/gho/data/node.main.LIFECOUNTRY?lang=en
- https://ghdx.healthdata.org/record/ihme-data/gbd-2019-life-tables-1950-2019

A Life tables and Life expectancy

B Tools used to make this book

To set up the environment for replicating the code used in this book the R language is needed as well as R and Rstudio IDE environments. The following sections contain the directions for installing **R** and **RStudio**, how to set up a book with **quarto** and how to use **GitHub** as a version saver source.

B.1 RStudio installation

Download and install R: Download and install RStudio IDE:

B.2 Info on how to setup this project in quarto

quarto is the new version of **Rmarkdown**, it can be used for making notes, presentations, websites, books and more.

In this project the book has been made in quarto and version saved on github.

https://quarto.org/docs/publishing/github-pages.html

B.2.1 GitHub useful commands

In RStudio create a new project on a new directory and in terminal type:

```
add git
quarto book project
```

The automated process will create a `_quarto.yml` file, the top of the file will look like this one:

```
project:
  type: book
```

On terminal type: `quarto preview`

It creates a folder `_book`

B.2.1.1 github later:

source: https://happygitwithr.com/existing-github-last.html

connect with github

create a github repo with the same name from github website

then type

usethis::use_git()

this pushes all files in R to a remote folder designed to github repo

in terminal connect with github repo

```
git init
git remote add origin
```

https://github.com/Fgazzelloni/infectious.git

```
git branch -M main
git push -u origin main
```

B.2.2 Publish your book on github pages

change the `quarto.yml` file into:

```
project:
  type: book
  utput-dir: docs
```

add a .nojekyll file …(terminal)

```
touch .nojekyll
```

then type

```
quarto render
```

some issues might arise if more than one calculation is made inside a single cunck split the cuncks!

quarto render creates a folder `docs`

Bibliography

[1] Robert C. Reiner and Simon I. Hay. "The overlapping burden of the three leading causes of disability and death in sub-Saharan African children". In: *Nature Communications* 13.1 (Dec. 6, 2022). Number: 1 Publisher: Nature Publishing Group, p. 7457. DOI: 10.1038/s41467-022-34240-6. URL: https://www.nature.com/articles/s41467-022-34240-6.