Geofaceting – allign small-multiples for regions in a spatially meaningful way

Ilya Kashnitsky

1. University of Groningen / NIDI
2. National Research University Higher School of Economics
ilya.kashnitsky@gmail.com; ikashnitsky.github.io

Jose Manuel Aburto

1. University of Southern Denmark
jmaburto@health.sdu.dk

2018-06-03

Abstract

In this paper we present a handy visualization technique of geofaceting, which was recently streamlined in geofacet R package. The idea is to align small-multiples for territorial units, often regions, according to approximate geographical location. Compared to other ways of small-multiples' arrangement, geofacetting improves the speed of regions' identification and exposes the broad spatial pattern. We illustrate the technique using data on young adult mortality in the 32 Mexcian states from 1990 to 2015.

Keywords: small-multiples; geofacet; ternary colorcoding; external mortality; homicide.

1   Introduction

In data visualization, it is often very challenging to represent multiple relevant dimensions preserving readability of a plot. This is especially true, when the task is to expose spatial variation of some complex phenomena. In such a case, geographical maps are the natural choice of a visualization framework – they are meant to show spatial patterns. Yet, the usual difficulty is that one can only represent one variable meaningfully with a choropleth utilizing colors. And what if the dataset at hand is much more complex and demands a balanced exposure of several dimensions?

Usually, the difficult to represent but very important for the story telling dimension of the data is time. Visualizing time series with choropleths is very challenging. One has to produce either small multiples for the years or animated pictures with maps for various years flashing sequentially. Both variants makes it difficult to compare regions across years, which is, in fact, the main goal of such a visualization. But what if one also needs to represent in addition yet another variable? In demographic studies, such super important variable is often age. In such a case, the basic choropleth visualization framework fails. And here comes *geofaceting*.

The idea of geofaceting is brilliantly simple: a "normal" plot is produced for each of the regions, and then all the small panels are arranged according to their approximate geographic location thereby making it easier to identify regions. The spatial logic of small-multiples alignment helps to identify the units of analysis,

usually regions of a country, faster and reveals the macro-level spatial pattern while keeping the flexibility of visualization technique choice for the small-multiples themselves. As a result, creating geofaceted visualizations gives all the advantages of standard plots in which one can easily display at least three dimensions of the dataset.

It seems that geofaceting has been invented and re-invented multiple times. Examples of a closely related idea to overlay small plots over geographical map can be seen in the works of 19-th century geographers (Palsky 1996). Geofaceting goes one step closer dropping the actual geographical map and just ordering the small-multiples in line with the spatial pattern of the corresponding areas. The approach was only recently formalized by Ryan Hafen, received its name and was consistently implemented in an R package `geofacet` (Hafen 2017). To put it in author's words:

> What's new about this R package is that it formalizes the "geofaceting" approach, gives it a name, and makes it available in a user-friendly way. Also, it provides the basis for creating a library of community-contributed grids, which can be used elsewhere outside the package. (Hafen 2018)

2 Data & Methods

2.1 Data

The application of our visualization proposal relies on the results from Aburto, Riffe, and Canudas-Romo (2018). These are based on cause-of-death information available from the Mexican Statistical Office from 1990 to 2015, and population estimates from the Mexican Population Council (Instituto Nacional de Estadística y Geografía 2015). Data were desegregated by single age, sex, and state. Population estimates were adjusted for age misstatement, undercounting, and interstate and international migration.

2.2 Demographic Methods

Cause-specific death rates were smoothed over age and time for each state and sex separately using the 2-d p-spline to avoid random variations between ages (Camarda 2012). Smoothed death rates were then constrained to sum to the unsmoothed all-cause death rates. Period life tables were constructed for males and females from 1990 to 2015 following standard demographic methods (Preston, Heuveline, and Guillot 2001). We calculate the average years lived between ages 15 and 49, temporary life expectancy (Arriaga 1984), and estimate cause-specific contributions to the difference between state-specific temporary life expectancy and a low mortality benchmark using standard decomposition techniques (Horiuchi, Wilmoth, and Pletcher 2008).

2.3 Low mortality benchmark

The low-mortality benchmark is calculated in the basis of the lowest observed mortality rates by age, cause of death, from among all states for a given sex and year. The resulting minimum mortality rate schedule has a unique age profile, and it determines our benchmark temporary life expectancy. The minimum mortality schedule can be treated as the best presently achievable mortality assuming perfect diffusion of the best available practices and technologies in Mexico (Vallin, Meslé, and Divinagracia 2008).

2.4 Visualization techniques

There exists substantial regional variation in young male mortality across Mexican states, therefore to visualize properly mortality patterns, it is necessary taking into account the spatial dimension of the dataset. For that we use *geofaceting*, a visualization technique recently streamlined in R with the package `geofacet` (Hafen 2017). As there was no geofacet layout for Mexico, we created it from the scratch. The produced grid for Mexican states was successfully submitted to `geofacet` package (Kashnitsky 2017).

To expose the various relevant dimension of the dataset at hand, we decided to produce three plots. The first two (Figures 1 and 2) complement each other. Figure 1 uses *stacked bar plot* technique to reveal the variation of young adult mortality in Mexican states over time. In Figure 2 we have created a separate *heatmap* plotted over a standard *Lexis surface* (Rau et al. 2017; Schöley and Willekens 2017) for each of the Mexican states; color shows the dominant cause of death.

Yet, focusing on one leading cause of death may mask it's relative importance compared with the second, third, and so on. Thus, in Figure 3 we apply the framework of *ternary colorcoding*, which was recently formalized and streamlined in the R package `tricolore` (Schöley 2018). Ternary colorcoding maximizes the amount of information conveyed by colors through representing each element in a three-dimensional array of compositional data with a single color. Each part of the ternary composition is assigned a hue (color characteristic), and the amount of hue for each data element is proportional to its weight in the ternary composition. For more technical details on the method check Schöley (forthcoming); for an indicative use case of ternary colorcoding see Kashnitsky and Schöley (forthcoming).

2.5 Reproducibility

The figures presented in this paper are easily reproducible using the replication material that we provide openly (Kashnitsky and Aburto 2018). R programming language (R Core Team 2018) was used for the analyses and data visualization; in addition to core packages we used: `tidyverse` (Wickham 2017), `tricolore` (Schöley 2018), `ggtern` (Hamilton 2017), `hrbrthemes` (Rudis 2018), `extrafont` (Chang 2014), `RColorBrewer` (Neuwirth 2014), and `geofacet` (Hafen 2017)

3 Application

To show the usefulness of our proposal, we analyze the contribution of homicide, road traffic accidents, and suicide, medically amenable mortality, and causes amenable to health behavior to the gap in temporary life expectancy between ages 15 and 49 of each of 32 Mexican states with a low mortality benchmark. These causes have emerged as leading among young people, and the first two recently had a sizable impact on life expectancy in Mexico (Aburto et al. 2016).
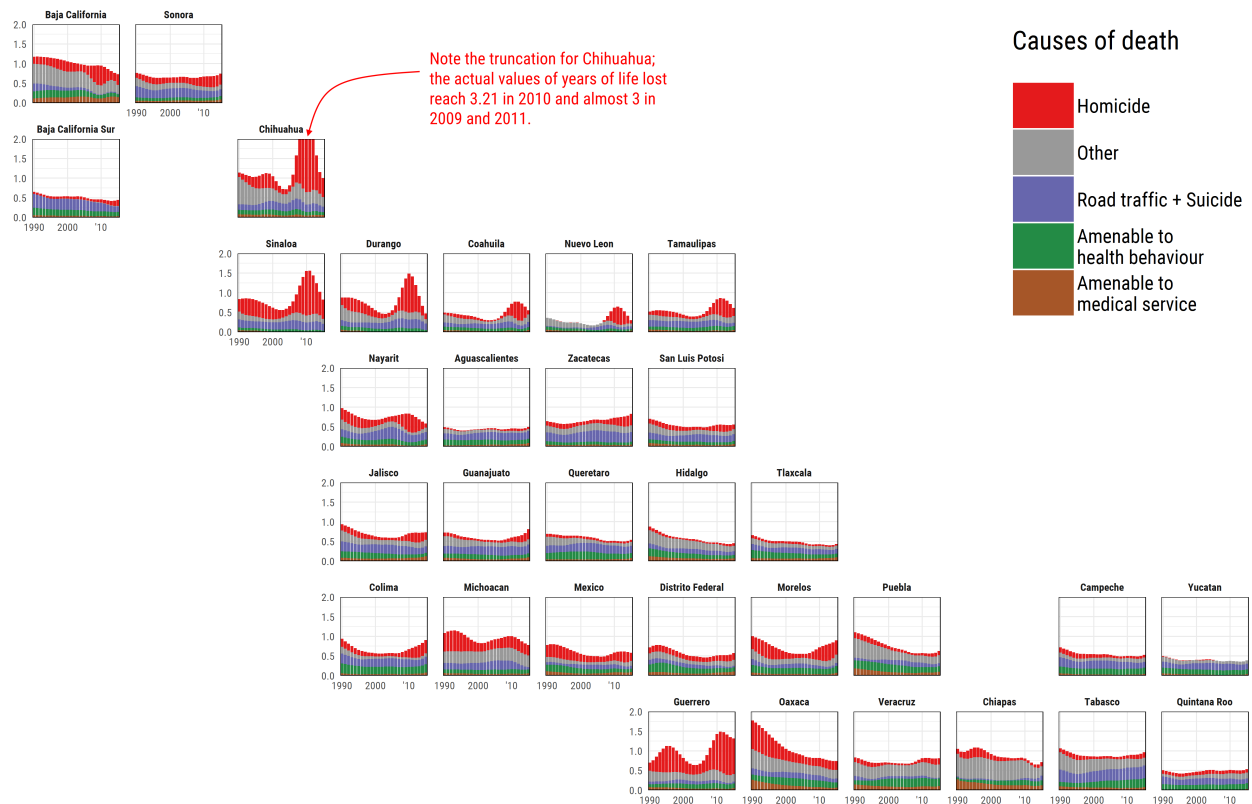
Three complementary geofaceted plots were created. Figure 1 shows the absolute impact of five causes of death on the difference between the observed life expectancy with the best-practice life expectancy (low benchmark) for young males. For example, it shows how the contribution of homicides (red color) increased substantially after 2005, particularly in the North, reaching a peak in 2011 for Chihuahua, Sinaloa, Durango, among others. It is also clear from this graph that the most affected state in the South is Guerrero since the early 1990s.

Figure 2 shows state-specific Lexis diagrams with the main cause of death at each age in a given year. It gives a full representation of the the main cause of death by age and period compromising on the actual values of the gap, i.e. Figure 1. For example, from this graph it is clear that homicides are contributing the most across ages between 15 and 49 in most states in the North. However, even though in Oaxaca (in the South) the contribution of homicide was decreasing (Figure 1), between ages 20 and 30 homicide remained the main contributor to the gap.

To enrich the plot with geofaceted Lexis surfaces (Figure 2), we use ternary colorcoding of the three main groups of causes of death: homicides, road traffic and suicides, and all other causes combined (Figure 3).

Basically, this plot highlights similar points as Figure 2, only it provides an opportunity to appreciate the relative importance of the two main causes of death compared with all others pulled together. For example, if we compare Federal District (the one, where the city of Mexico is located) with the neighboring Guerrero, their mortality pattern at ages 20-30 seem very similar if we look at Figure 2 and focus only at the leading cause of death, homicide. Yet, when we consider the relative importance of homicide in the mortality regime
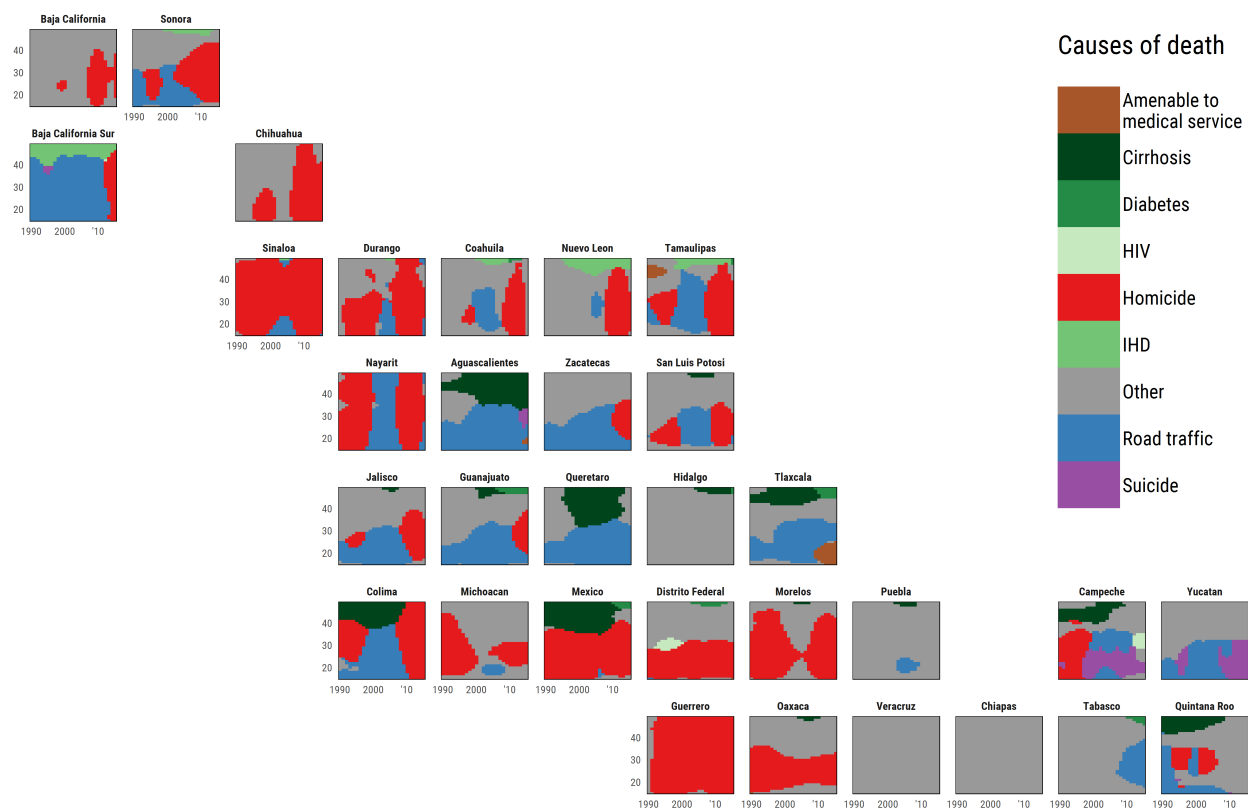
Gap between observed and best-practice temporary life expectancy for Mexican males (15-49)
Years of life lost by cause of death across time (1990-2015)

Note the truncation for Chihuahua; the actual values of years of life lost reach 3.21 in 2010 and almost 3 in 2009 and 2011.

Causes of death
- Homicide
- Other
- Road traffic + Suicide
- Amenable to health behaviour
- Amenable to medical service

Ilya Kashnitsky and Jose Manuel Aburto, 2018; replicate: https://github.com/ikashnitsky/demres-2018-geofacet

Figure 1: Gap between observed and best-practice life expectancy for Mexican states: Cause of death contributing the most by age (15-49) and time (1990-2015).

Gap between observed and best-practice temporary life expectancy for Mexican males (15-49)
Cause of death contributing the most by age (15-49) and time (1990-2015)

Ilya Kashnitsky and Jose Manuel Aburto, 2018; replicate: https://github.com/ikashnitsky/demres-2018-geofacet

Figure 2: Gap between observed and best-practice life expectancy for Mexican states: Years of life lost by cause of death across time (1990-2015).

Gap between observed and best-practice temporary life expectancy for Mexican males (15-49)

Colorcoded ternary compositions of the three leading causes of death by age (15-49) and time (1990-2015)

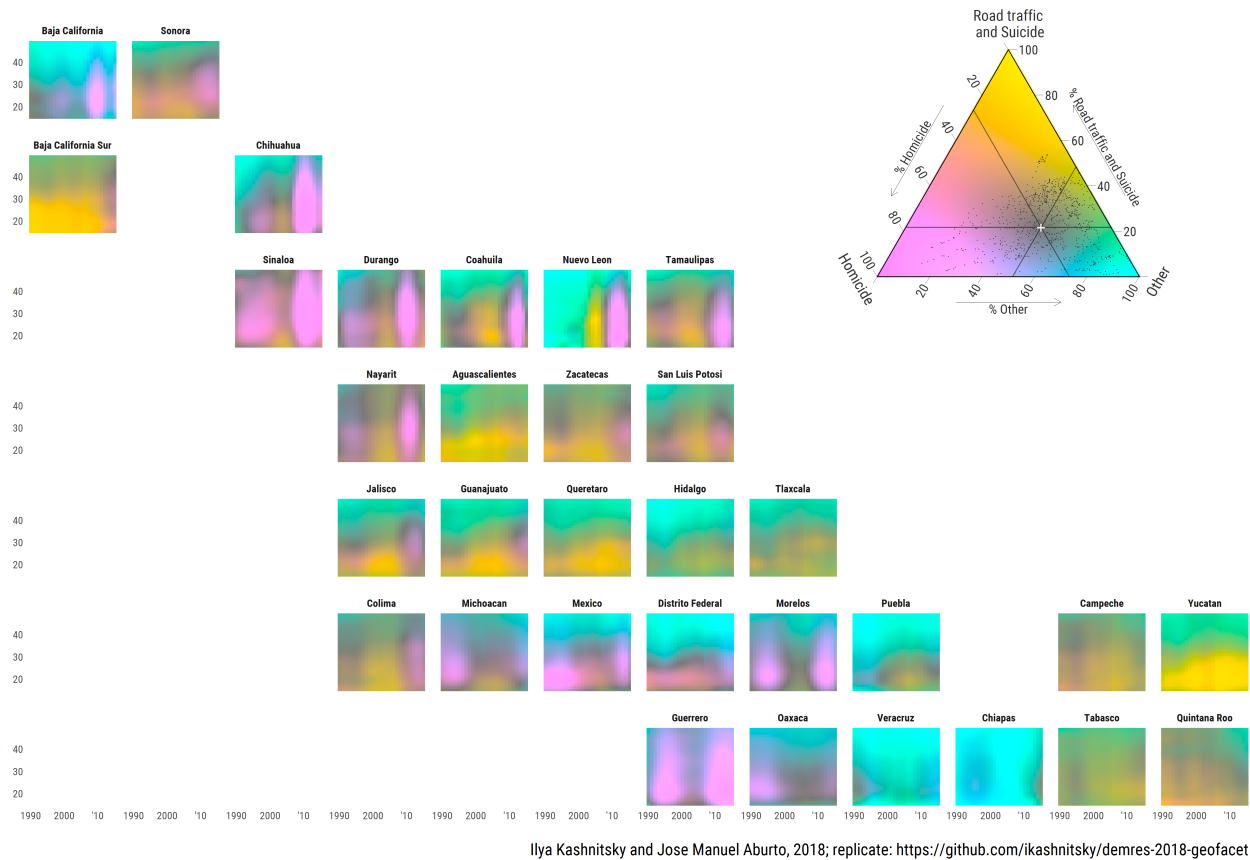Ilya Kashnitsky and Jose Manuel Aburto, 2018; replicate: https://github.com/ikashnitsky/demres-2018-geofacet

Figure 3: Gap between observed and best-practice life expectancy for Mexican states: Colorcoded ternary compositions of the three leading groups of causes of death by age (15-49) and time (1990-2015).

of the two states (Figure 3), it becomes clear that homicide is by far a much bigger problem in the state of Guerrero.

4   Discussion

Geofaceting regions to analyze different dimensions of a phenomenon can improve on their graphical representation. Here, we propose using a multi-panel arrangement according to regional location to depict different dimensions of a dataset. We demonstrate its usefulness showing the specific case of Mexico and mortality patterns over a fairly large period, 1990-2015. The main advantage of our proposal is that the reader can easily interpret complex phenomena, while being able to identify regional variations. This is particularly important in the case of young males in Mexico that have experienced an unprecedented period of rising homicidal mortality. Moreover, the changing dynamics of violence in the country is a dimension hard to represent graphically, nevertheless with the geofaceting framework the reader can easily get a sense of this. For example, while most of the historically violent states are in the northern part of the country, in figures 1-3, an upsurge of violence in the South is clear, albeit with different intensities, i.e. absolute gap between states and best-practice life expectancy. Being able to identify variations regionally, but also in terms of intensity, is a great advantage of proposed visualization technique.

There are some limitations of the approach. For example, if a territory is divided in a large, or very small, number of regions the geofaceting might not be the ideal approach to show complex phenomena. Moreover, if a territory is oddly shaped, or unevenly distributed, getting the proper regional representation might be impossible. Nevertheless, we believe that in the case shown here, geofaceting was a useful tool, which combined with ternary colorcoding provided a macro-representation of a phenomenon, i.e. cause-specific contributions to the gap between states and best-practice life expectancy, and still being able to account for regional variations.

5   Acknowledgement

References

Aburto, J.M., Beltrán-Sánchez, H., García-Guerrero, V.M., and Canudas-Romo, V. (2016). Homicides In Mexico Reversed Life Expectancy Gains For Men And Slowed Them For Women, 2000–10. *Health Affairs* 35(1):88–95. doi:10.1377/hlthaff.2015.0068.

Aburto, J.M., Riffe, T., and Canudas-Romo, V. (2018). Inequalities and deterioration in average lifespan among adults in Mexico, 1990-2015: A cross-sectional demographic cause-of-death analysis. *Forthcoming.*

Arriaga, E.E. (1984). Measuring and explaining the change in life expectancies. *Demography* 21(1):83–96. doi:10.2307/2061029.

Camarda, C.G. (2012). MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines. *Journal of Statistical Software* 50(1):1–24. doi:10.18637/jss.v050.i01.

Chang, W. (2014). *Extrafont: Tools for Using Fonts.* https://CRAN.R-project.org/package=extrafont.

Hafen, R. (2017). *Geofacet: 'Ggplot2' Faceting Utilities for Geographical Data.* https://cran.r-project.org/package=geofacet.

Hafen, R. (2018). Introducing geofacet [electronic resource]. http://ryanhafen.com/blog/geofacet.

Hamilton, N. (2017). *Ggtern: An Extension to 'Ggplot2', for the Creation of Ternary Diagrams.* https://CRAN.R-project.org/package=ggtern.

Horiuchi, S., Wilmoth, J.R., and Pletcher, S.D. (2008). A decomposition method based on a model of continuous change. *Demography* 45(4):785–801. doi:10.1353/dem.0.0033.

Instituto Nacional de Estadística y Geografía (2015). Deaths microdata.. http://www3.inegi.org.mx/sistemas/microdatos.

Kashnitsky, I. (2017). New grid: 'Mex_grid1' · Issue #31 · hafen/geofacet [electronic resource]. https://github.com/hafen/geofacet/issues/31.

Kashnitsky, I. and Aburto, J.M. (2018). Geofaceting: Allign small-multiples for regions in a spatially meaningful way – replication materails.. https://github.com/ikashnitsky/demres-2018-geofacet.

Kashnitsky, I. and Schöley, J. (forthcoming). Regional population structures at a glance. *The Lancet.*

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes.* https://CRAN.R-project.org/package=RColorBrewer.

Palsky, G. (1996). *Des chiffres et des cartes: Naissance et développement de la cartographie quantitative française au XIXe siècle.* Ministère de l'enseignment supérieur et de la recherche, Comité des travaux historiques et scientifiques.

Preston, S.H., Heuveline, P., and Guillot, M. (2001). *Demography: Measuring and Modeling Population Processes.* Oxford ; Blackwell Publishers.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rau, R., Bohk-Ewald, C., Muszyńska, M.M., and Vaupel, J.W. (2017). *Visualizing Mortality Dynamics in the Lexis Diagram.* Springer.

Rudis, B. (2018). *Hrbrthemes: Additional Themes, Theme Components and Utilities for 'Ggplot2'.* https://CRAN.R-project.org/package=hrbrthemes.

Schöley, J. (2018). *Tricolore: A Flexible Color Scale for Ternary Compositions.* https://cran.r-project.org/package=tricolore.

Schöley, J. (forthcoming). The centered ternary balance scheme: A technique to visualize surfaces of unbalanced three part compositions. *Demographic Research.*

Schöley, J. and Willekens, F. (2017). Visualizing compositional data on the Lexis surface. *Demographic Research* 36(21):627–658. doi:10.4054/DemRes.2017.36.21.

Vallin, J., Meslé, F., and Divinagracia, E. (2008). Minimum Mortality: A Predictor of Future Progress? *Population* 63(4):557–590. https://www.cairn-int.info/article-E_POPU_804_0647--minimum-mortality-a-predictor-of-future.htm.

Wickham, H. (2017). *Tidyverse: Easily Install and Load the 'Tidyverse'.* https://CRAN.R-project.org/package=tidyverse.