

Identifying Demographic Changes in Europe: Empirical Application of Soft Clustering

Marcell P. Granát and Marcell Örsi

Advisor: Zoltán Madari

Corvinus University of Budapest, Hungary

Abstract

Our goal is to identify demographic characteristics of European regions (NUTS-2) focussing on fertility and mortality rates, life expectancy, and female employment rate. Traditionally, sorting areas into groups is done by clustering. Clustering is an unsupervised machine learning technique that divides the population into several groups in which points in the same group are similar to each other and points in different groups are dissimilar. Repeating the algorithm in each year would cause confusing jumps in the cluster centres. Thus, instead of putting each data point into separate clusters, the probability of that point being in that cluster is assigned. In soft clustering or fuzzy clustering, each data point can belong to multiple clusters along with its probability score or likelihood.

Keywords: Fuzzy clustering, Soft clustering, Demographic trends, Fertility rate

1. Introduction

Decreasing fertility rates are one of Europe's biggest problems in the early decades of the 21st century. After World War II, the willingness to have children in developed countries declined significantly. By the 1970s and 1980s, the phenomenon had reached the countries of the socialist bloc, which had been an exception until then (World Bank 2022). Fertility rates reached another low point in the 1990s and then in the 2000s, and some countries had values below 1.2. The dramatic situation is well illustrated by the statement from D'Addio and d'Ercole (2006), that in the first years of the twentieth century only Turkey and Mexico had achieved a fertility rate of 2.1 for self-reproduction. It is also important to note that the total fertility rate of 2.1 is only sufficient in the most developed countries. In some less developed African countries, the required fertility rate can reach 4.9 according to Espenshade, Guzman, and Westoff (2003). In the 2010s, the decline had slowed or stopped in most countries, and in some places there was significant growth (KSH 2022), but in the same time the number of children born did not grow.

At an age when the world population is expected to grow by 40% in the next 80 years, the question arises of whether it is worth addressing the declining fertility of developed countries. However, it is important to note that according to United Nations (2019), half of the children are expected to be born in Africa in 2100, which, together with the warming of the continent, will significantly increase the risk of mass migration to Europe and Asia. However, it is particularly important to

address the dramatic fertility situation in Europe. The reason is that although many people will come to Europe, declining fertility is making a significant contribution to the ageing of society (Botev 2012). Although the constant increase in life expectancy is also a significant factor in the latter, this change can be assessed as positive in contrast to declining fertility.

From an economic point of view, the development of fertility is extremely important due to the positive relationship between population growth and economic growth. The political leadership of several European countries has decided to promote childbearing through economic incentives (Neyer, Caporali, and Gassen 2017). These incentives are primarily aimed at having children, which can violate the principle of equality, as they do not consider those who cannot have children for health reasons. However, after intervening with incentives, fertility rates grew significantly in numerous countries (e.g. Hungary).

However, the most important argument in favour of economic incentives can be found in the individual preferences of potential parents. The number of children desired and born in developed countries differs significantly, which parents have largely explained by factors related to financial and quality of life (D'Addio and d'Ercole 2006; Günther and Harttgen 2016). On the contrary, the problem in developing countries is that around 20% of the children born due to deficiencies in contraception and sexual education were unwanted, and, in fact, there are countries (Bolivia, Swaziland, Lesotho) where this rate is close to 40% (Westoff 2010). This duality confirms that, in addition to solving the problems of developing countries, the low fertility of developed countries must also be addressed at the individual level.

Addressing the problem requires a complete overview of all the factors that have a significant effect on the individual decision about whether or not to have children. However, it is important to note that despite all the effort there is a chance that even if fertility rates start to grow significantly, they will remain below replacement level in the long term (Lesthaeghe and Willems 1999).

Education and employment are crucial factors in this decision. Logically, those who have invested almost two decades into their education and spent five years at a university instead of making money are less willing to give up all their labour market opportunities to have children in their 20s. In developed countries, women generally have a higher proportion of access to secondary and higher education, so there are more opportunities for them on the labour market. However, women often suffer disadvantages compared to their male counterparts with the same qualifications. These come in part from their role in the family, but in part they suffer from unfavourable assumptions from employers about women. These include the issue of having a child; as in the case of a woman of childbearing age, employers often take the possibility of pregnancy into account when determining their salary, as the replacement might cause the company additional expenses. Higher average educational attainment amplifies these effects, as women must give up higher salaries and better career prospects if they choose to have children. Making career development more attractive thus contributes greatly to the postponement of childbearing and lower fertility in developed countries.

On the other hand, people also want to have a stable financial background before having children to be able to provide them with a high level of education and opportunities. To understand this part of the decision, it is important to look at the distribution of incomes in a society because that has a major effect on the number of children being born. In some southern European countries (e.g. Greece, Spain), youth unemployment is a serious issue, which further strengthens the postponement of childbearing.

There is also an inverse relationship between lower fertility and the ageing of society. The lower number of children born will cause a higher mean age of society in the long term, assuming that life expectancy will be higher over time. This means that after a generation, the proportion of women of childbearing age will be lower, which together with their higher employment rate will cause a further decline in fertility. This is what is called an *aging spiral* by Lennert (2019), which in the long term will make the pension system unsustainable (Heinrich 1997).

The purpose of this study is to provide an exploratory analysis of the trends mentioned above between European regions. More specifically, how have fertility rates, ageing, life expectancy, the proportion of women of childbearing age, and female employment changed over the past decades and can these be classified into demographic clusters? Our contribution to the literature is that we use soft clustering methodology to avoid radical, unexplained jumps in the middle of groups and members, as well.

2. Explanatory data analysis

In this study, our main objective is to identify regional patterns in the EU along demographic and economic indicators. To achieve this goal, we used European regional data, which are easily accessible in the Eurostat databases for all levels of NUTS regions. The definitions of the NUTS levels given by Eurostat (2022) are shown in *Table 1* below.

Table 1. Definitions of the NUTS region levels (Eurostat 2022)

Region level	Definition
NUTS-1	major socio-economic regions
NUTS-2	basic regions for the application of regional policies
NUTS-3	small regions for specific diagnoses

Based on these definitions, we chose NUTS-2 level data to perform our analyses. The main aspect of the decision was the heterogeneity of the different levels. The population of NUTS-1 regions is approximately 5 million, making it impossible to separate urban and rural regions. Unlike the extremely heterogeneous NUTS-1 regions, NUTS-3 regions are mostly homogeneous and have the same characteristics. However, in this case, many of the data are incomplete; therefore, we chose NUTS-2 regions as the basis of our analyses. These have lower heterogeneity compared to NUTS-1 regions and are also the basis of regional policies, making it easier to find regional patterns among them.

To perform our analyses, we chose five demographic variables based on the relevant literature. Our data cover 30 years between 1990 and 2020. *Total fertility rate* (tfr) expresses the number of children a woman is expected to give birth to between the ages of 15 and 50 given the current fertility structure. However, it is important to note that it is unable to measure the effect of postponement in childbirth, which is an important issue according to Sobotka (2004). *Life expectancy* (lifeexp) shows the expected average age of death of those who are born now, based on the current death structure. Although this is inaccurate, because the average number of years lived is expected to grow over time, life expectancy is a useful measure of general living conditions (e.g. health, war). *Ageing index* (ageing) is the usual way to compare the proportion of the population older than 65 years and younger than 14 years. In a society with growing population, this index is well below 1, as there are more people under 14 than over 65 years. The *rate of women of childbearing age* (motherrate) indicates well the other aspect of ageing. If there are fewer women of this age in a society, then in the long term less children will be born even with a higher total fertility rate, which explains why there is a delayed relationship between TFR and the number of children born. The *percentage of women employed* (emp) shows the general state of women aged 15 to 64 years on the labour market. Although studying for women aged 15 to 25 is an alternative that is not taken into account, this ratio shows a satisfying image of the state of women.

Before performing any analysis, we looked at the main characteristics (mean, minimum, maximum) of the distributions of our variables. The results are shown in *Table 2*.

The mean values are close to what was expected prior and support the fact of a demographic crisis. The mean TFR of 1.551 shows that fertility in Europe is far from 2.1 what is needed for reproduction.

Table 2. Descriptive statistics of the demographic variables

	mean	min	max
tfr	1.551	0.800	4.950
lifexp	78.998	66.600	85.800
ageing	1.099	0.059	2.573
motherrate	0.461	0.321	0.603
emp	58.349	21.300	80.500

Life expectancy is close to 79 years, which together with the TFR contributes significantly to the ageing of society. The measure of ageing is well shown by the fact that on average there are 0.99% more people in the age group over 65 years than in the age group under 14. Although this may seem not that big of a difference, it is significant compared to countries outside Europe. (In Africa, for example, the ageing index is around 0.086.)

It is important to understand the extreme values of the distributions. The large spread of values for life expectancy and employment of women was expected, but the other three variables show how big the European demographic crisis really is. If we do not count the overseas regions of France and Spain, there are not many regions that have a TFR above 2.0, or an ageing index below 0.4. These extreme values seem to be atypical considering the characteristics of continental Europe, but the most shocking values are on the opposite ends of these scales. A TFR around 0.8 and an ageing index above 2.5 draw a shocking picture of one of the possible outcomes of the current processes in Europe. This is one of many reasons why it is extremely important to understand the regional characteristics and patterns in the European demographic crisis.

3. Methodology

Clustering is a methodology of unsupervised machine learning procedures where the aim is to sort the observations into disjoint subsets so that the members of the set are as homogeneous as possible to each other, but as heterogeneous as possible from others.

A frequently applied technique is the k-means clustering. In this framework k points are randomly selected and in each iteration points are assigned to the cluster that has the closest centre. Centres are calculated as the average of observations belonging to the given cluster; thus, they have the same number of dimensions as the number of variables used for clustering. The difference is most commonly measured with the Euclidean distance (Equation 1).

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{j=1}^p (a_j - b_j)^2}, \quad (1)$$

where the distance is calculated between the a and b points and p is the number of dimensions. These measurements would rely on the scale of the variables; thus, standardisation of the variables is suggested. Iterations when points are assigned to the cluster which has the closest centre are repeated, while the decrease in the R^2 is smaller than the given hyperparameter (or reaches an upper limit of the number of iterations).

The target function is to maximise the R^2 (Equation 2).

$$R^2 = 1 - \frac{SSE}{SST}, \quad (2)$$

where SSE denotes the total sum of squared distances between the points and the cluster centre

belonging to them, while the SST denotes the total sum of squared distance of the average of the observations.

Since the initial selected points are random, the outcome varies at each modelling, even if the data are the same. The typical solution is to run K-means clustering N times and choose the one with the highest R^2 (Likas, Vlassis, and Verbeek 2003).

However, using K-means clustering, the problem of "jumping" emerges if the model runs on a longitudinal (panel) dataset. Academics often apply the K-means clustering method as the model is run on subsets separated annually (see, for example, Benedek, Kiss, and Bakucs (2017)).

By "jumping", we mean that even a small change in the data in the case of annual calibration can move observations from one cluster to another for the next year. An extreme case of this is when the a data point changes (significant displacement, disappearance, or appearance) and this causes change in the cluster centre; as a result, another data point is now closer to another cluster centre, even if none of its variables changed. Figure 1 demonstrates a simulation to provide an example. The simulation shows that if the bottom 20% of the simulated observations are omitted, then some points (highlighted with purple circles) change their clusters.

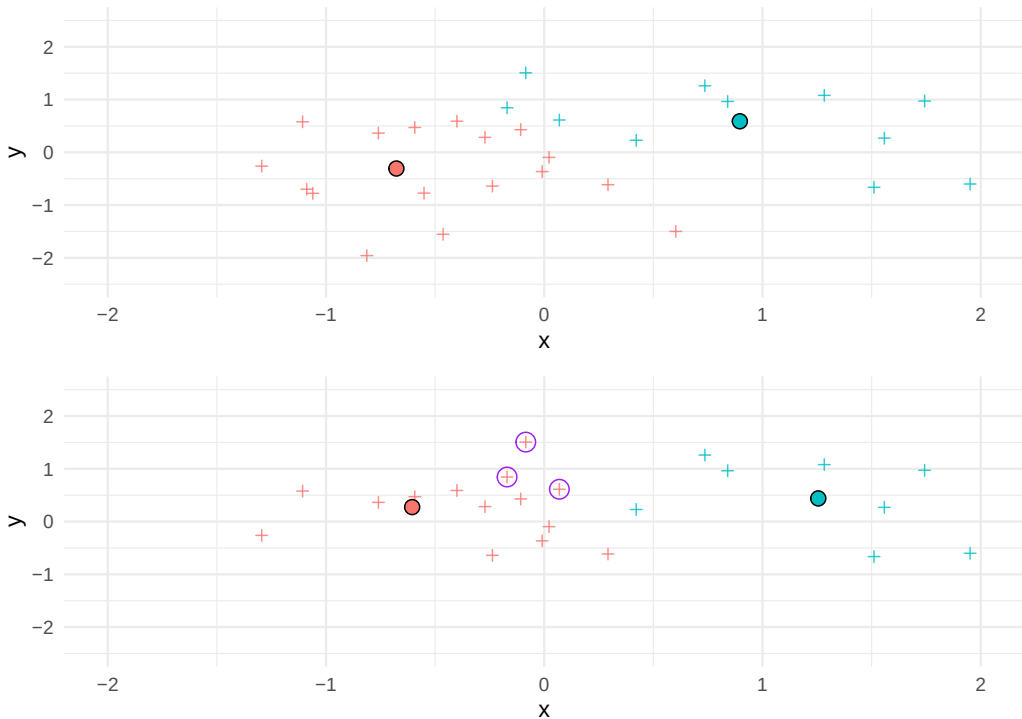


Figure 1. Simulation of K-means clustering.

The reason to use soft clustering instead of hard clustering (such as k-means clustering) is to avoid the presented "jumping" issue. Soft clustering is appropriate in situations where objects do not have a clear assignment (Ferraro 2021). In this paper, we use fuzzy k-means algorithm developed by Bezdek (2013).

The main difference between this algorithm and k-means clustering is that each observation belongs to each cluster but with varying proportion. This results in *instead of jumping we see a moderate change in proportions*.

Similarly to K-means clustering, the number of clusters is a hyperparameter what can be optimised based on silhouette criterion (Campello and Hruschka 2006).

Putting all this together, our methodological steps are as follows: (1) we standardise all the variables involved in the analysis and then (2) run the clustering for the observations for each year with cluster numbers between 2 and 10. The results are interpreted along with the number of clusters with the (3) highest average silhouette coefficient.

4. Results

4.1 Results from clustering

We began our analysis with deciding on the number of clusters to work with based on the patterns in our data. As shown in Figure 2, there is no optimal solution due to the time scale of our data. Therefore, we decided to rely on the individual optimums of each year to determine the optimal number of clusters. Until 2010 there seems to be no doubt that three is the optimal number, but in the following years there is no clear optimum. Thus, we decided to work with three clusters.

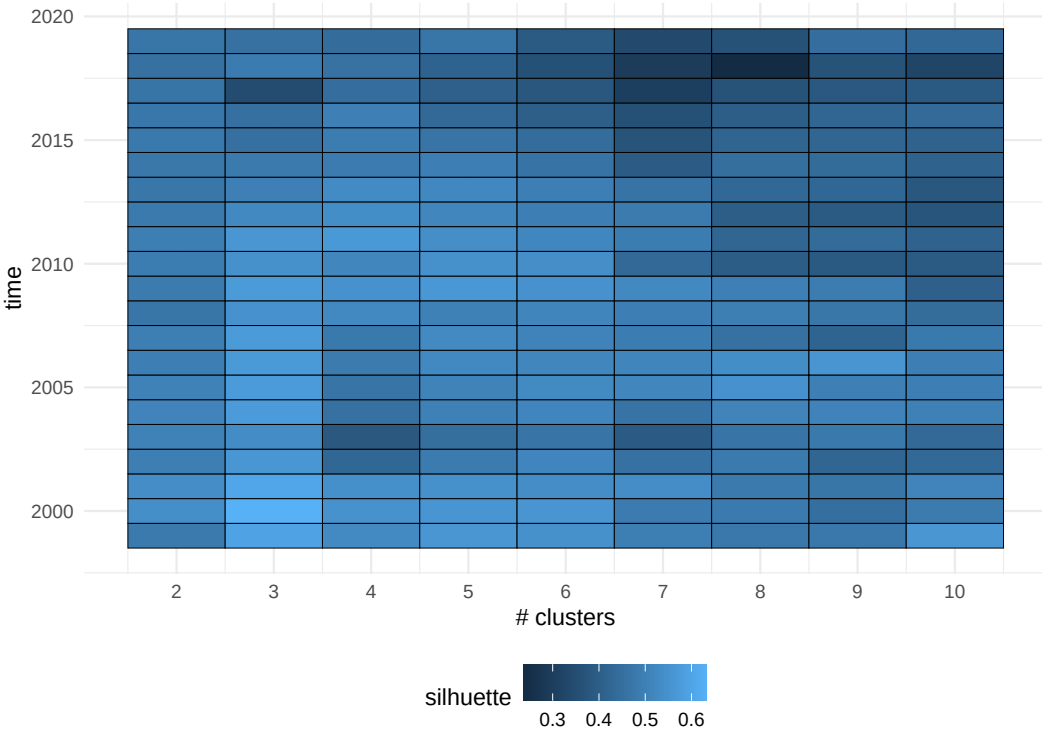


Figure 2. The optimal number of demographic clusters based on the silhouette over time (1990-2020).

We ran the K-means clustering process multiple times, which made our results robust. In this way, only a few regions changed clusters over time, and it was easier to draw conclusions from our results. The classification from the first () and last () years examined is shown in Figure 3.

Our results show that the usual classification of European regions also applies in the case of demographic characteristics. A possible way to label these clusters is *Western EU* (blue), *Central EU* (green), and *Eastern EU* (red), which accurately depicts the way Europe is usually divided into parts.

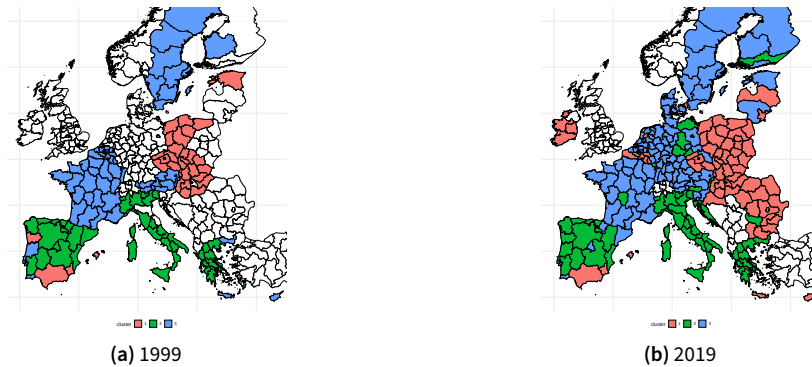


Figure 3. Classification of European NUTS-2 region in three clusters based on demographic characteristics in 1999 and 2019.

4.2 Socio-economic correlations and convergence

After classifying the NUTS-2 regions into three clusters, we examined if the clusters also depict fault lines based on socio-economic characteristics. We chose four variables to describe the socio-economic situation of a region. The first is the income of households, which shows the financial possibilities of the average citizen. The second is the income quintile share ratio of the fifth and first quintiles; this shows the concentration of incomes. The number of research and development personnel gives us insight into the state of scientific life, which can be a measure of how developed a region is. The last variable chosen, the proportion of students in the 17-year age group, gives us insight about how general secondary education is in that region.

The results of our analysis are shown in Figure 4. The results show that there is a significant convergence between the clusters in terms of socio-economic characteristics. The exception is the income of households, where the Central EU cluster is significantly left behind the Western EU, getting closer to the Eastern EU. Although numerically the difference is similar to what was 20 years ago, in terms of proportions the convergence is inevitable. This is especially important in case of those countries who joined the EU after 2000, because according to these results Eastern EU-regions are provided the opportunity to get closer to or even reach the level of the core of the EU in terms of socio-economic characteristics. However, it is important to note, that this convergence is not present in case of demographic characteristics, the reason being the clustering itself, as those who converged to other clusters changed clusters over time.

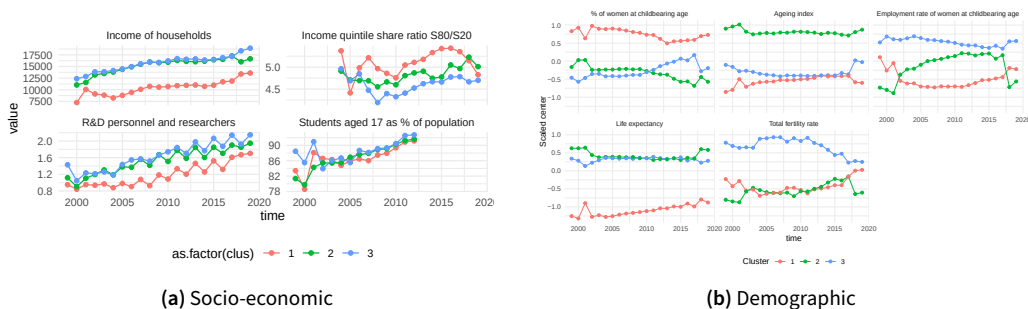


Figure 4. The convergence in terms of socio-economic and demographic characteristics of the demographic clusters over time

5. Conclusion

In this paper we performed a clustering of the European NUTS-2 regions based on demographic characteristics. These clusters seemed to match the socio-economic characteristics of these regions as well. The soft clustering method provided a robust solution to examine changes in clusters, which allowed us to draw conclusions in terms of convergence. We found, that given the same clusters over time, socio-economic characteristics of these clusters tend to converge, which is an important result for those who joined the EU later and who are less developed socio-economically.

References

- Benedek, Zsófia, Veronika Kiss, and Zoltán Bakucs. 2017. A magyar háztartások élelmiszer-fogyasztásának alakulása a 2007 és 2009 közötti pénzügyi válság tükrében. *Statisztikai Szemle*.
- Bezdek, James C. 2013. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.
- Botev, Nikolai. 2012. Population ageing in central and eastern europe and its demographic and social context. *European Journal of Ageing* 9 (1): 69–79.
- Campello, Ricardo JGB, and Eduardo R Hruschka. 2006. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems* 157 (21): 2858–2875.
- D'Addio, Anna Cristina, and Marco Mira d'Ercole. 2006. Policies, institutions and fertility rates, https://doi.org/10.1787/eco_studies-v2005-art9-en.
- Espenshade, Thomas J, Juan Carlos Guzman, and Charles F Westoff. 2003. The surprising global variation in replacement fertility. *Population Research and Policy Review* 22 (5): 575–583.
- Eurostat. 2022. *Background - nuts - nomenclature of territorial units for statistics*. <https://ec.europa.eu/eurostat/web/nuts/background>.
- Ferraro, Maria Brigida. 2021. Fuzzy k-means: history and applications. *Econometrics and Statistics*, issn: 2452-3062. <https://doi.org/https://doi.org/10.1016/j.ecosta.2021.11.008>. <https://www.sciencedirect.com/science/article/pii/S2452306221001398>.
- Günther, Isabel, and Kenneth Harttgen. 2016. Desired fertility and number of children born across time and space. *Demography* 53 (1): 55–83.
- Heinrich, Georges. 1997. Pension reforms in central and eastern europe: yet another transition...?
- KSH. 2022. *Télsz termékenységi arányszám (1990-2016)*. https://www.ksh.hu/docs/hun/eurostat_tablak/tabl/tsdde220.html?lang=hu.
- Lennert, József. 2019. A magyar vidék demográfiai jövőképe 2051-ig, különös tekintettel a klímaváltozás szerepére a belső vándormozgalom alakításában. *Területi Statisztika* 59 (05): 498–525.
- Lesthaeghe, Ron, and Paul Willems. 1999. Is low fertility a temporary phenomenon in the european union? *Population and Development Review* 25 (2): 211–228. issn: 00987921, 17284457. <http://www.jstor.org/stable/172423>.
- Likas, Aristidis, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36 (2): 451–461.
- Neyer, Gerda, Arianna Caporali, and Nora S Gassen. 2017. Eu-policies and fertility: the emergence of fertility-related family policies at the supra-national level. *Families and societies. Working paper series*, 79.
- Sobotka, Tomáš. 2004. *Postponement of childbearing and low fertility in europe*. Dutch University Press Amsterdam.
- United Nations. 2019. *World population prospects - population division*. <https://population.un.org/wpp/>.
- Westoff, Charles F. 2010. *Dhs comparative reports 25 - dhs program*. <https://dhsprogram.com/pubs/pdf/CR25/CR25.pdf>.
- World Bank. 2022. *World development indicators*. <https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.TFRT.IN&country=>.