Functional analysis in R: A new approach to sustainability indicators in the context of regional development

Antonín Hořčica[1]

Abstract: Digitalisation, together with new possibilities for data analysis, is currently gaining importance for regional development. This study focusses on the use of advanced statistical methods in R software, such as generalised linear models (GLM) and envelope methods from the GET package, to analyse time series of sustainability indicators in EU regions at the NUTS level. Analysis of selected sustainability indicators, such as the Human Development Index in its variants from the United Nations and the EU, compared to GDP shows that these methods enable effective identification of key trends and differences between groups of regions, while also identifying regions with similar trajectories. They also allow for advanced visualisation of the results, illustrating the effectiveness of EU cohesion policies at the regional level and their impact on reducing regional disparities. The analysis shows that functional analysis tools in R can be used for further research on sustainability in EU regions.

Keywords: functional analysis, statistics in R, sustainability, sustainability indicators, GDP, regional development, NUTS, GET.
JEL Classification: C14, Q01, R11, R58

1 Introduction

In the modern digital era, new data analysis capabilities are gaining prominence for regional development. This article introduces the use of R statistics for evaluating sustainability indicators, which is becoming increasingly important to assess the implementation status of sustainability strategies at both national and regional levels.

The concept of sustainable development, created in response to addressing environmental and social issues in the second half of the twentieth century, states that development should *"meet the needs of the present without compromising the ability of future generations to meet their own needs"* (WCED, 1987, p. 41). Sustainability, developed further at UN conferences (UNCED, 1992; United Nations, 2002; United Nations, 2012), has become a global discourse, especially after the adoption of the 2030 Agenda and 17 Sustainable Development Goals (SDGs) in 2015 (United Nations, 2015).

With the development of the concept of sustainability, indicators are used to verify the achievement of the set goals. The United Nations developed a set of Sustainable Development Indicators (SDGs) that assess progress at the country level. Gross Domestic Product (GDP), although still used as a universal indicator of economic success and prosperity, has proven inadequate for measuring sustainable development (Costanza, 2014). The most important alternative indicators include HDI (Human Development Index), ISEW (Sustainable Economic Well-Being), and GPI (Genuine Progress Index). In the EU, a regional variant of the HDI, the Regional Human Development Index (RHDI), has been developed for which data are available for NUTS2 regions in EUROSTAT databases (Hardeman & Dijkstra, 2014).

In European Union practise, sustainability manifests itself through models such as the green economy, the bioeconomy, and the circular economy (D'Amato & Korhonen, 2021). These models are integrated into the current European-wide plan known as the Green Deal for Europe (2019). The EU has developed its own tools to implement its strategies and policies, including the NUTS (Nomenclature of Units for Territorial Statistics) nomenclature managed by EUROSTAT, which allows monitoring of progress on sustainable development at the EU regional level.

This paper will present the statistical methods used in the analysis of regional development. Examples of time series data processing of the HDI, its regional variant RHDI, and GDP for the EU countries and their NUTS2 regions will be used to present the possibilities of functional analysis. The statistical methods provided by the open-source software tool R are used. The analysis presented in this paper demonstrates the potential of R to assess sustainability in regions.

This work aims to find an answer to the question of how advanced statistical methods such as generalised linear models (GLMs) and envelope methods can effectively enhance the analysis and interpretation of time-series data of sustainability indicators in the context of regional development. In this regard, a working hypothesis can be formulated

[1] University of South Bohemia in České Budějovice, Faculty of Economics, Department of Regional Management and Law, Studentská 13, 370 05 České Budějovice, Czech Republic, ahorcica@ef.jcu.cz

in the sense that these advanced statistical methods provide better opportunities for the analysis of time series of sustainability indicators at the regional level.

2 Methods

This chapter will begin with descriptive statistical methods for data overview and trend identification. It will then dive into advanced statistical techniques, including generalised linear models (GLMs) and envelope methods, to better understand sustainability indicator dynamics and improve data interpretation. Finally, the use of statistical software R will be described.

2.1 Descriptive statistical methods

In statistical analysis, commonly used methods allow deeper insight into the data examined, providing a foundation for informed decisions and strategies. Basic descriptive statistics serve as the initial step in data analysis, enabling researchers to quickly comprehend the fundamental characteristics of a dataset.

The key characteristics of descriptive statistics include measures of central tendency, such as mean, median, and modus, along with measures of variability, such as variance, standard deviation, and range. Position measures, such as quartiles and percentiles, are also vital. Visualisation tools used in descriptive statistics include histograms, graphs, and box plots, which display data ranges and their quartiles, allowing rapid identification of key data characteristics.

Higher levels of analysis involve regression analysis and hypothesis testing. Although descriptive statistics are essential for understanding the basic properties of datasets, advanced statistical methods are necessary for a deeper understanding of extensive data sets, such as time series of sustainability indicators at the regional level.

2.2 Generalised linear models and envelope methods

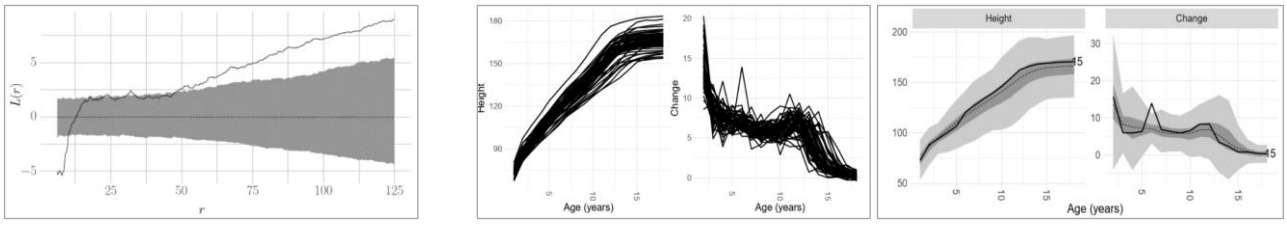
Time-series analysis is used to monitor the evolution of sustainability indicators at the regional level. These time series show linear or nonlinear functions over time. Advanced approaches using functional statistical methods and generalised linear models (GLMs) are used to process the time series, focussing on functional analysis and data structures.

Global envelope methods are a technique used in conjunction with GLM to test hypotheses about the data structure. These methods create an "*envelope*" around expected data values, and subsequently examine whether the actual data fall within this envelope. If the data falls outside the envelope, it may indicate that the model is not suitable for the data and the null hypothesis about the data and model agreement should be rejected. Figure 1a shows an example of an envelope test of the null hypothesis for a Poisson process, where the grey area defines a 95% global envelope based on simulations.

Myllymäki and Mrkvička (2020) provide several examples of utilising global envelopes from the GET package in R. Figure 1b shows one such example, highlighting the difference between visualisation using standard methods (left) and the envelope method (right). This study used the annual height of girls, evaluating their growth trajectories from 1 to 18 years, and compared annual heights and year-over-year changes, underscoring the advantage of the envelope method.

Figure 1 Examples of envelope method application and comparison with standard processing

a) Envelope test b) Standard method Envelope Method



Source a) Mrkvička, 2017 Source b) Myllymäki a Mrkvička, 2020

The use of global envelopes in functional data analysis and generalised linear models serves as a valuable tool for identifying the central region ("*normal values*"), comparing data with a reference distribution, performing tests between groups, and creating confidence intervals. Functional cluster analysis, proposed as another GLM method, is discussed in the work of Dai et al. (2022). This method can be used to identify groups of countries and regions with similar economic trends with respect to sustainability indicators, helping to predict economic developments and planning strategies.

2.3 Statistics in R

The open-source software R is widely recognised for its ability to perform a wide range of statistical methods, including descriptive statistics. For this work, methods related to functional statistics were used. R provides a number of tools and methods for performing these analyses and is a popular tool for scientists and analysts around the world to work

with data. According to the website of *R: The R Project for Statistical Computing*, R is a statistical software that was developed in 1995 at Bell Laboratories under the direction of John Chambers and his team.

RStudio serves as the interface that allows users to work with R. It is user-friendly and contains tools to work efficiently with R, including a code editor, a console for direct access from R to the file manager, and visualisation tools. RStudio can be downloaded from the Posit website. Source packages for various applications and tasks can be found at *CRAN (n.d.), a* detailed description is given in *'The R Book'* by Crawley (2013).

The R software tool provides basic descriptive statistics and data visualisation tools. Basic functions include:

- Basic functions:**mean()**, **median()**, **sd(), var()**, **min()**, **max()**, **quantile()**, **summary()**, **IQR()** interquartile range for the identification of outlier values, and for determining order, the functions **sort()** or **order**.
- Visualisation of data using graphs: **plot()**, **barplot()**, **boxplot(), hist(),**
- Graphic packages: **graphics**, **ggplot2, lattice, plotly** and tools for 3D visualisation and geospatial analysis,
- The **eurostat** package serves for working with data from EUROSTAT, see Lahti et al. (2017) or the tutorials (Tutorial for the Eurostat R Package, n.d.; Stavrakoudis, n.d.).

In addition, a number of specialised packages can be used for various analyses. One of them is the GET package, which contains tools to analyse functions according to GLM. A description of this package can be found in Myllymäki & Mrkvička (2020) and Mrkvička (2017). The concept of envelope methods can be used to detect outliers and anomalies in time series and to test hypotheses about the fit of a statistical model. Another method is the Functional Clustering Method (FCM), described in Dai et al. (2022), which is used to identify groups of similar functions.

3 Research results

Initially, basic statistical calculations were performed using descriptive statistics and basic visualisation functions. Subsequently, functional analysis methods were applied. The basic statistics were derived from data for the EU27 member states for the UN HDI, obtained from the UNDP database. Data for GDP per capita and the calculated RHDI indicator from the Eurostat database were also used for functional analysis.

3.1. Basic statistics and visualisation in R

The **summary** function provides overall statistics for the entire data set (minimum, maximum, 1st quartile, median, mean, and 3rd quartile) - see Table 1. The **aggregate** (data, summary) function yields basic statistics for individual years, with only the first five displayed here, achievable with the head command - see Table 2.

Table 1 Example of summary function output

Indicator	Min.	1st Qu.	Median	Mean	3rd Qu	Max.
UN HDI	0.6790	0.7993	0.8515	0.8414	0.8920	0.9480

Source: Own processing in R from UNDP data from UNDP data (HDI values are dimensionless).

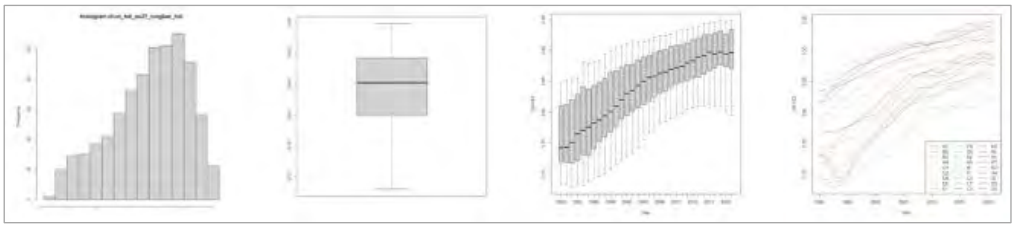
Table 2 Example of the output of the summary function

Indicatorr	Year	Min.	1st Qu.	Median	Mean	3rd Qu	Max.
UN HDI:	1990	0.6840	0.7200	0.7420	0.7593	0.8100	0.8470
	1991	0.6840	0.7190	0.7430	0.7615	0.8130	0.8520
	1992	0.6790	0.7180	0.7510	0.7640	0.8200	0.8530
	1993	0.6800	0.7220	0.7650	0.7703	0.8200	0.8550
	1994	0.6820	0.7310	0.7710	0.7780	0.8410	0.8810

Source: Own processing in R from data from UNDP and EUROSTAT (HDI values are dimensionless).

Furthermore, graphical processing of the UN HDI was carried out using the **hist** function to construct a histogram for the whole period and a boxplot for the whole period and for each year and using the **plot** and **boxplot** functions to plot the values over the monitoring period for each EU country. The graphs are shown in Figure 2.

Figure 2 Graphical visualisation in R for the UN HDI for EU countries



Source: Own processing in R from data from UNDP and EUROSTAT

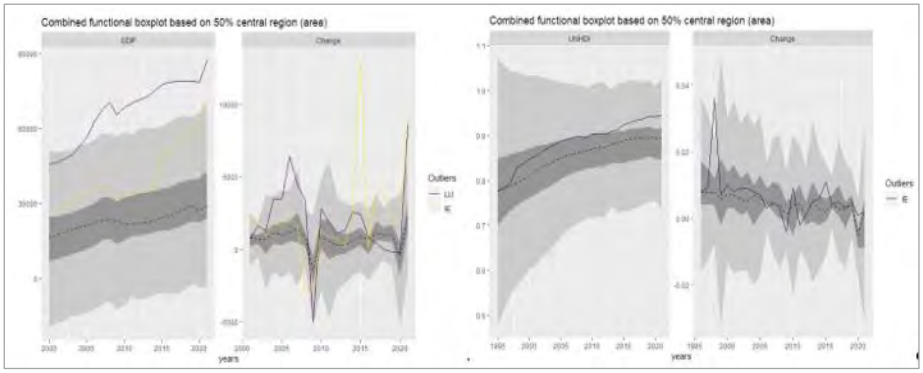
3.2. Functional statistics in R

To test the feasibility of using functional statistical methods, the methodologies described in Mrkvička (2017) and Myllymäki and Mrkvička (2020) were used.

3.2.1 Functional boxplot for aggregate indicator values

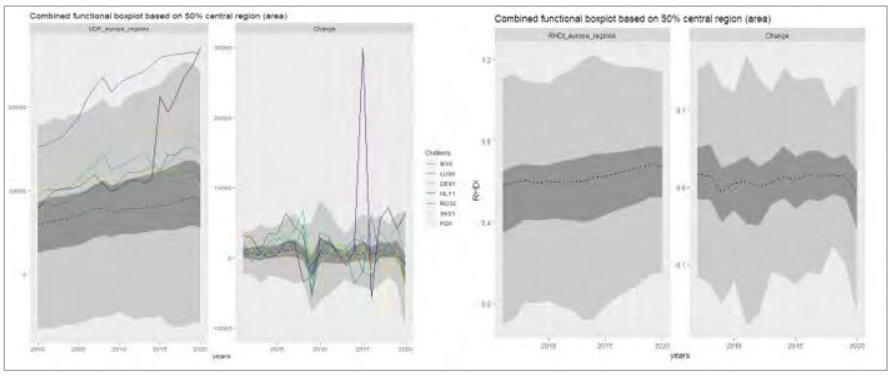
Figures 3 and 4 illustrate the application of the envelope method to analyse time series of values and annual changes in GDP, UN HDI, and RHDI using a functional boxplot (via the **create_curve_set** function). The median value (dashed line), a dark grey area representing 50%, a light grey envelope for 'normal' values, and coloured outliers are shown.

Figure 3 Functional boxplot for GDP per capita and UN HDI values for EU countries



Source: Own processing in R from data from UNDP and EUROSTAT

Figure 4 Functional boxplot for GDP/capita and UN HDI values for NUTS2 regions

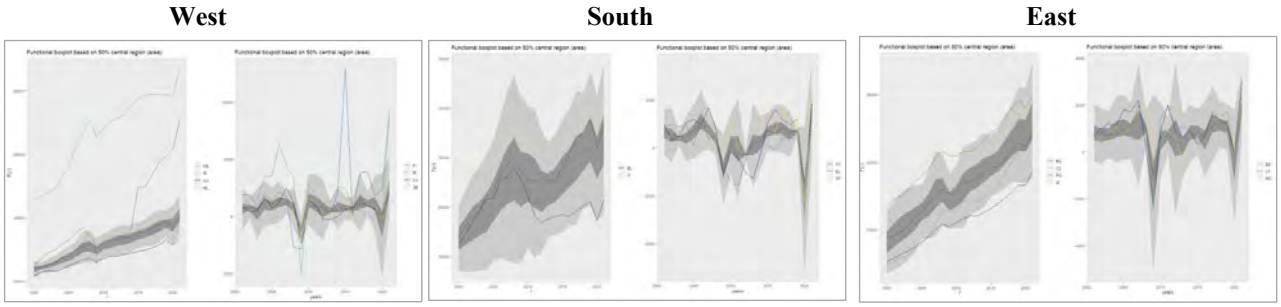


Source: Own processing in R from data from UNDP and EUROSTAT

From the graphs in Figure 4 can be deduced:

- different trends for different indicators, but also the level of hierarchical ranking of regions,
- to identify outlying trends for countries or regions,
- the shape of the envelope suggests a tendency towards convergence, i.e. whether there is a reduction of disparities;
- outliers can be found in the visualisation for countries and for regions,
- for year-on-year change, we can see fluctuations around 2008 (financial crisis) and after 2019 (covid pandemic).

Figure 5 Functional boxplot of GDP per capita by geographical grouping



Source: Own processing in R from data of EUROSTAT

3.2.2 Functional boxplot for geographical groups of countries

Geographical groups of EU countries were created; for simplicity, the Baltic countries were grouped in the east, and the Scandinavian countries in the west. The analysis was carried out as a demonstration of the possibility of functional analysis for the GDP per capita indicator. The geographical breakdown of the EU countries is as follows:

- West (+ North): Belgium (BE), France (FR), Ireland (IE), Luxembourg (LU), Denmark (DK), Finland (FI), Sweden (SE), Austria (AT), Germany (DE), Netherlands (NL),
- South: Cyprus (CY), Malta (MT), Greece (EL), Italy (IT), Portugal (PT), Spain (ES),
- East: Bulgaria (BG), Czech Republic (CZ), Hungary (HU), Poland (PL), Romania (RO), Slovenia (SI), Croatia (HR), Lithuania (LT), Latvia (LV), Estonia (EE).

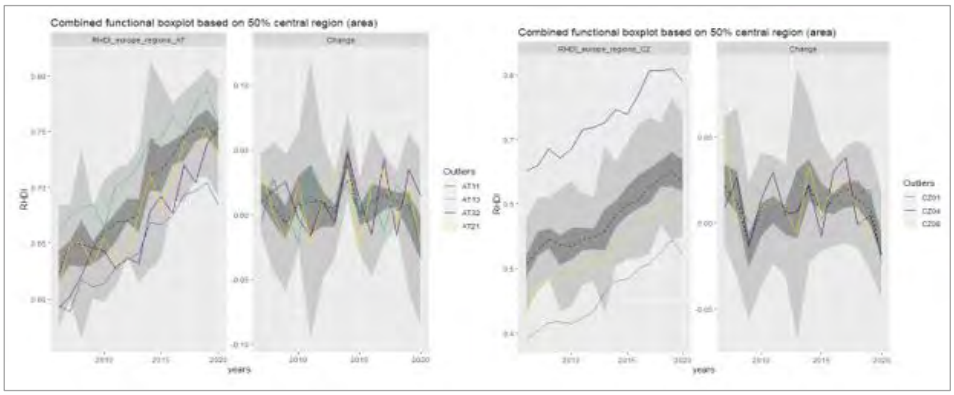
The results are shown in Figure 9, they are analogous to those of the previous case. From the graphs, it can be deduced:

- within each geographic group of countries, countries as outliers can be identified,
- the median curves for all groups are growing, but their character is different; the west showing a steady rise, the south showing a distinct peak, and the east showing a steep increase with a smaller peak than the south,
- the shape of the envelopes varies quite a lot, it is not possible to clearly conclude on convergence or divergence in each group, the shapes of the envelopes change quite a lot during the observed period,
- there are also significant differences between year-on-year changes, there are clear fluctuations around 2008 and 2019 as before, but not the same for each group of countries.

3.2.3 Functional boxplot for NUTS2 regions within a country

Figure 6 shows the time series for the RHDI indicator in a NUTS 2 region within a country. For the demonstration, two neighbouring countries, namely Austria and the Czech Republic, were chosen.

Figure 6 Functional boxplot for RHDI indicator values for NUTS 2 regions of Austria and the Czech Republic



Source: Own processing in R from data from UNDP and EUROSTAT

It is possible to complement the previously identified deductions to derive more region-specific findings. For example,

- regions AT13 Wien and CZ01 Praha show above-average deviations in the development of RHDI, and
- regions AT11 Burgenland, AT21 Kärnten, AT22 Steiermark in Austria, and CZ04 Severozápad and CZ08 Moravskoslezsko in the Czech Republic deviate below the average,
- From the graphs, it is evident, except for CZ01 Praha, that Austrian regions achieve a better RHDI level than Czech regions, but the latter do not manage to significantly reduce the gap,
- no clear convergence or divergence can be inferred for both countries,

- From the year-on-year changes can be deduced that the impact of 2008 was more significant in the Czech Republic than in Austria, the impact of COVID was similar, other significant fluctuations require more detailed analysis.

4 Conclusions

In this paper, the descriptive and functional statistics tools in R have been used to analyse regional development. Time series of the UN HDI and RHDI indicators, supplemented with a GDP indicator, were used for both EU countries and their regions at the NUTS2 level. Statistics in R, including methods of descriptive statistics and generalised linear models with the innovative envelope method, allow in-depth data analysis and visualisation, identifying key trends and data characteristics for EU countries and regions.

This study focused on answering the basic question. How can advanced statistical methods, such as generalised linear models (GLM) and envelope methods, improve the analysis and interpretation of sustainability indicators' time series in the context of regional development? Based on the results, it could be stated that the use of these methods can significantly contribute to a deeper and more accurate understanding of sustainability dynamics at the regional level. In particular, the innovative envelope method from the GET package provides intuitive and easily interpretable outputs, which are valuable for broader discussions and communication of research results. This method allows application for countries, groups of countries, or NUTS regions, revealing specificities and general trends crucial to understanding interregional disparities and development within countries.

At the same time, it can be stated that our working hypothesis that advanced statistical methods provide better options for analysing time series of sustainability indicators was supported. For example, analysing the RHDI indicator in Austria and the Czech Republic reveals specifics for each country and simultaneously identifies general trends, which are key to understanding regional disparities and developments within countries.

Although this paper does not encompass the entire scope of sustainability indicators' time series analysis, it opens new possibilities and provides a scope for further research. Future research directions might include using cluster analysis and functional ANOVA to assess statistical differences between regions and employing advanced visualisation techniques in R, such as 3D graphs and geospatial analysis, to gain more insight into the spatial aspects of research and visualisation of regional and national trends and differences in sustainability.

Acknowledgement:

I would like to express my thanks to doc. Ing. Eva Cudlínová, CSc. and prof. RNDr. Tomáš Mrkvička, Ph.D., for their advice and consultations related to the topic of this work. This work was supported by the GA JU 129/2022/S.

References

Costanza, R., Hart, M., Kubiszewski, I., & Talberth, J. (2014). A Short History of GDP Moving Towards Better Measures of Human Well-being. *Solution*, 5(1), 91-97. [Online]. Avalaible at https://www.thesolutionsjournal.com/article/a-short-history-of-gdp-moving-towards-better-measures-of-human-well-being/.

CRAN. (n.d.). CRAN. [Online]. Avalaible at https://cran.r-project.org/.

D'Amato, D., & Korhonen, J. (2021). Integrating the green economy,, circular economy and bioeconomy in a strategic sustainability framework. *Ecological Economics*, 188, 107143. [Online]. DOI 10.1016/j.ecolecon.2021.107143.

Dai, W., Athanasiadis, S., & Mrkvička, T. (2022). A New Functional Clustering Method with Combined Dissimilarity Sources and Graphical Interpretation. In R. López-Ruiz (Ed.), *Computational Statistics and Applications*. IntechOpen. [Online]. DOI 10.5772/intechopen.100124.

European Commission. (2019). The European Green Deal. [online]. [Online]. Avalaible at https://ec.europa.eu/info/sites/default/files/european-green-deal-communication_en.pdf.

EUROSTAT. European Statistical Office. (n.d.). [Online]. Avalaible at https://ec.europa.eu/eurostat.

Eurostat. NUTS. Nomenclature of Territorial Units for Statistics. [Online]. Avalaible at https://ec.europa.eu/eurostat/web/nuts/background.

Hardeman, S., & Dijkstra, L. (2014). The EU Regional Human Development Index. Luxembourg: Publications Office of the European Union. [Online]. Avalaible at https://publications.jrc.ec.europa.eu/repository/bitstream/JRC90538/online%20version%20a4.pdf.

Lahti, L., Huovari, J., Kainu, M., & Biecek, P. (2017). Retrieval and Analysis of Eurostat Open Data with the eurostat Package. *The R Journal*, 9(1), 385. [Online]. DOI 10.32614/RJ-2017-019.

Mrkvička, T. (2017). Globální obálkové testy aneb jak otestovat vhodnost statistického modelu na základě funkcionální charakteristiky. *Pokroky matematiky, fyziky a astronomie,* 62(1), 17–23.

Myllymäki, M., & Mrkvička, T. (2020). GET: Global envelopes in R (arXiv:1911.06583). arXiv. [Online]. Avalaible at http://arxiv.org/abs/1911.06583.

Posit. (n.d.). Posit. [Online]. Avalaible at https://www.posit.co/.

R: The R Project for Statistical Computing. (n.d.). [Online]. Avalaible at https://www.r-project.org/.

Redclift, M. R., & Springett, D. (2015). *Routledge international handbook of sustainable development*. London: Routledge, Taylor & Francis Group.

Stavrakoudis, A. (n.d.). Using Eurostat with R. [Online]. Avalaible at http://stavrakoudis.econ.uoi.gr/r-eurostat/index.html.

Tutorial for the eurostat R package. (n.d.). [Online]. Avalaible at https://ropengov.github.io/eurostat/articles/eurostat_tutorial.html.

UNCED. (1992). United Nations Conference on Environment and Development: Agenda 21. Programme of Action for Sustainable Development. New York: United Nations. [Online]. Avalaible at https://sustainabledevelopment.un.org/content/documents/Agenda21.pdf.

UNDP. United Nations Development Programme. (2019). Human Development Reports. Technical Notes [Online]. Avalaible at http://hdr.undp.org/en/content/human-development-index-hdi.

United Nations. (2002). World Summit on Sustainable Development (WSSD): Johannesburg Declaration on Sustainable Development. [Online]. Avalaible at http://www.un-documents.net/jburgdec.htm.

United Nations. (2012). Report of the United Nations Conference on Sustainable Development (UNCSD): Rio de Janeiro, Brazil, 20-22 June 2012. New York: United Nations [Online]. Avalaible at https://www.un.org/ga/search/view_doc.asp?symbol=A/CONF.216/16&Lang=E.

United Nations. (2015b). Transforming our World: The 2030 Agenda for Sustainable Development. New York: United Nations [Online]. Avalaible at ttps://sustainabledevelopment.un.org/post2015/transformingourworld/publication.

WCED. (1987). Report of the World Commission on Environment and Development: Our Common Future. [Online]. Avalaible at http://www.un-documents.net/our-common-future.pdf.