Master's Thesis – Master Sustainable Business and Innovation

JULY 15, 2022

Exploring regional patterns of scientific knowledge production on the Sustainable Development Goals

RUESEN, C. (CEDER)
5793696
c.ruesen@students.uu.nl

Geosciences, Utrecht University

Supervisor: Dr. Gaston Heimeriks
Second reader: Dr. Jarno Hoekman

Universiteit Utrecht

CWTS
Meaningful metrics

Abstract

To address the pressing issues human overpopulation has created on earth, such as climate change and poverty, the United Nations (UN) introduced the Sustainable Development Goals (SDGs) as a global agenda to guide the world onto a more sustainable trajectory. To achieve the goals, the UN and scholars urge for local action, support, and collaboration of scientific research institutes. Additionally, embedding this structural transformation locally, requires new knowledge to address region specific challenges. Important in this process is to gain insights into the current knowledge base of regions, to find opportunities for new knowledge development. This study therefore aims to explain differences among European regions in complex knowledge production on SDG related research, specifically looking at the knowledge complexity of a region, it's scientific relatedness to the SDGs, and several regional characteristics based on the SDG indicators. These concepts are drawn from literature on Evolutionary Economic Geography that argues that complex knowledge production is influenced by mechanisms of path- and place dependency. Using the CWTS wos_2113 database, scientific publications are retrieved that represent a region's knowledge base. Data from the STRINGS project is used to identify SDG related publications. Findings show that North-Western Europe produces the most complex knowledge and has the highest relatedness to the SDGs. Following from this, four regression models are estimated, to find relationships between the variables. These models include data from before the introduction of the SDGs, 2010-2014, hereafter, 2015-2020, and with and without the inclusion of the regional characteristic variables, as these are only selected for a limited number of SDGs. The findings show that the SDGs are not equally well explained by the different variables, suggesting that there is not one model that fits all the SDGs. Nevertheless, for most SDGs a positive relationship is discovered between the knowledge complexity of regions and their scientific relatedness to the SDGs. This indicates that path- and place-dependent mechanisms also apply to the SDGs and proximity advantages should be considered. In addition, for the regional characteristic variables no general conclusion is drawn, as the indicators vary per SDG. However, surprising to see is that several SDGs show a negative relationship with the SDG research share, indicating that there is a misalignment between research priorities and societal needs. This study thus provides several promising insights and recommendations to expand the knowledge base of regions through research and collaboration, thereby aiming to accelerate the sustainability transition.

Table of Content

1. Introduction

Society is increasingly becoming aware of the pressure it has created on the earth (Sachs, 2012). Climate change, poverty, and the financial crisis are just a few of the phenomena that result from the exponential growth of the human population (Uniyal et al., 2020). To address these pressing issues, the United Nations (UN) introduced the Sustainable Development Goals (SDGs) in 2015 as a global agenda to guide the world onto a sustainable trajectory and embracing the triple bottom line, people, planet, and prosperity (Sachs, 2012). Overall, the SDGs consist of 17 goals and 169 targets that are integrated, indivisible, and incorporate the three pillars of sustainable development: the Economic, Social, and Environmental pillar, as shown in Figure 1 (UN General Assembly, 2015). Currently, the SDGs are adopted by all UN Member States and although progress is made on the different goals, overall, the speed and scale that are required to deliver the goals by 2030 is not yet sufficient. Therefore, the UN Secretary-General called for amongst others local action by embedding the sustainability transition more in policies, institutions, and regulatory frameworks (United Nations, n.d.). In addition, the UN acknowledges the requirement of support and resources for scientific institutions to advance the existing knowledge base (United Nations, 2019a). New knowledge development is thus crucial to this transformation pathway and regions can play an important role in this transition.

Figure 1:
Sustainable Development Goals, clustered into the three pillars of sustainability

Note: From *A Novel ICT Framework for Sustainable Development Goals,* by Kostoska and Kocarev, 2019, p. 2.

Previous research has shown that knowledge is an important aspect of regional development as the accumulation of knowledge largely determines the economic prosperity of a region and is important for addressing societal challenges (Bretschger, 1999; Schikowitz, 2020). Additionally, research found that knowledge production is unevenly distributed and lagging regions often don't have the capabilities to replicate and exploit knowledge used by leading regions (Fritsch & Slavtchev, 2005; Heimeriks et al., 2019). This is especially the case with complex knowledge, which, due to its tacit nature, is hard to imitate and often spatially bounded (Pintar & Scherngell, 2021). Complex knowledge production is thus subject to path- and place-dependency, which means that each region has its own specific knowledge base that could both constrain and foster the development of new knowledge (Heimeriks & Boschma, 2014). Opportunities with regards to this knowledge base depend on the relatedness between scientific topics, as regions are more likely to develop new knowledge that is related to their existing knowledge base. In addition, if regions lack any useful related knowledge this could cause them to lock-in as diversification options are limited (Hidalgo et al., 2007; Li, 2020). This relatedness, with regards to the sustainability transition, was found to also apply for renewable energy technologies (Li et al., 2021). Consequently, there is an urgency to also understand the dynamics of complex knowledge production with regards to the SDGs.

Earlier research on complex knowledge production with regards to the SDGs stresses that knowledge production on the SDGs is different from knowledge production in other areas in several ways (Romero-Goyeneche et al., 2021); First, addressing the SDGs involves directionality as the transformative agenda requires important choices and directions for investment in research. This includes a deep analysis of what type of research and which areas will be most likely to induce large scale transformative change (Ramirez et al., 2019). Secondly, research on the SDGs requires a variety of approaches, due to the complexity of the goals. The SDGs are multidisciplinary and involve environmental, social, and technological systems. To be able to manage such complexity, knowledge, and expertise across a variety of disciplines is required to reach the goals (Arroyave et al., 2021). Finally, when it comes to integrated research and decision-making related to the SDGs, one should always consider the interactions, positive ("co-benefits) or negative ("trade-offs"), between the SDGs and how these contribute to transformative change (Nilsson et al., 2018; Schot & Kanger, 2018). Considering this, analyzing region's current knowledge bases, and identifying regions that were capable to produce this complex knowledge on the SDGs, provides the potential to map and locate promising areas for transformative research (Romero Goyeneche et al., 2022). A better understanding of the current state of the knowledge bases and identifying leading and lagging regions thus provides the opportunity to gain deeper insights into the dynamics behind the complex knowledge production and to further expand knowledge production on the SDGs.

Additionally, in order to reach the sustainability goals, set for 2030, and increase knowledge production on SDGs, scholars call for action and urge scientists from different scientific fields to share knowledge and collaborate more to fill the knowledge gaps in critical areas (McCollum et al., 2018; Messerli et al., 2019). The promotion of structural transformation requires more research on the SDGs and policies specifically designed for regions (Hidalgo et al., 2007). This is another reason why better insights are required into the current knowledge bases of regions and the adjacent possibilities they have for diversification, as this could provide decision-makers with more concrete handholds to foster knowledge production on the SDGs (Autant-Bernard et al., 2013). The research by Romero-Goyeneche et al. (2021) already provides a good starting point for identifying the state of knowledge production on SDGs, although results are only applicable to the University of Utrecht. Therefore, a good opportunity to follow-up on their activities would be to gain better insights on how different regions in Europe are performing compared to each other. Especially, empirical evidence is required to identify which factors promote or constrain the capability of regions to produce complex knowledge on SDGs, which in turn can strengthen transformative research and lead to breakthrough innovations. To address this gap in the literature, this research aims to answer the following research question:

What explains differences among European regions in complex knowledge production on SDG related research?

This research question is twofold, on the one hand it tries to identify how the non-SDG knowledge base of a region could determine the possibilities for diversification in SDG research. It thereby aims to identify whether the capabilities of a region to produce complex knowledge non-related to the SDGs also enables them to produce complex knowledge on the SDGs. This leads to the following sub questions:

What is the relation between the complexity of a region's knowledge base and its ability to produce SDG related research?

What is the relation between the relatedness of a region's knowledge base to the SDGs and its ability to produce SDG related research?

On the other hand, this research aims to deepen the knowledge on other factors than place- and path dependency mechanisms that could promote the research on SDGs. Therefore, this research will highlight several SDGs with the aim to identify whether local effects related to one of the SDGs increases the knowledge production in that SDG. This leads to the following sub question:

How do region-specific characteristics influence the complex knowledge production on SDGs?

This study will follow a quantitative approach, building upon the theory of Evolutionary Economic Geography scholars (Balland & Rigby, 2017; Boschma & Martin, 2010; Hidalgo et al., 2018; Martin & Sunley, 2006). Scientific publications from the Web of Science (WoS) are retrieved from the CWTS database to give an estimation of the current knowledge base of regions. Besides, a distinguishment will be made between non-SDG and SDG related research with data from the STRINGS project, to identify the ability of regions to produce complex knowledge regarding the SDGs. Other indicators that are suspected to either constrain or promote complex knowledge development within this research are gathered from the Eurostat database and the Global SDG Indicators Database. First a descriptive analysis will be performed, to map and visualize the current state of the knowledge base of European regions. Hereafter multiple regression analyses are performed in order to estimate the relationship between the SDG research share and knowledge complexity, scientific relatedness, and regional characteristics.

Through exploring the factors influencing the complex knowledge production on SDGs this study aims to contribute to the existing body of knowledge within the Evolutionary Economics Geography domain. By extending the concepts towards opportunities and constraints imposed by the current scientific and technological trajectories of the SDGs, this study provides a more holistic and explicit understanding of regional knowledge development. Besides, this study aims to contribute to an increased understanding of the current state of the European knowledge base with regards to the SDGs. By mapping the knowledge production per region, policy makers are enabled to identify lagging and leading regions. This will allow them to identify knowledge gaps and give them the opportunity to provide incentives and stimulate further knowledge production and diversification in specific areas. Ultimately, this research attempts to provide other scholars with the opportunity to apply the used methodology and develop further insights that allow for comparison. From a social perspective, this study attempts to increase knowledge production on the SDGs within the most critical areas. Thereby contributing to reaching the goals of the 2030 Agenda to transform our world, with the ultimate goal of solving poverty and creating a healthy and secure planet (UN General Assembly, 2015).

2. Theoretical framework

This research builds upon the theory of Evolutionary Economic Geography. The aim of this research field is to understand how geography can determine and shape the economic landscape over time. Innovation and knowledge are central in this process, where internal knowledge development is seen as an important factor of the adaptive and transformative characteristics of economic evolution (Boschma & Martin, 2010). Literature that is central to this theory is strongly centered around knowledge complexity and scientific relatedness (Balland et al., 2019; Balland & Rigby, 2017). In the following sections these concepts are further elaborated upon. First, the importance of SDG related research is discussed, followed by a conceptualization of knowledge complexity. Second the principles of scientific relatedness are explained and their implications on SDG related research. Finally, literature on the influence of region-specific characteristics on regional knowledge production are discussed and the transformative lens is introduced.

2.1 SDG related research

Research, innovation, and raising awareness through substantial private and public investment, are important mechanisms to achieve the SDGs (Salvia et al., 2019). The goals can function as an opportunity to foster research on sustainability as humans are currently still exceeding the planetary boundaries and have not stopped depleting natural resources (Leal Filho et al., 2018). Scholars could contribute to the goals by translating the SDGs in local and feasible goals. Previous research shows that Europe is mainly concerned with SDGs related to education, industry, innovation and infrastructure, and sustainable consumption and production (Salvia et al., 2019). Advancing this knowledge base and diversifying to other SDGs is of main importance to achieve the SDGs. Further knowledge development enables continued and improved research, possibly leading to breakthrough innovations (Chandra & Dong, 2018), and can provide a foundation for policy measures that target societal change. SDG related research can thus result in or at least contribute to the achievement of the 2030 Agenda.

2.2 Knowledge Complexity

Knowledge production is dependent both on knowledge complexity and the ability of a country or region to accumulate knowledge. This is also a reason why knowledge production is distributed unevenly over regions (Heimeriks & Balland, 2016). This phenomenon is often attributed to the tacitness of knowledge, which is knowledge that is highly personal and often only learned by experience, such as expertise (Mitchell & Leach, 2019). Tacit knowledge is thus hard to communicate and cannot be simply written down, therefore some degree of geographical proximity is required for the effective communication of tacit knowledge. Tacit knowledge can also be described as complex knowledge as it is difficult to imitate and thus on the one side can provide organizations with a competitive advantage, however, on the other side, it can also constrain research institutions to use this knowledge and accumulate it (Kim & Anand, 2018). SDGs are especially interesting in the light of complexity as they are wicked problems. Wicked problems are known to be complex as they often involve a multitude of stakeholders, cross organizational boundaries, and no perfect solution exists (Kolk, 2013). Consequently, the inter- and transdisciplinarity of the goals and the interrelatedness between them makes the goals incredibly complex (Arroyave et al., 2021; Romero-Goyeneche et al., 2021).

The complexity of knowledge is determined by the difficultness for others to imitate it. This is mainly dependent on the diversity and ubiquity of the knowledge. This means that a region should possess a variety of different knowledge sources and should be capable to develop unique knowledge, in order to produce complex new knowledge (Pintar & Scherngell, 2021). Thus, high diversity and low ubiquity contribute to the complexity of knowledge production in a region. Besides, the ability of a region to produce complex knowledge and combine this with existing knowledge and capabilities into new knowledge allows them to diversify more easily into other scientific fields (Hidalgo & Hausmann, 2009). Diversification to unrelated field is, however, more important for the long-term development of countries and regions, to keep competitive and ensure economic growth (Hidalgo et al., 2007; Li, 2020). From this reasoning it follows that the ability of regions to diversify into more SDG related research, either related or unrelated, is dependent on the complexity of the knowledge base of that region. Expected is that this is critical considering the complexity of the SDGs, which leads to the following hypothesis.

H1: Regions with a more complex knowledge base are more likely to engage in SDG related research.

2.3 Scientific relatedness

Knowledge production in regions is characterized by path- and place-dependency mechanisms, as earlier mentioned (Heimeriks & Balland, 2016; Li, 2020). Consequently, regions have different knowledge assets, such as skills and expertise, public-private relationships and formal and informal institutions that support knowledge creation within regions (Lönnqvist et al., 2014). The degree and nature of path-dependency can thus be seen as locally embedded and is largely place-dependent (Martin & Sunley, 2006). The characteristics of knowledge production, thus, are distinct, but interact with each other. The cumulative and tacit nature of knowledge leads actors to build on existing knowledge they have required in the past. This process of building on existing knowledge leads eventually to a trajectory that defines the technological opportunities for further development. This path-dependent mechanism often takes place at the regional level as some degree of geographical proximity is necessary (i.e., place-dependency) for the transfer of tacit knowledge (Heimeriks & Boschma, 2014).

The ability of a region to exploit the existing knowledge base, but also to utilize and evaluate external knowledge, is determined by a region's absorptive capacity. This means that a region needs a degree of prior related knowledge to be able to assimilate and exploit new knowledge (Cohen & Levinthal, 1990). This related knowledge can also be explained as scientific relatedness that involves on the one hand scientific similarity, which is the degree to which scientific research in regions is based on the same knowledge. On the other hand, it involves scientific complementarity between regions, which is the degree to which scientific research focuses on different areas of knowledge within a larger shared knowledge area so that they complement each other (Makri et al., 2010). Regions that have a knowledge base with a certain level of similarity and complementarity to external knowledge, will thus be more capable to assimilate this new knowledge (Hidalgo et al., 2018). Besides, the existing knowledge base of a region leads to an increased likeliness to diversify in related knowledge bases and discourages knowledge production on unrelated topics (Heimeriks et al., 2019). Thus, if the scientific field of an SDG is more related to the existing scientific knowledge base of a region, the region will be more likely to diversify into this SDG. This leads to the following hypothesis:

Hypothesis 2: Regions with a knowledge base closer related to SDGs are more likely to engage in SDG related research.

2.4 Regional characteristics

The innovativeness, competitiveness, and productivity growth of knowledge-intensive industries tends to be determined by regional characteristics, such as the resource base, innovation policy, and R&D expenditure. Besides, also more cultural, and social regional characteristics are identified, such as the quality of local government and a 'thick' labor market, that enable regions to attract and integrate expertise from individuals (Reichert, 2006). Knowledge production is shown to increase in regions with a high level of social interaction as well as the ability of regions to exploit externally acquired R&D (Laursen et al., 2007). Heimeriks and Balland (2016) argue that the institutional context also influences knowledge replication and accumulation. It affects both the degree of interactive learning between actors and the ability of a region to transform activities, organizations, and institutions to accumulate and develop new knowledge (Boschma, 2004). The tacit nature of the institutional context and the cultural embeddedness of these mechanisms, make them strongly geographically bounded and difficult to imitate and copy for other regions (Rodríguez-Pose, 2013). Additionally, research has shown that market structures influence knowledge development. Liberal market economies are more likely to develop radical innovations, while coordinated markets economies are more characterized by incremental innovations (Hall & Soskice, 2001).

Furthermore, a study by Barbieri et al, (2020) shows that early-stage environmental innovations tend to come from regional knowledge bases that are diverse, but where the scientific subfields in the knowledge base are unrelated. This is because these types of innovations require a multidisciplinary approach. Considering this, knowledge on the SDGs can also be regarded as early-stage since knowledge that addresses the SDGs is still in development and multidisciplinary by nature (Arroyave et al., 2021). The possibility that knowledge production on SDGs is dependent more on unrelated variety, indicates that other factors than knowledge complexity and scientific relatedness, such as the social and cultural regional characteristics discussed earlier, could stimulate the development of knowledge production on SDGs. Besides, research has shown that developing countries are more likely to adopt the most effective technology rather than the most environmentally friendly technology (Perruchas et al., 2020). Furthermore, Soini et al. (2018) point out that academic research is often motivated by a researcher's interest in a topic. This might indicate that regions choose to develop knowledge and implement innovations that benefit them the most. For instance, a region that suffers from massive insect deaths, might be more inclined to do research with regards to SDG 15 Life on Land. Consequently, this might lead to an underinvestment in knowledge production related to SDGs that are of less importance to a region. Additionally, Ciarli and Ràfols (2019) stress that there is often a misalignment between science and innovation, i.e., research 'priorities', and societal 'needs', thereby leading to sub-optimal configurations. The authors acknowledge the potential of SDGs in setting relevant directions and reversing the misalignment. Building on the above, it is expected that regional characteristics in the form of local adverse effects influence the knowledge production on SDGs. This leads to the following hypothesis:

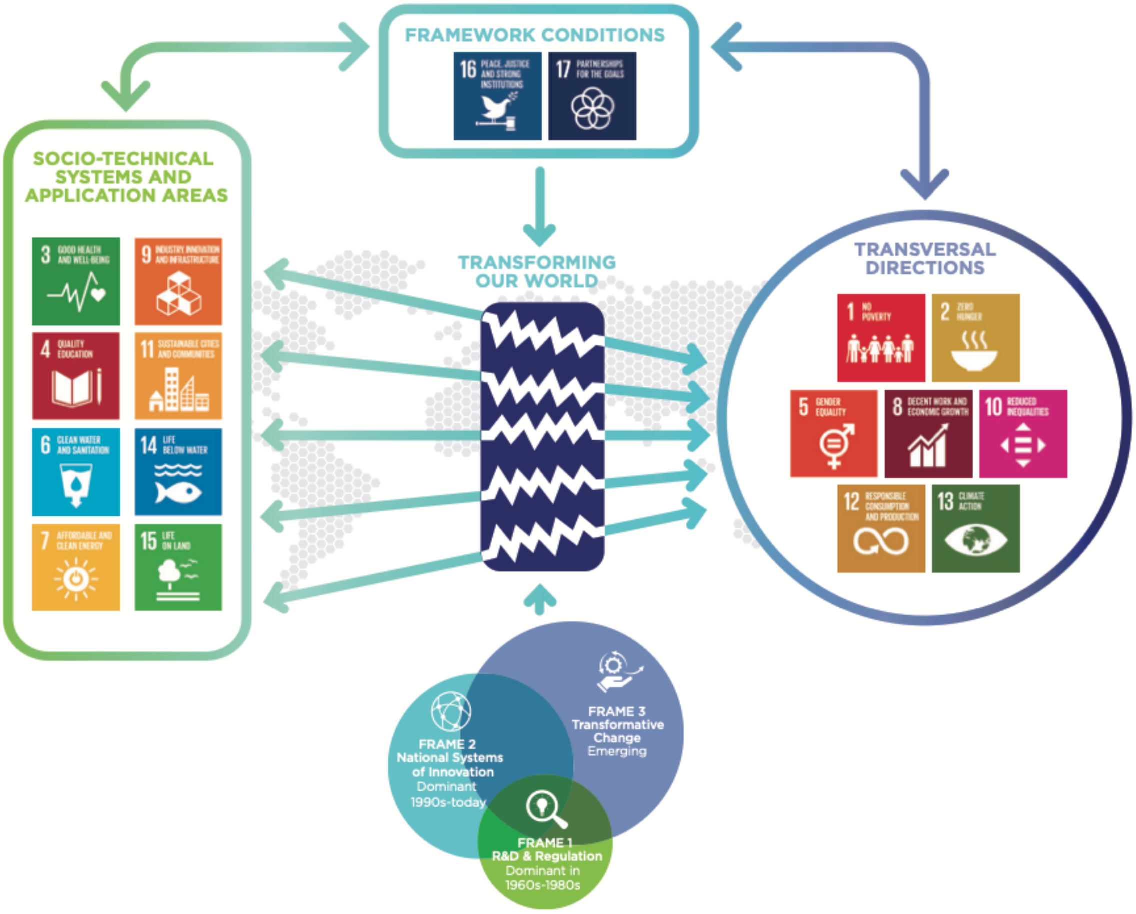
H3: Regions that experience relatively more adverse regional effects related to an SDG are more likely to engage in research related to that SDG.

2.5 Transformative lens

An important aspect of the SDGs is that they differ greatly from each other, as each is set to target another wicked problem in society. Besides, they can't be seen independently, as trade-offs and synergies between the different goals exist (Fuso Nerini et al., 2018). Consequently, knowledge production on SDGs is affected by these underlying relations. Ramirez et al. (2019) stress that to understand and account for this phenomenon a focus is needed on system transformation. Identifying knowledge production on the SDGs should be focused on transformative change, researching those SDGs that together have the greatest transformative potential. Taking this transformative lens approach, three different types of SDGs can be distinguished: First, SDGs that involve socio-technical systems and application areas (ST), which provide a basic need or service to society such as Clean Water and Energy (SDG6 and 7). Second, SDGs that provide directionalities to the future trajectories of socio-technical systems such as No Poverty and Gender Equality (SDG1 and 5), also called transversal directions (TD). Finally, to induce transformative change the participation of multiple actors is required. This is covered by the SDGs on Peace, Justice, and Strong Institutions and Inclusive Partnerships (SDG16 and 17), labeled as framework conditions (FC) (Ramirez et al., 2019; Schot et al., 2018).

Figure 2:

The Three Frames of Innovation



Note: From *Addressing the Sustainable Development Goals through Transformative Innovation Policy,* by Schot et al., 2018, p. 7.

Figure 2 gives an overview of this transformative perspective. Overall, transformative change is thus reached if research integrates all three types of SDGs. Research should include a focus on one or more socio-technical systems (ST), that need to be transformed in a certain direction (TD), thereby considering the partnerships (FC) necessary to reach the goal (Romero-Goyeneche et al., 2021). ST SDGs include both social and technological elements and transforming these elements involves knowledge production and diffusion that provide policymakers with potential directions of change based on TD SDGs. Besides, FC SDGs involve the change of governance and are thus less dependent on technological innovation (Schot et al., 2018). This differentiation between the SDGs allows for a more in-depth analysis of the results. As the SDGs differ greatly among each other, results are expected to vary among the different types of SDGs. ST SDGs involve systems required for the basic needs of society, including technological innovation, and are thus knowledge intensive. From this reasoning it is expected that regional patterns of complex knowledge production are more evident in ST SDGs compared to the others. This leads to the following hypothesis:

H4: Research on Socio-Technical SDGs is more affected by complexity, scientific relatedness, and adverse regional effects than research on other SDGs.

3. Methodology

3.1 Research design

This study followed a quantitative approach and focused at obtaining data from scientific publications and European statistics. Quantitative research provides the opportunity to use a large-scale sample and allows for results to be generalized and replicated (Eyisi, 2016). An appropriate approach to this research was the descriptive repeated cross-sectional research design. A cross-sectional study is an approach that allows the researcher to collect data from many different individual entities in order to collect quantitative data on several variables, which can be examined to identify certain patterns between them (Bryman, 2008). This design fitted well within this research as it enabled estimations on the prevalence of certain outcomes of interest and several outcomes and risk factors could be assessed (Levin, 2006). In this research, this allowed for a deeper understanding of the regional patterns and differences of SDG related research and factors underlying these regional patterns.

To operationalize the theoretical framework indicators were developed to explore the relationship between SDG related research and mechanisms of path- and place dependency together with regional characteristics. This allowed for a multiple regression analysis. The unit of analysis of this research was European regions, specifically as identified in the current NUTS 2021 classification. This is a geographical system developed by the European Union and divides the EU territory into hierarchical levels (European Union, 2020). The system only covers the member states of the EU, which means only these states were included in this research. This includes overseas departments and regions of European countries such as French Guiana. The EU contributed constructively to the development of the 2030 Agenda and thus provided an interesting case to understand how knowledge production has evolved since the adoption of this agenda (European Commission, n.d.-b). Data was collected on 16 out of 17 SDGs in two non-overlapping periods, to identify changes in regional patterns of complex knowledge production before and after the adoption of the 2030 Agenda. All analyses have been done using the statistical software R.

3.2 Data collection

Data was collected through desktop research on scientific publications and statistical databases. Scientific publications from the European regions were retrieved from the Centre for Science and Technology Studies (CWTS) wos_2113 database. This database contains cleaned bibliometric data from the Web of Science (WoS) of all publication up until 2021 week 13. Besides, it allows for easy coupling of publications to geographic locations based on the NUTS codes, which made it a very credible data source in this research. In addition, WoS is a high-quality database with a broad coverage that includes many journals and conference proceedings. Although social sciences and humanities are sparsely represented in the database, as well as books, it was still deemed an appropriate data source as one of the most used databases for peer-reviewed documents (Mingers & Leydesdorff, 2015). Scientific publications were collected for the time-period 2010-2020 to be able to make a distinction between publications in the period leading up to the adoption of the 2030 Agenda and the period hereafter. The time-periods were thus divided in two non-overlapping periods: 2010-2014 and 2015-2020, in order to capture different stages of the complex knowledge development on SDGs.

Furthermore, the specific publication on SDGs were collected with help of the CWTS wos_2113 database that allowed the linking of publications to results of the STRINGS-project. STRINGS stands for Steering Research and Innovation for Global Goals and the project, conducted by seven leading universities, research centers and the UNDP, contributes to mapping and visualizing research related to SDGs (Rafols et al., 2021). The thesaurus used by the project is based on policy agendas, has a focus on individual SDGs, consists of 3718 keywords, and describes 16 SDGs (excluding SDG 17) (Romero-Goyeneche et al., 2021). The project already assigned numerous publications in the CWTS database to be either SDG or non-SDG related, this allowed publications relevant for this research to be linked to the results of this mapping process. This means that only publications and SDGs that were considered within the STRINGS project were used in this research and therefore SDG 17 was not included. Although earlier research by Romero-Goyeneche et al. (2021) showed that the STRINGS thesaurus was somewhat more biased towards SDG 3, it is the focus of the thesaurus on individual SDGs and the fact that there is not one single preferred consensus on mapping SDGs to publications (Rafols et al., 2021), that made the use of this thesaurus still deemed feasible for this study.

To collect data on the adverse local effects of the different European regions, the Global SDG Indicator Database was consulted. This database is compiled through the UN System and contains data on 115 of the 230 SDG indicators (UN-iLibrary, n.d.). Data was only collected for indicators that relate to four of the Socio-Technical SDGs (SDG 3, 7, 14, and 15) and the Transversal Directions SDGs (SDG 1, 2, 10 and 13), due to time constraints. Besides, for the control variables the database from Eurostat was consulted based on the NUTS-2 level codes of regions. Eurostat is the European statistical office that publishes a variety of statistics and indicators that enable comparisons between regions over time (European Commission, n.d.-a). Data from the same time period as the scientific publications was selected and if data was not available for every single year than only the available years were used and accounted for. In case data was only available on country level, all regions from that specific country were attributed the same value.

3.3 Operationalization

In this section the operationalization of the concepts from the theoretical framework are presented. The SDG related research was identified as the dependent variable, where the other concepts of knowledge complexity, scientific relatedness and the regional characteristics together represent the independent variables. To control for other factors that might influence knowledge complexity, several control variables were included, as can be seen from table 1.

Table 1:
Overview operationalization of the theoretical framework

Category	Concept measured	Indicator	Description	Database
Dependent variable	SDG Related research	Share of SDG research	Percentage of total scientific output	WoS
Independent variables:	Knowledge complexity	Knowledge complexity index (KCI)	The diverse and unique knowledge that a region possesses	WoS
	Scientific relatedness	Relatedness density	The degree to which subfields are related to a region's knowledge base	WoS
	Adverse local effects	See table 2	Local effects that have an adverse effect on the region	SDG Indicators Database
Control variables	Level of economic development	GDP per capita	A country's GDP divided by its population	Eurostat
	Education level	Percentage of population with tertiary education	Relative level of education	Eurostat
	Population	Number of inhabitants	The number of inhabitants of the region	Eurostat

3.3.1 Dependent variable

SDG related research done per region was measured by calculating the amount of research on SDGs relative to the total amount of research done by the region. This results then in a percentage of the total scientific research output ($Pub_{S,r,t}$) as shown in formula (1).

$$Pub_{S,r,t} = \frac{\#Pub_{S_{rel},r,t}}{\#Pub_{tot,r,t}} * 100\% \qquad (1)$$

The total scientific research output *Pub* for a given SDG *S* in a region *r* at a certain time period *t* is thus given by the total amount of SDG related publications S_{rel} for a region in that time period divided by the total number of publications Pub_{total} of that region in a given time period. This formula was used to give the average scientific research output for all SDGs together and for individual SDGs. Using the research share instead of the exact amount of SDG related publications allowed for better comparison between regions, as some regions produce considerably more publications than others.

3.3.2 Independent variables

3.3.2.1 Complexity score

Knowledge complexity was measured through the knowledge complexity index (KCI) which is an extension of the model proposed by Hidalgo and Hausmann (2009) in their Method of Reflections and introduced by Balland and Rigby (2017). In the original model the level of complexity of a country is used to estimate the capability of a country to develop new products, which are tangible. However, the model is extended to incorporate knowledge, which is an intangible asset. In this research, scientific publications were clustered into scientific subfields and used to calculate the complexity of a region's knowledge base. Calculations were done through the EconGeo package in R (Balland, 2017). The first step was to distinguish the scientific subfields in the knowledge base of a region in which it has a Revealed Comparative Advantage (RCA). For this instance, the RCA measures whether a region *r* produces more clusters of scientific publications *p* related to a scientific subfield, as a share of its total clusters of scientific publications, than the 'average' region. The knowledge base of a region consists of several clusters of scientific publications that represent different (non) SDG related research areas. A region is considered to be a significant producer of scientific publications in a research area if $RCA_{rp} \geq 1$, where the RCA can be written as follows:

$$RCA_{rp} = \frac{S_{rp}}{T_p} \qquad (2)$$

Where,

$$T_p = \sum_r S_{rp} \qquad (3)$$

In the calculation of the KCI only those regions with an RCA in a certain subfield were considered, to prevent bias from regions with negligible knowledge. From this a bipartite network *M* could be created that separates the regions and clusters of subfields and is described by the adjacency matrix M_{rp}. Between these two groups links could be drawn that reflect in which clusters a certain region has an RCA, meaning that only a link is drawn between a cluster and a region if $RCA_{rp} \geq 1$ and thus M_{rp} gets a value of 1. Following from this, the diversity, the production of specific knowledge, and the ubiquity, the commonness of specific knowledge, could be calculated to get the KCI of a region. The diversity of a region is measured by its degree centrality ($K_{r,0}$) that is equal to the number of scientific subfields a region has an RCA in (4). The ubiquity of a region is also measured by its degree centrality ($K_{p,0}$) that is equal to the number of other regions that also have an RCA in a specific subfield (5).

$$K_{r,0} = \sum_p M_{rp} \qquad (4)$$

$$K_{p,0} = \sum_r M_{rp} \qquad (5)$$

Following the Method of Reflections, the measures of diversity and ubiquity could be expanded. The average value of the previous-level properties of a node's neighbor were calculated iteratively *N*-times leading to a more specific estimate (Hidalgo & Hausmann, 2009). The *N*-th iteration was reached when the value of *K* remained the same, so that saturation was reached, and no additional information could be retrieved. Regions with a higher KCI could then be said to have a more complex knowledge base, meaning that the region is diverse and relatively few other regions are able to reproduce their knowledge base (Balland & Rigby, 2017). This leads to the following equations for the average diversity ($K_{r,N}$) and ubiquity ($K_{p,N}$):

$$KCI = K_{r,N} = \frac{1}{K_{r,0}} \sum_p M_{rp} K_{p,N-1} \quad (6)$$

$$KCI = K_{p,N} = \frac{1}{K_{c,0}} \sum_p M_{rp} K_{r,N-1} \quad (7)$$

3.2.2.2 Relatedness density score

Scientific relatedness can be measured by calculating the relatedness density. This indicator measures the relatedness shown in a region between scientific subfields, from which diversification opportunities can be deduced. Before the relatedness between a scientific subfield and a region could be calculated, the scientific relatedness between scientific subfields had to be calculated first. This study followed the approach by Hidalgo et al. (2007) that base the relatedness ($\varphi_{i.j.t}$) between scientific subfields *i* and *j* on the conditional probability of having an RCA, thereby taking the minimum of the pairwise conditional probabilities as shown in equation 8. Only regions with an RCA in a scientific subfield were considered in order to avoid the inclusion of regions with a negligible knowledge base in a certain subfield.

$$\varphi_{i.j.t} = min\{P(RCAx_{i,t} | RCAx_{j,t}), P(RCAx_{j,t} | RCAx_{i,t})\} \quad (8)$$

Following, the knowledge base of a region could then be determined by looking at co-occurrences of scientific publications from that region to identify in which subfields it has an RCA. The knowledge base was given by the co-occurrences of publications related to that topic. Consequently, topics were considered to be related if they co-occurred together in different regions. Next, the relatedness density could be calculated that combines the relatedness between scientific subfields with the knowledge base of regions (i.e., the subfields in which a region publishes). This measure allowed to identify the share of a scientific subfield within a region that is built upon existing knowledge from a region. The relatedness density index was calculated following Heimeriks et al (2019) and Hidalgo et al. (2007):

$$RD_{i,r,t} = \frac{\sum_{j \in r, j \neq i} \varphi_{ij}}{\sum_{j \in r} \varphi_{ij}} * 100 \quad (9)$$

The relatedness of a subfield *i* in a region *r* at a certain time period *t* is equal to the sum of the scientific relatedness (φ_{ij}) of subfield *i* to all other subfields in which the region has an RCA, divided by the sum of the scientific relatedness of subfield *i* to all other subfields at time period *t* (9) (Heimeriks et al., 2019). The equation is multiplied by 100 to obtain a percentage between 0% and 100%. The higher the score the more related a subfield is to a region's knowledge base. Finally, the relatedness density between a scientific subfield and a region was further expanded to specify for SDG related research. The relatedness density between an SDG and a regions scientific knowledge base can be calculated as follows:

$$RD_{S1,r,t} = \frac{\sum_{S2 \in r, S2 \neq S1} \varphi_{S1S2}}{\sum_{S2 \in r} \varphi_{S1S2}} * 100 \quad (10)$$

The relatedness density of a certain SDG and a region was determined in the same way as it was for scientific subfields. The relatedness of an SDG *S1* in a region *r* at a certain time period *t* is equal to the sum of the scientific relatedness (φ_{S1S2}) to all other SDGs in which the region has an RCA, divided by the sum of the scientific relatedness of SDG *S1* to all other SDGs at time period *t.* The time period equals the two time periods indicated earlier from 2010-2014 and 2015-2020. Calculations were done through the EconGeo package in R.

3.2.2.3 Regional characteristics

The regional characteristics in a region were measured through several different indicators related to each selected SDG. This research considered two indicators per SDG to give a more robust insight. An overview of the indicators can be found in table 2. The indicators were selected based on a list, containing 230 indicators, developed by the Inter-Agency and Expert Group on SDG Indicators (UN General Assembly, 2022). In the table the different types of SDG are indicated, as well as which target from an SDG the goal comprises and the unit in which results are presented. All units give a relative value, which allowed for comparison between countries, as all indicators are presented on national level and no regional data was available. Only target 13.2 was an exception as this indicator provides an absolute value. Therefore, data was retrieved from the United Nations (2019-b) on the total population of countries on July 1st. This data was used to calculate the relative greenhouse gas emissions per year based on tonnes CO_2 equivalent per capita. Important to note is that for target 2.1.1 and 14.1.1 no data was available for the period 2010-2014, therefore these indicators were only used in the second time period (2015-2020).

Table 2:

Description of the SDG indicators used

SDG	Target	Indicator	Unit
1 (TD)	1.1.1	Proportion of population below the international poverty line	Percentage (%)
	1.4 .1	Proportion of population living in households with access to basic services (basic drinking water and sanitation services)	Percentage (%)
2 (TD)	2.1.1	Prevalence of undernourishment	Percentage (%)
	2.1.2	Prevalence of moderate or severe food insecurity in the population	Percentage (%)
3 (ST)	3.1.2	Proportion of births attended by skilled health personnel	Percentage (%)
	3.4.2	Suicide mortality rate	Per 100,000 population
7 (ST)	7.1.1	Proportion of population with access to electricity	Percentage (%)
	7.2.1	Renewable energy share in total final energy consumption	Percentage (%)
10 (TD)	10.2.1	Proportion of people living below 50 per cent of median income	Percentage (%)
	10.7.4	Number of refugees	Per 100,000 population
13 (TD)	13.1.1	Number of deaths and missing persons attributed to disasters	Per 100,000 population
	13.2.2	Total greenhouse gas emissions per year	Million tons of CO_2-equivalent
14 (ST)	14.1.1	Beach litter per square kilometer	Number
	14.5.1	Average proportion of Marine Key Biodiversity Areas (KBAs) covered by protected areas	Percentage (%)
15 (ST)	15.1.2	Average proportion of Terrestrial Key Biodiversity Areas (KBAs) covered by protected areas	Percentage (%)
	15.5.1	Red list Index, shows the conservation status of major species groups and measures trends in extinction risk over time	Index

Note: Based on indicators developed by the Inter-Agency and Expert Group (UN General Assembly, 2022)

3.3.3 Control variables

In addition to the independent variables, three control variables were added that could have an impact on complex knowledge production related to the SDGs, an overview is given in table 3. First, the economic development of a region measured through the Gross Domestic Product (GDP) per capita. The GDP is an indicator that reflects the market value of all final goods and services minus the value of goods and services that are used in intermediate consumption (Callen, 2020). By expressing GDP in PPS (Purchasing Power Standards), living standards can be more easily compared between regions, as differences in price levels between them are eliminated (Eurostat, 2022). Research has shown that complex knowledge production is correlated with a country's level of income, arguing that countries with a higher complexity score show a higher income per capita (Hidalgo & Hausmann, 2009). Therefore, GDP per capita is used to control for the effects of the economic development in each region.

Second, the educational level of a region, measured through the percentage of inhabitants with tertiary education. Tertiary education is an indicator that reflects the percentage of the population of a region that has obtained the highest level of education. This includes both theoretical and vocational programs (OECD, 2021). The higher the percentage of the population that obtained tertiary education, the more likely it is that a region can contribute to economic and social development as tertiary education is seen as a fundamental pillar of sustainable development (Salmi, 2017). Besides, earlier research on the economic complexity also controlled for education level as a factor that mediates economic complexity (Hidalgo, 2021). Therefore, also the education level was used as a control variable.

Finally, Balland et al. (2020) and Nomaler et al. (2014) showed that complex knowledge production concentrates disproportionately more in large cities as organizations in complex environments tend to agglomerate. Consequently, population in these areas will increase. In addition, Balland et al. (2020) found a strong positive linear relationship between knowledge complexity and metropolitan cities, with a higher urban concentration. Therefore, the effects of the increase in population around metropolitan cities were controlled for by including the number of inhabitants in a region.

Table 3:
Overview of control variables, their indicators, and units

Control Variable	Indicator	Unit
Population	Population on January 1st	Number
GDP	Regional gross domestic product	Purchasing Power Standard (PPS) per inhabitant
Education	Tertiary educational attainment, age group 25-64	Percentage (%)

3.4 Data overview

This section gives a brief overview of the data that was collected for this research. Overall, the NUTS 2021 Classification identifies 334 NUTS2-level regions, from which 333 regions were included in this research. The region Jan Mayen and Svalbard (NO0B) was excluded as there were no publications attributed to this region. In the case of Jan Mayen, this does not come as a surprise as it is a Norwegian volcanic island in the Arctic Ocean, without any permanent residents. Svalbard, however, is another Norwegian island in the Arctic Sea, also known as Spitzbergen, which does offer a research facility. The facility is however a state-owned limited company, and most researchers are not from Svalbard (UNIS, n.d.). This means the researchers are most likely affiliated to other regions, which might explain why Svalbard has not contributed to any publications in both time periods. Nevertheless, this was not seen as an impediment to this study.

Before linking the publications to the different SDGs with the results of the STRINGS project, all publications that belonged to either one of the time periods and contained a NUTS2-level code were retrieved from the CWTS wos_2113 database. The database assigns each publication a unique identifier (UT) to avoid double counting, which allows to identify the number of unique publications. Table 4 gives an overview of the number of publications retrieved for both time periods. As can be seen from the table, the final data sample used differs from the initial data sample. This is due to the fact that after linking the unique identifiers to the SDGs several publications could not be attributed to any of the SDGs or labeled as non-SDG. In order to avoid making estimations about the SDG share of these publications, they were excluded from the data sample. This means that this research was conducted with 77% of the initial data sample for time period 1 and with 80% of the initial data sample for time period 2, which was deemed feasible.

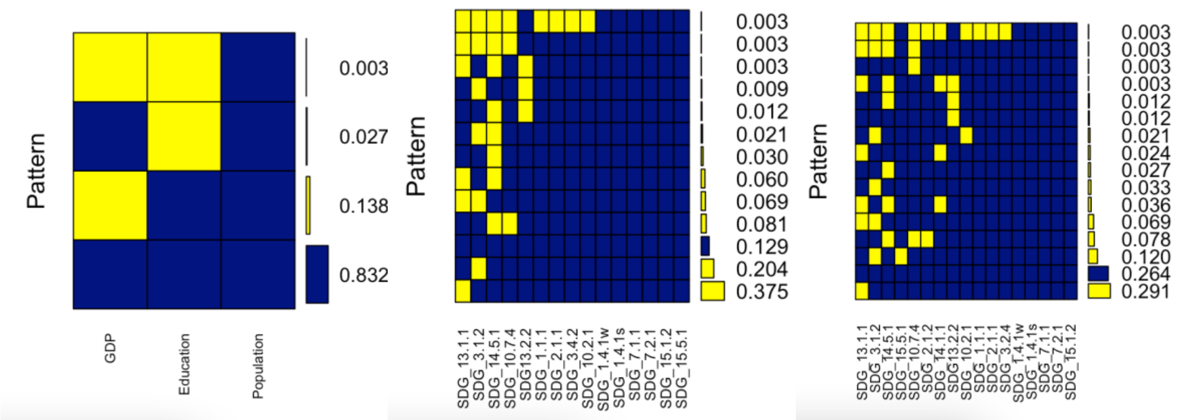
Table 4:
Overview data sample

	2010-2014	2015-2020
Initial data	3 497 996	5 108 484
Final data	2 702 630	4 061 550

3.5 Imputation of missing values

Figure 3 shows the missing value patterns for the control variables and the variables that explained the regional characteristics of the regions. On the left side of the figure, it illustrates that there were both missing values in GDP and Education, whereas population was complete. Besides, it shows that 83% of the data sample was complete, whereas 14% of the data contained missing values in GDP, 3% contained missing values in Education, and 0.3% of the data contained missing values in both GDP and Education. This pattern was the same for both time periods. In contrast, the regional characteristics, in the middle and the right of the figure, show that for the first time period only 13% of the data sample was complete and for the second time period 26%. Most missing values were found for both periods in the variable that indicated SDG target 13.1.1, number of deaths and missing persons attributed to disaster. The large difference in complete variables between the control variables and the regional characteristics was not surprising, as the regional characteristics comprised of far many variables then the control variables did.

Figure 3:
Missing data patterns



Note: The left side of the figure represents the control variables, the middle and right pattern in the figure belong to the regional characteristic variables for the period 2010-2014 and 2015-2020 respectively.

In order to deal with these missing values Multivariate Imputation via Chained Equations (MICE) was used in R to create complete datasets for the analyses. Besides, it allowed for an extra robustness check, as the use of multiple imputation reduces bias and increases efficiency (Azur et al., 2011). MICE was deemed an appropriate method as it imputes data on variable-by-variable basis, does not depend on normally distributed data and it performs well under most missing data conditions (Granberg-Rademacker, 2007). As all variables were numeric the default method used by MICE was predictive mean matching (PMM). PPM allows for the imputed variables to be more like 'real' values, thus if the variable is skewed the imputed values are skewed as well (Allison, 2015). This resulted in five multiple imputation datasets with slightly varying values for the missing data. These were pooled together into a complete dataset in R and imputed in the data sample.

3.6 Regression analysis

Data was analysed through a variety of methods. After a descriptive analysis of the results, a multiple regression analysis was performed. This is a statistical technique that allows a researcher to analyse the relationship between a dependent variable and multiple independent variables (Moore et al., 2006). The use of a multiple regression analysis allowed this study to identify to which extent the independent variables influenced the production of SDG related research. Besides, all variables were continuous which made them suitable for performing the analyses. Multiple regression analysis starts with the assumption that a linear relation between the dependent and independent variables exists (Tranmer et al., 2020). Therefore, to get a first impression of the data, scatterplots were created to visualize the relationship of the dependent and independent variables (see appendix A). A first glance at the plots showed that there was a somewhat positive linear relation between the SDG share and Knowledge Complexity and Scientific Relatedness. The regional characteristics, however, varied and no clear relationship could be found. To test the research question properly the standard equation for multiple linear regression (equation 11) was adjusted to fit the research (equation 12):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i \qquad (11)$$

$$SDG_{s,r,t} = \beta_0 + \beta_1 KCI_{r,t} + \beta_2 RD_{s,r,t} + \beta_3 RCa_{s,r,t} + \beta_4 RCb_{s,r,t} + \beta_5 P_{r,t} + \beta_6 GDP_{r,t} + \beta_7 E_{r,t} + e \qquad (12)$$

Where *SDG* is the SDG research share for a given SDG *s* attributed to a region *r* in time period *t*. β_0 is a constant that gives a predicted value of the SDG share when all independent variables are zero (Tranmer et al., 2020). β_p is the estimated regression coefficient that represent the change in SDG research share relative to a one unit change in any of the independent variables, while all others are held constant (LaMorte, 2016). The first independent variable is KCI, which represents the knowledge complexity of a region *r* at time period *t*. Following is RD, that gives the scientific relatedness density of an SDG *s* related to a certain region *r* in time period *t.* Hereafter comes RC, these are the regional characteristics *a* and *b* of an SDG *s* in a region *r* at a given time period. The formula varies per SDG as only for a selected group of SDGs the regional characteristics were studied. This means that this term was not present for some SDGs, for others it varied whether there were 1 or 2 terms for regional characteristics, dependent on the number of variables measured. The final three terms encompass the control variables population *P,* GDP, and the educational level *E* for a given region *r* in time period *t*.

For the regression analysis, four models were constructed in order to test the hypotheses. The analysis was conducted for two non-overlapping periods: 2010-2014 and 2015-2020. The first model was run for the time period 2010-2014, where the dependence of the SDG research share on the knowledge complexity and scientific relatedness, including the control variables, was studied. This was done for the combined SDG share of all SDGs and for those SDGS that were excluded from the further study into regional characteristics, an overview can be found in table 5. The second model involves the same variables as the first model; however, this regression model was estimated for time period 2015-2020. The third model involves again time period 2010-2014, and besides the knowledge complexity, scientific, relatedness and control variables, also the regional characteristics variables were included. In this model only those SDGs were included, that were selected to further analyse regional characteristics. The fourth model was again the same as the third model, but with data from time period 2015-2020.

Table 5:
Overview regression models

	Independent variables	Dependent	SDGs	Control
Model 1 2010-2014	Knowledge complexity Relatedness density score SDGs	SDG Research Share	4, 5, 6, 8, 9, 11, 12, and 16	GDP Population Education
Model 2 2015-2020	Knowledge complexity Relatedness density score SDGs			
Model 3 2010-2014	Knowledge complexity Relatedness density score SDGs Regional characteristics	SDG Research Share	1, 2, 3, 6, 10, 13, 14, and 15	GDP Population Education
Model 4 2015-2020	Knowledge complexity Relatedness density score SDGs Regional characteristics			

Before running these regression analyses, it was checked whether the variables fulfilled the other key assumption of multiple linear regression, namely no multicollinearity and multivariate normality. Multicollinearity occurs when two variables in a model are highly correlated. This could pose serious problems to a model, because as an effect the standard error of the regression coefficients becomes very large (Alin, 2010). In order to test for multicollinearity, the Pearson correlation coefficient was calculated for each model (see Appendix B). In general, it is assumed that severe multicollinearity is present between two variables if the correlation coefficient is larger than 0.8. Within the datasets no such large signs of multicollinearity were present. However, important to note, is that for SDG 7 the Pearson correlation could not be calculated for target 7.1.1 (Proportion of the population with access to electricity) as the standard deviation was 0, meaning all values were approximately equal. As this also posed problems for the regression analysis, it was decided to exclude this indicator from the analyses. Multivariate normality states that residuals should be normally distributed in order to fulfill the key assumptions of the analysis. Therefore, the assumption was tested by analyzing the P-P plots and histograms of the different regression models (see Appendix C). This showed that only SDG 3 and the combined SDGs for both time periods approached a close to normal distribution in its residuals.

In order to deal with the non-normal distribution of the variables the skewness coefficient was calculated to estimate the direction and value of the skew (see Appendix D). In general data is considered highly skewed if the skewness is outside the range of -1 till 1 (Wade, 2018). Highly skewed data was then to be found in all SDG research shares, population, and in the regional characteristics of SDG 1, 2, 7, 10, 13, 14 and 15. Therefore for the SDG research share and population a logarithmic transformation, by default a natural logarithm, was applied to the skewed data, in order to pull outliers inwards and overcome the effects of the skewed variables (Leydesdorff & Bensman, 2006). All variables regarding the SDG research share contained zeros in the datasets, therefore a small constant (k), ranging from 0.001 till 10, was added to avoid undefined values. The constant was selected based on a value equal to 10^N and chosen in such a way that skewness was the smallest possible value.

Looking at the variables that expressed the regional characteristics it could be seen that, besides some variables showing a severe skew, also left-skewed data was present. In order to overcome this problem, the data was reversed in order to make it right skewed. As all left-skewed variables were expressed in percentages this entailed that all values were subtracted from 1. Consequently, the variables should be interpreted differently, for example RC_2 and RC_3 that covered SDG indicator 1.4.1 now should be interpreted as the proportion of the population living in household that don't have access to basic services. Hereafter the variables could also be logarithmically transformed. However, for RC_3 from SDG1, and RC_1 from SDG2 the skewness could not be decreased using this method. A closer look at the data revealed that for SDG1.4.1w 76% and 82% of the sample could be attributed to a single value and for SDG2.1.1 94% and 97% of the sample, for 2010-2014 and 2015-2020 respectively. This posed a significant problem for the skewness of the data and consequently for the further analysis, therefore it was decided to exclude these indicators from the regression analysis.

An overview of the constants used in the logarithmic transformation and the skewness coefficient after the transformation can be found in table 6 and 7. The variables that have been transformed are shown in italics. In addition, there are several variables indicated with an asterisk, these variables were not logarithmically transformed, as this did not decrease skewness enough. Instead, a reciprocal square root transformation was used to decrease skewness. Variable SDG14.1.1, beach litter per square kilometer, decreased in skewness but did not reach a value between -1 and 1, as can be seen from table 7 for SDG14. As the indicator for beach litter varied greatly per country, the data was very skewed, nevertheless this data was deemed important for the regression analysis. Therefore, this indicator was not excluded from the analysis. For those regional characteristics variables that contained zeros, again a small constant was added, indicated between brackets in the columns. This transformation led to the following multiple regression equations for model 1 and 2 and for SDG 2 in model 3 (13):

$$\ln{(SDG)_{s,r,t}} = \beta_0 + \beta_1 KCI_{r,t} + \beta_2 RD_{s,r,t} + \beta_3 \ln{(P_{r,t})} + \beta_4 GDP_{r,t} + \beta_5 E_{,r,t} + e \tag{13}$$

For the other regression analyses in model 3 and 4 the regression equations varied, based on the transformations of the regional characteristic variables. For SDG1, 7, and 14 from model 3 and SDG1, 2, 7, and 14 equation 14 applied. For SDG3 from model 3 equation 15 applied. For SDG13 and 15 from model 3, and SDG 15 from model 4 equation 16 applied. Finally, for those regional characteristics that were transformed using the reciprocal square root, namely in SDG10 for model 3 and SDG3 and 10 for model 4, the regression equation is shown in equation 17.

$$\ln(SDG)_{s,r,t} = \beta_0 + \beta_1 KCI_{r,t} + \beta_2 RD_{s,r,t} + \beta_3 \ln\left(RCa_{s,r,t}\right) + \beta_4 \ln\left(RCb_{s,r,t}\right) + \beta_5 \ln{(P_{r,t})} + \beta_6 GDP_{r,t} + \beta_7 E_{,r,t} + e \tag{14}$$

$$\ln(SDG)_{s,r,t} = \beta_0 + \beta_1 KCI_{r,t} + \beta_2 RD_{s,r,t} + \beta_3 RCa_{s,r,t} + \beta_4 RCb_{s,r,t} + \beta_5 \ln{(P_{r,t})} + \beta_6 GDP_{r,t} + \beta_7 E_{,r,t} + e \tag{15}$$

$$\ln(SDG)_{s,r,t} = \beta_0 + \beta_1 KCI_{r,t} + \beta_2 RD_{s,r,t} + \beta_3 \ln\left(RCa_{s,r,t}\right) + \beta_4 RCb_{s,r,t} + \beta_5 \ln{(P_{r,t})} + \beta_6 GDP_{r,t} + \beta_7 E_{,r,t} + e \tag{16}$$

$$\ln(SDG)_{s,r,t} = \beta_0 + \beta_1 KCI_{r,t} + \beta_2 RD_{s,r,t} + \beta_3 \, {}^{1}\!/\!\sqrt{RCa_{s,r,t}} + \beta_4 RCb_{s,r,t} + \beta_5 \ln{(P_{r,t})} + \beta_6 GDP_{r,t} + \beta_7 E_{,r,t} + e \tag{17}$$

Universiteit Utrecht

Table 6:
Skewness coefficient after transformation (2010-2014)

SDG	Share	K	KCI	RD	RC_1	RC_2	RC_3	Pop	GDP	Edu
All	_0.28_	0	0.45	0.13				_-0.82_	0.76	0.47
1	_0.053_	0.001	0.45	0.88	_0.68 (0.001)_	_-0.43 (0.001)_	-2.66	_-0.82_	0.76	0.47
2	_0.070_	0.1	0.45	-0.13	4.74			_-0.82_	0.76	0.47
3	_-0.25_	10	0.45	0.57	-0.61	0.46		_-0.82_	0.76	0.47
4	_0.30_	0.1	0.45					_-0.82_	0.76	0.47
5	_-0.093_	0.1	0.45	0.81				_-0.82_	0.76	0.47
6	_-0.48_	0.1	0.45	0.11				_-0.82_	0.76	0.47
7	_-0.44_	0.1	0.45	0.02		_0.16_		_-0.82_	0.76	0.47
8	_-0.12_	0.1	0.45	0.29				_-0.82_	0.76	0.47
9	_-0.21_	0.1	0.45	0.27				_-0.82_	0.76	0.47
10	_0.42_	0.1	0.45	0.92	0.37	_-0.32*_		_-0.82_	0.76	0.47
11	_-0.29_	0.1	0.45	0.20				_-0.82_	0.76	0.47
12	_-0.23_	0.1	0.45	0.04				_-0.82_	0.76	0.47
13	_-0.35_	0.1	0.45	-0.03	_-0.03_	0.72		_-0.82_	0.76	0.47
14	_-0.27_	0.01	0.45	0.11		_0.77 (1)_		_-0.82_	0.76	0.47
15	_-0.25_	0.1	0.45	0.41	_-0.17_	-0.26		_-0.82_	0.76	0.47
16	_-0.48_	0.01	0.45	0.59				_-0.82_	0.76	0.47

Note: The coefficients displayed in italics where logarithmically transformed, the coefficient in italic with an asterisk was transformed using a reciprocal square root transformation

Table 7:
Skewness coefficient after transformation (2015-2020)

Model	Share	K	KCI	RD	RC_1	RC_2	RC_3	Pop	GDP	Edu
All	_0.54_	0	0.39	0.19				_-0.79_	0.78	0.44
1	_0.34_	0.01	0.39	0.80	_0.95 (0.001)_	_-0.28 (0.001)_	-2.68	_-0.79_	0.78	0.44
2	_0.32_	0.1	0.39	0.19	6.19	_0.98_		_-0.79_	0.78	0.44
3	_-0.20_	10	0.39	0.40	_0.11 (0.01)_	0.18		_-0.79_	0.78	0.44
4	_0.19_	0.1	0.39	0.79				_-0.79_	0.78	0.44
5	_-0.050_	0.1	0.39	0.95				_-0.79_	0.78	0.44
6	_-0.55_	0.1	0.39	0.12				_-0.79_	0.78	0.44
7	_0.13_	1	0.39	0.04		_0.83_		_-0.79_	0.78	0.44
8	_-0.46_	0.1	0.39	0.24				_-0.79_	0.78	0.44
9	_-0.53_	0.1	0.39	0.10				_-0.79_	0.78	0.44
10	_0.075_	0.1	0.39	0.96	0.03	_-0.67*_		_-0.79_	0.78	0.44
11	_0.67_	1	0.39	0.20				_-0.79_	0.78	0.44
12	_-0.45_	0.1	0.39	0.10				_-0.79_	0.78	0.44
13	_-0.24_	0.1	0.39	-0.05	_0.26*_	0.46		_-0.79_	0.78	0.44
14	_-0.17_	0.01	0.39	0.01	_-1.05_	_-0.75 (1)_		_-0.79_	0.78	0.44
15	_-0.21_	0.1	0.39	0.38	0.05	-0.25		_-0.79_	0.78	0.44
16	_0.41_	0.1	0.39	0.67				_-0.79_	0.78	0.44

Note: The coefficients displayed in italics where logarithmically transformed, the coefficients in italic with an asterisk were transformed using a reciprocal square root transformation

4. Results

In this section the findings of the study are presented. First the results of the descriptive analysis, that gives a general understanding of the knowledge production within regions, is summarized, and visualized. Next, the results of the analysis on SDG related research within the regions is explored and visualized, where also the regression models are estimated. Finally, the difference between the SDGs is further analyzed and the hypotheses are explored and to what extend they can be accepted.

4.1 Regional knowledge production

To give an overview of the knowledge production in the different European regions, the number of publications on a regional level for the two time periods (2010-2014 and 2015-2020) were determined, from which the results are visualized in figure 4. Fractional counting was applied to allocate publications to regions. This was deemed the most appropriate method as some publications were affiliated with multiple regions and this avoided double counting when aggregating the data and comparing the publications across regions. In general, the figures show a skewed distribution of publications among regions, where Western Europe produces on average more publications than Eastern Europe. The least productive regions in terms of knowledge production have only produced a maximum of 2600 publications over a time period of six years, while the most productive regions have produced at least six times this amount. Besides, it shows that regions that produce a large number of publications often border each other and can be considered neighbors. This can also be seen from table 8, that shows the top 10 most publishing regions. Here it becomes apparent that the number of publications is geographically bounded, so that high-publishing regions belong to the same country, as is the case for the United-Kingdom, Italy, Spain, and Germany. Finally, an observation that can be made from table 8 is that all high-publishing regions can be connected to metropolitan areas in the region, such as London, Paris, and Milan. This confirms the fact that knowledge production tends to reside in larger cities since organizations tend to agglomerate due to proximity advantages (Balland et al., 2020; Nomaler et al., 2014).

Figure 4:

Maps of Europe showing the total number of publications per region for two time periods.

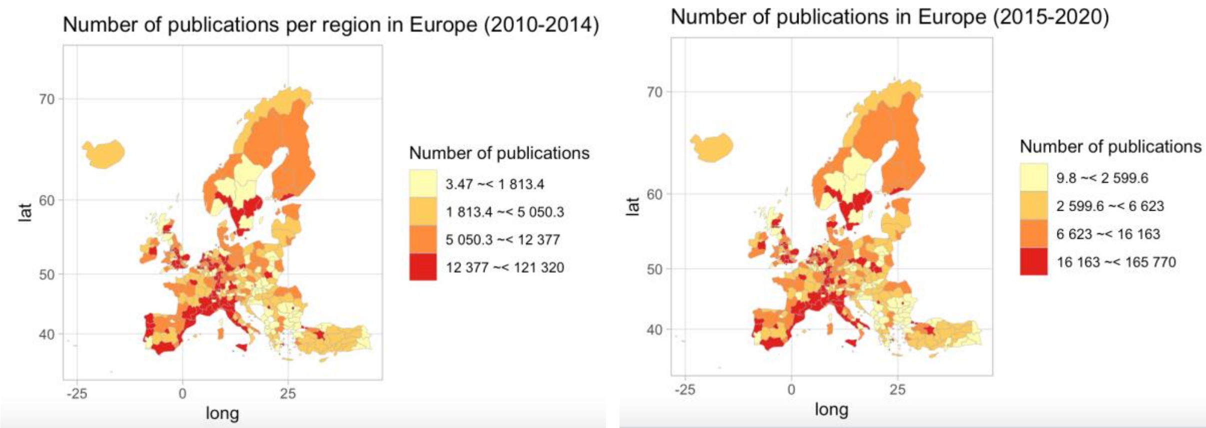


Table 8:
Top 10 most publishing regions in Europe for two time periods including NUTS2-level codes.

2010-2014			2015-2020		
NUTS	**Region**	**Number of publications**	**NUTS**	**Region**	**Number of publications**
UKI3	Inner London-West	121 317	UKI3	Inner London-West	165 771
FR10	Île-de-France	101 869	FR10	Île-de-France	122 866
ES30	Madrid	55 397	ITC4	Lombardia	75 367
ES51	Catalonia	53 589	ES30	Madrid	74 129
ITC4	Lombardia	51 255	ES51	Catalonia	69 257
UKJ1	Berkshire	44 743	ITI4	Lazio	61 093
ITI4	Lazio	44 255	UKJ1	Berkshire	58 533
DE21	Oberbayern	42 721	DE21	Oberbayern	55 807
DE30	Berlin	40 887	DE30	Berlin	51 424
UKH1	East Anglia	40 056	NL33	Zuid-Holland	50 717

The total number of publications was, however, not an appropriate indicator for the quality of the knowledge produced per region or its ability to produce knowledge regarding the SDGs. Therefore, the knowledge complexity and average relatedness density were also determined per region, as shown in figure 5 and 7. The knowledge complexity index provides more insight into the complexity of the knowledge produced in regions. If a region possesses a complex knowledge base it is more likely to be capable to produce knowledge on other complex phenomena such as the SDGs. Besides, the average relatedness density gives an indication of how well connected a region is to other scientific subfields. Regions with a high relatedness density can more easily diversify into a variety of subfields as their knowledge produced is highly related to these. In contrary, if regions have a low relatedness density their knowledge base is limited to a specific field, with little existing knowledge in other subfields. This makes it harder for a region to create 'new' knowledge in other scientific subfields.

The scientific subfields with the highest knowledge complexity scores are given in table 9. From this table it can be seen that the most complex subfields belong to the disciplines of Language, Information and Communication, Law, Arts and Humanities, Social and Behavioral Sciences and Medical and Life Sciences. These are predominantly social subfields and are likely to belong to the most complex subfields due to their multidisciplinary nature. Besides, the data source used, namely WoS, is known to represent these subfields only to a limited extent, which also could explain the high complexity scores. Remarkable is that disciplines of Engineering Sciences and Natural Sciences don't belong to the most complex subfields, this might be due to the fact that relatively a lot of regions operate in these subfields. The knowledge complexity score of a region is dependent on de complexity of the different subfields, as regions that engage in research related to more complex scientific subfields are likely to have a higher complexity score.

Table 9:
Top 10 most complex scientific subfields in Europe

2010-2014		2015-2020	
Subfield	**KCI Score**	**Subfield**	**KCI Score**
Poetry	99.59	Poetry	97.28
Social Sciences, Biomedical	89.36	Social Sciences, Biomedical	91.59
Law	88.23	Health Care Sciences & Services	85.72
Literature, African, Australian, Canadian	87.56	Law	85.10
Social Work	85.18	Humanities, Multidisciplinary	84.77
International Relations	84.95	Criminology & Penology	84.42
Criminology & Penology	84.73	Political Science	82.47
Ethnic Studies	84.58	Social Work	82.14
Film, Radio, Television	84.21	Cultural Studies	81.51
Primary Health Care	83.39	Primary Health Care	81.47

The knowledge complexity score per region is given in figure 5 for both time periods, which differ only slightly from each other. The figure clearly illustrates that North-Western European regions have a higher degree of knowledge complexity than regions in the South and East of Europe. Especially large parts of the United Kingdom have a highly complex knowledge base. This can also be seen from table 10 that shows the top 10 KCI scores of the regions, where most of the regions are located in the UK. Other regions that perform well are Limburg and Noord-Holland in the Netherlands and the Eastern and Midland of Ireland. Knowledge complexity is dependent on the amount of knowledge a region can accumulate and is therefore often associated with productivity (Rodrigues & Breach, 2021). This also explains why regions that produce a high number of publications, such as London (UKI3) and Berkshire (UKJ1), are considered highly complex. Besides, the table shows that the knowledge complexity scores slightly increase over time, indicating that regions grow to be more complex. This is not unexpected as more complex regions can benefit from knowledge networks and expand their research fields into other complex areas.

Figure 5:
Map of Europe showing the KCI scores per region

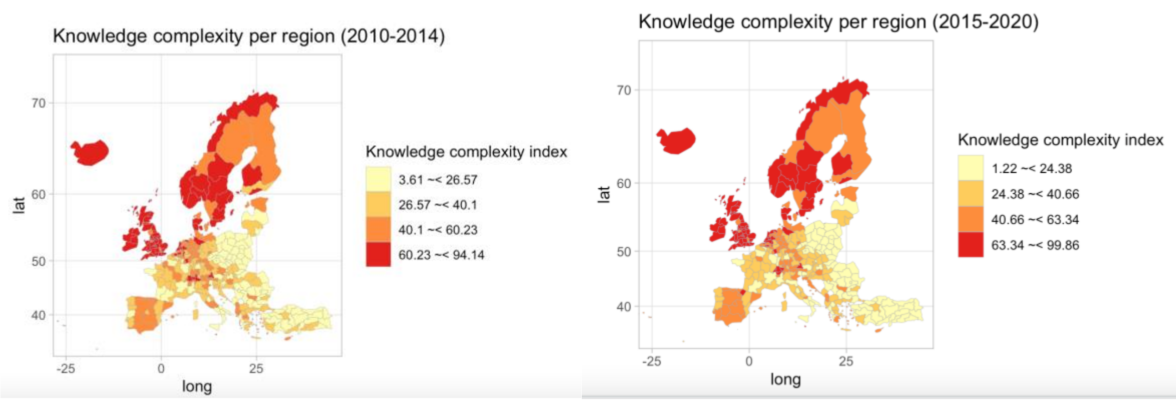
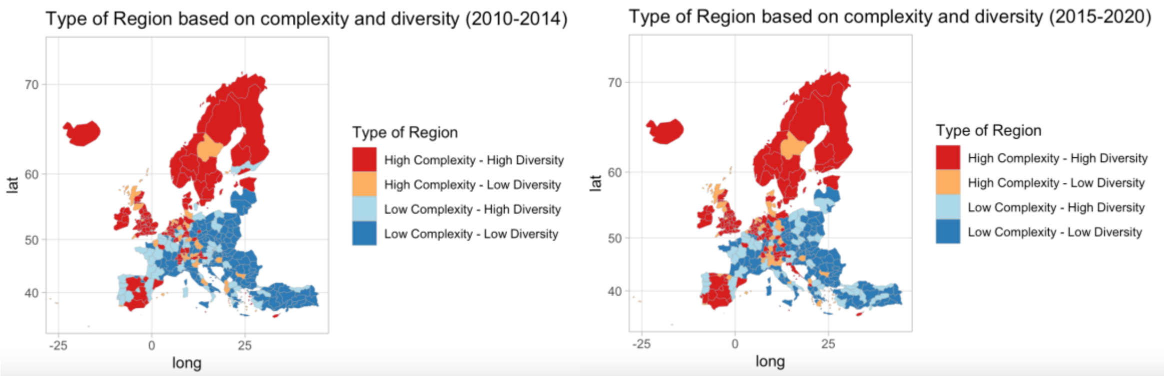


Table 10:
Top 10 most complex regions of Europe

2010-2014			2015-2020		
NUTS	**Region**	**KCI Score**	**NUTS**	**Region**	**KCI Score**
UKI3	Inner London-West	94.14	UKI3	Inner London-West	99.86
NL32	Noord-Holland	88.77	NL32	Noord-Holland	96.48
UKG2	Shropshire and Staffordshire	88.19	UKH3	Essex	91.94
UKJ2	Surrey, East, and West Sussex	87.75	UKJ2	Surrey, East, and West Sussex	91.87
UKG3	West Midlands	87.32	UKG2	Shropshire and Staffordshire	91.51
UKK4	Devon	86.90	UKK4	Devon	91.00
UKE1	East Yorkshire and Norther Lincolnshire	86.42	NL42	Limburg	90.43
UKD4	Lancashire	86.00	UKJ1	Berkshire, Buckinghamshire, and Oxfordshire	90.22
NL42	Limburg	85.93	UKD4	Lancashire	89.78
UKJ1	Berkshire, Buckinghamshire, and Oxfordshire	85.12	IE06	Eastern and Midland	89.50

Besides the knowledge complexity, it is also important to look at the diversity of the knowledge base of regions in order to better understand the capability of a region to diversify to other (more complex) subfields. Regions can be highly complex, but if their knowledge base is very specialized their ability to diversify will be compromised. The data retrieved on the complexity scores of the regions was therefore split to identify the overall scientific complexity and diversity. For each time period the median value of knowledge complexity and diversity was used, to be able to distinguish regions based on four categories, namely high complexity and diversity, low complexity and diversity, low complexity and high diversity, and high complexity and low diversity. Results of this analysis are shown in figure 6 and indicate that complex knowledge production is indeed very concentrated in the North-Western regions of Europe and to some extent in Spain. Spain has over the last decade shifted its priorities towards scientific research and increasingly understood the importance of knowledge and innovation (Bertero, 2009). This might be a reason as of why Spain scores exceptionally well in this area, compared to neighboring countries. Several regions in France, Italy, and Portugal, are mainly characterized by low complexity of knowledge, while a high diversity of this less complex knowledge exists. In contrast, some regions in Northern-Italy and Swiss show levels of high complexity, while their knowledge base is not diverse. This indicates that these regions are more specialized but do possess some highly complex knowledge. Finally, Western Europe produces, as expected, knowledge of low complexity and diversity. Overall, this indicates that regions with a diverse and complex knowledge base should be more likely to diversify into complex fields of research such as the SDGs. Consequently, regions that are not diverse and/or complex will not be able to diversify into SDGs and should focus on taking incremental steps towards more complex knowledge and from there on develop capabilities for future research into the SDGs.

Figure 6:
Map of European regions based on their overall scientific complexity and diversity



In order for a region to be capable to diversify into other subfields it is important that they are related to each other, so that existing knowledge can be used to build upon the new knowledge from these scientific subfields. The relatedness density score of a region determines this and indicates the average relatedness of a region to any given scientific subfield. The higher the score the easier it is for a region to connect to new fields. Figure 7 shows the results of this analysis, which are similar to earlier results on the scientific complexity. The highest relatedness density scores can be found in North-Western Europe, with the exception of some regions in Scotland, Spain, and Portugal. Regions with a low relatedness density score can be mainly found in Central and Eastern Europe. Table 11 confirms this and shows that the regions with the highest average relatedness score are located mostly in the United Kingdom, and to some extent in Belgium and Portugal. Striking is to see that again the UK scores very high, but these regions differ from those that scored high on the knowledge complexity, except for London (UKI3). However, if results are checked beyond the top 10, it becomes apparent that most of these regions, with the exception of Shropshire and Staffordshire (UKG2) that is on place 101, are within the top 50 of best performing regions in terms of average relatedness. Finally, in general it can be said that the average relatedness density has decreased somewhat over time, indicating that regions might specialize more.

Figure 7:
Maps of Europe showing the average relatedness density score per region

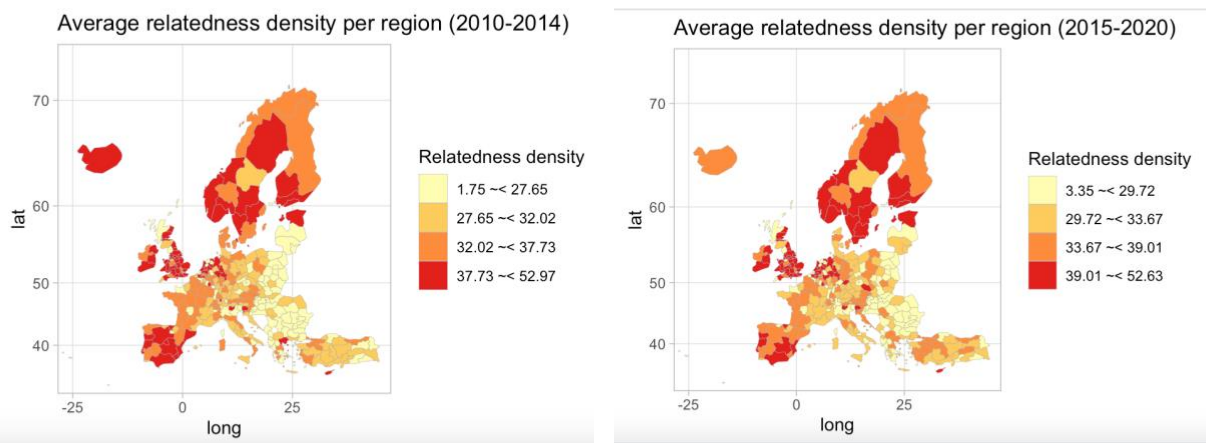


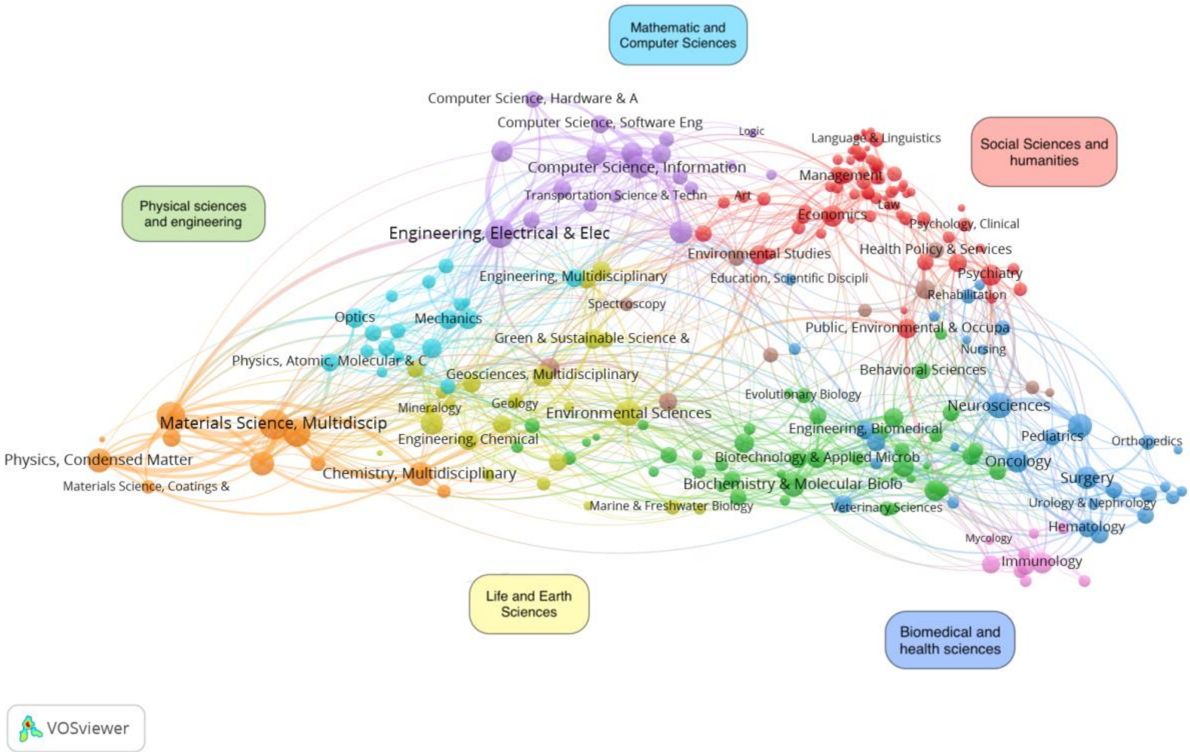
Table 11:
Top 10 regions with the highest relatedness density score in Europe

2010-2014			2015-2020		
NUTS	**Region**	**RD Score**	**NUTS**	**Region**	**RD Score**
BE23	Prov. Oost-Vlaanderen	52.97	UKK1	Gloucestershire, Wiltshire, and Bristol	52.63
UKM7	Eastern Scotland	52.95	BE23	Prov. Oost-Vlaanderen	52.36
UKF1	Derbyshire and Nottinghamshire	52.73	UKM7	Eastern Scotland	52.11
UKL1	West Wales and The Valleys	52.60	UKF1	Derbyshire and Nottinghamshire	51.62
UKK1	Gloucestershire, Wiltshire, and Bristol	52.60	UKF2	Leicestershire, Rutland, and Northamptonshire	50.44
UKF2	Leicestershire, Rutland, and Northamptonshire	49.88	UKL1	West Wales and The Valleys	50.14
UKE4	West Yorkshire	49.59	UKG3	West Midlands	49.85
UKE3	South Yorkshire	49.36	UKE2	North Yorkshire	49.59
UKI3	Inner London-West	48.75	PT17	Metropolitan area of Lisbon	49.16
UKJ3	Hampshire and Isle of Wight	48.71	UKE3	South Yorkshire	48.58

Another observation that resulted from comparing the maps of the number of publications, the knowledge complexity, and the relatedness density is that the best performing regions in terms of knowledge complexity and relatedness are somewhat more located in North-Western Europe, while the number of publications map is very skewed across Europe and shows a slightly higher number in Southern France and Northern Italy. This reinforces the suspicion that the quantity of knowledge produced does not necessarily say anything about the quality of this knowledge. Besides, if the high-scoring regions from the knowledge complexity map are compared to the relatedness density map, it can be seen that the average relatedness is somewhat more distributed across Europe, while the high-scoring regions in the complexity map are mainly located in North-Western Europe. This indicates that especially regions in the South of France, Spain and Portugal show a strong relatedness to the average scientific subfield but have not been capable to produce more complex knowledge over time despite of this. Opposite observations were seen in some Swiss regions, where knowledge complexity was high but average relatedness density relatively low. This might indicate, as observed earlier, that these regions are highly specialized and do not rely on a broad set of capabilities but use specialized skills and resources (Heimeriks et al., 2019).

The relatedness of regions to different scientific subfields is dependent on the co-occurrences of these subfields. To gain deeper insight into these co-occurrences a scientific relatedness space was constructed with the use of VOSviewer (van Eck & Waltman, 2010), based on the occurrence of scientific subfields in the publications of the different European regions. This field is visualized for period 2015-2020 in figure 8 and shows which scientific subfields are often found together in a publication. The closer these subfields are located in the map, the more often they appear together. Additionally, dots represent the different scientific subfields, the size of them gives an indication of the number of co-occurrences. Thus, the larger the dot, the more often a scientific subfield co-occurs and thus the more publications belong to this scientific subfield. The scientific relatedness space can be subdivided into five broad research areas: Physical Sciences and Engineering (the orange and light blue dots), Mathematics and Computer Sciences (purple dots), Social Sciences and Humanities (red dots), Biomedical and Health Sciences (green and dark blue dots), and Life and Earth Sciences (yellow dots). Although these subfields are connected, they are highly distinct, and it would thus be rare if a region would have capabilities in each of these broad research areas. More likely is that regions specialize in only one or two and develop these capabilities through related scientific subfields, rather than trying to diversify into all areas.

Figure 8:
Scientific relatedness space of scientific subfields (2015-2020).



Finally, to interpret the data of the relatedness space map better a table was created that shows the top 10 scientific subfields that co-occur the most with each other for both time periods, see table 12 and 13. What becomes apparent is that all co-occurrences in the tables are related to either natural sciences or engineering sciences. This also explains why these research fields where not among the most complex scientific subfields, as most research is thus done in these research fields. Consequently, this increases the ubiquity of these subfields and therefore lowers their complexity score.

Table 12:
Top 10 most co-occurring scientific subfields (2010-2014)

Subfield 1	Subfield 2	N
Physics, Applied	Physics, Condensed Matter	51 432
Materials Science, Multidisciplinary	Physics, Applied	46 329
Clinical Neurology	Neurosciences	39 232
Computer Science, Interdisciplinary Applications	Computer Science, Software Engineering	37 686
Chemistry, Physical	Materials Science, Multidisciplinary	35 315
Materials Science, Multidisciplinary	Physics, Condensed Matter	33 649
Engineering, Electrical & Electronic	Telecommunications	29 577
Materials Science, Multidisciplinary	Nanoscience & Nanotechnology	28 944
Biochemistry & Molecular Biology	Cell Biology	25 802
Engineering, Electrical & Electronic	Physics, Applied	25 342

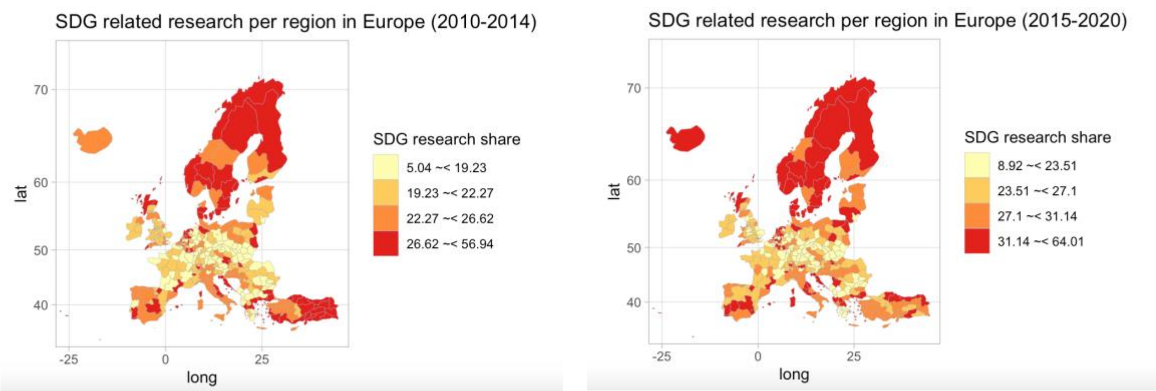
Table 13:
Top 10 most co-occurring scientific subfields (2015-2020)

Subfield 1	Subfield 2	N
Materials Science, Multidisciplinary	Physics, Applied	75 466
Clinical Neurology	Neurosciences	64 255
Physics, Applied	Physics, Condensed Matter	62 146
Chemistry, Physical	Materials Science, Multidisciplinary	56 462
Engineering, Electrical & Electronic	Telecommunications	52 502
Materials Science, Multidisciplinary	Nanoscience & Nanotechnology	50 852
Materials Science, Multidisciplinary	Physics, Condensed Matter	45 329
Computer Science, Interdisciplinary Applications	Computer Science, Software Engineering	38 863
Chemistry, Multidisciplinary	Materials Science, Multidisciplinary	34 008
Engineering, Electrical & Electronic	Physics, Applied	33 428

4.2 Regional knowledge production on SDGs

To give an overview of the knowledge production in the different European regions regarding the SDGs, the research share of SDG related research for the two time periods (2010-2014 and 2015-2020) was determined. In order to properly count the number of SDG related publications to determine the research share, publications were attributed a 1 for SDG related and a 0 for non-SDG related research. This way double counting was avoided, as some publications were assigned to multiple SDGs. These weights were then combined with the NUTS2 level codes of the regions and the affiliation weight of the unique identifier (UT) of the publications. This allowed for SDG related research to be counted fractionally. Findings were then combined with the fractionally counted overall publications to determine the research share. Results of this analysis are displayed in figure 9. In general, the figure shows a somewhat skewed distribution of the SDG research share, although Northern Europe seems to outperform the rest of Europe. Over time it seems as Northern Europe increased in research share, as several Scandinavian regions and Iceland moved from the orange to the red sector. Regions in Turkey on the other hand seem to have a decreasing SDG research share, as they have moved from red to orange. Overall, it can be concluded that the SDG research share increased over time, as the lowest share was 5.04 for the earlier time period and 8.92 for the following period. Besides, maximum research share increased from 56.94 to 64.01 percent.

Figure 9:
Research share of SDG related research per region for both time periods



A closer interpretation of the data is given in table 14 and 15. That show the top and bottom 10 regions regarding their SDG share for both time periods. Remarkable is that most regions in the top 10 SDG share are exclaves, overseas departments, or islands from European countries, such as Mayotte, Melilla, and Corse. A possible explanation for these observations might be that these regions do not produce a lot of publications, but the publication that they produce are SDG related. Guyana, Réunion, Mayotte and the Azores belong to Europe's Outermost Regions (ORs), to which the obligations of the European Treaties fully apply. Besides these regions know several constraints compared to regions in Europe due to their remoteness, small size, and adverse climatic conditions (European Commission, 2017). This might have as a consequence that the limited knowledge developed is much more steered towards complying with the European goals for sustainable development and therefore these regions have a higher research share compared to other European regions. In contrary, the bottom 10 regions are mostly located in central and Western Europe, such as Bulgaria, Albania, and Poland. This does not come as a surprise, as these regions are known for their low complexity and diversity and thus probably lack the capability to diversify into more complex research such as the SDGs. More striking is the fact that some regions in Western Europe also perform relatively poor, while these regions are known to be more complex and diverse. This can be however explained by the fact that the research share is used. When the SDG related publications count is used it can be seen that the Western regions actually produce more SDG related publications and that some of the Outer Regions are at the bottom of the publication count (see Appendix E).

Table 14:
Top 10 regions regarding SDG research share in Europe

2010-2014			**2015-2020**		
NUTS	**Region**	**SDG Share**	**NUTS**	**Region**	**SDG Share**
FRY5	Mayotte	56.94	FRY5	Mayotte	64.01
ES64	Melilla	54.49	ES64	Melilla	62.28
NL34	Zeeland	49.26	FRY4	La Réunion	58.52
FRY4	La Réunion	49.10	NL34	Zeeland	58.25
FI20	Åland	47.43	FRY3	Guyane	57.44
PT20	Azores	46.20	FRM0	Corse	57.06
FRM0	Corse	46.19	NL23	Flevoland	53.59
UKM6	Highlands and Islands	44.47	PT20	Azores	51.12
FRY3	Guyane	39.47	UKM6	Highlands and Islands	49.81
NO07	Nord-Norge	48.71	ES63	Ceuta	48.58

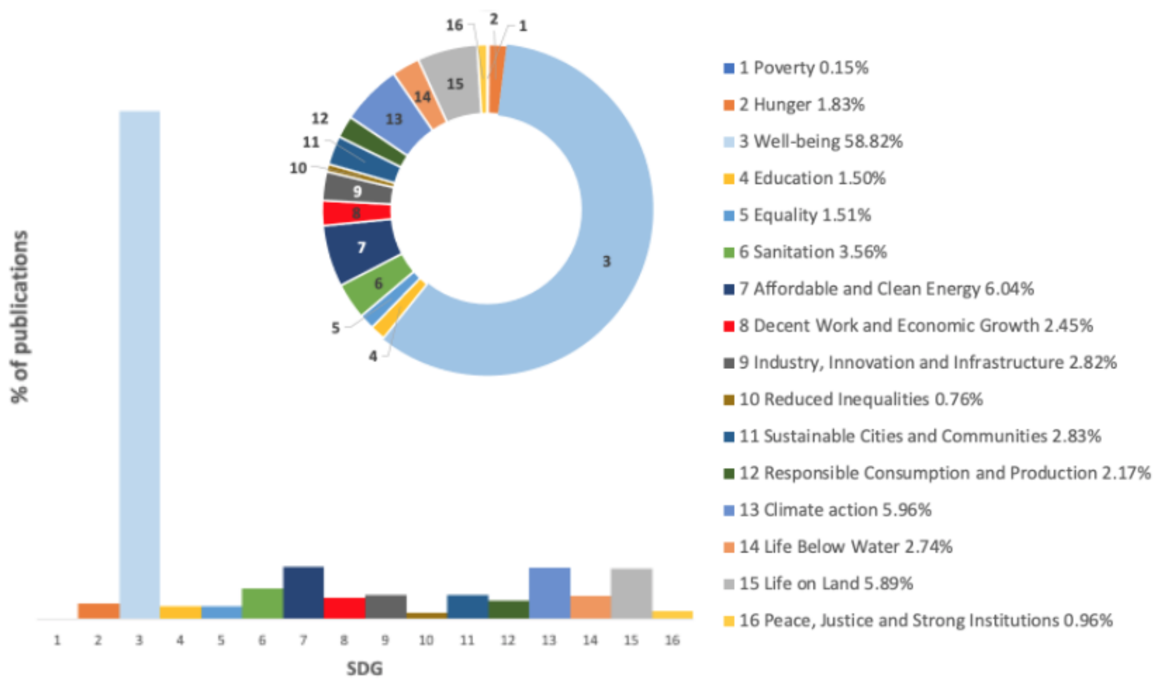
Table 15:
Bottom 10 regions regarding SDG research share in Europe

2010-2014			**2015-2020**		
NUTS	**Region**	**SDG Share**	**NUTS**	**Region**	**SDG Share**
BG32	Severen Tsentralen	5.04	DED4	Chemnitz	8.92
DED4	Chemnitz	7.18	BG32	Severen Tsentralen	11.39
AL03	Jug	8.61	BG31	Severozapaden	12.48
AL01	Veri	9.07	AT31	Oberösterreich	13.21
NL13	Drenthe	10.71	AL01	Veri	14.21
PL43	Lubuskie	10.73	AT21	Kärnten	14.69
AT31	Oberösterreich	11.25	NL13	Drenthe	16.21
DEA4	Detmold	11.66	DED2	Dresden	16.30
PL52	Opolskie	11.91	UKC1	Tees Valley and Durham	16.63
CZ08	Moravskoslezsko	12.22	DEA4	Detmold	16.93

In order to get insights into the distribution of the individual SDGs, a visual representation is given in figure 10. This representation is only shown for time period 2015-2020, as it shares great similarity with the other time period (see Appendix F). These results show that the largest research share among the SDGs can be attributed to SDG3 (Health and Wellbeing) and to some extent to SDG7 (Affordable and Clean Energy), SDG13 (Climate action), and SDG15 (Life on Land). The fact that SDG3 is greatly overrepresented does not come as a surprise, as earlier was indicated that the thesaurus used to distinguish the SDG related research has a bias towards SDG3. Other thesauruses have shown that many publications labeled as SDG3 could also be assigned to SDG1 (Poverty) or SDG2 (Hunger) (Romero-Goyeneche et al., 2021), who only contributed to a small share of the SDG related research done in this study. Besides, this dataset is based on publication from the WoS, which is known to favor natural science publications over social sciences and humanities (Mingers & Leydesdorff, 2015). Consequently, some more social SDGs such as SDG5 (Gender Equality) and SDG10 (Reduced Inequalities), are also underrepresented in the data sample.

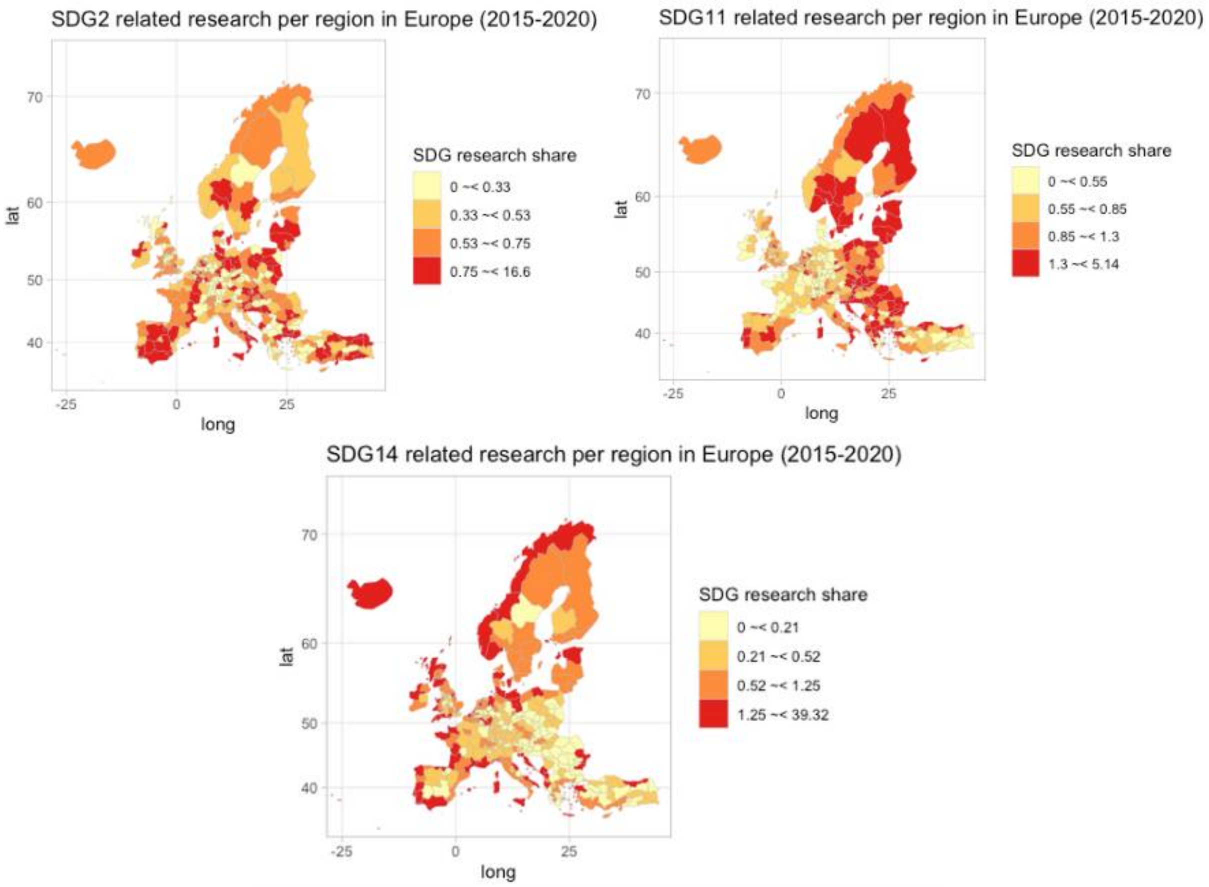
In addition, looking at the most represented subfield in all publications, could also give an explanation about why some SDGs are more researched than others (see Appendix G). Over both time periods it can be seen that life sciences together with physics and engineering belong to the most published subject categories. Biochemistry & Molecular Biology, Medicine, Oncology, and Neurosciences belong to the most represented subfields, and are all related to SDG3. Challenges related to communicable diseases, leading causes of death, and mental illnesses play a key role in our society and are at the heart of SDG3. Besides SDG3 plays a significant role in the achievement of many other SDGs (Pettigrew et al., 2015), which could explain the fact that SDG3 is by far the most represented. Another subfield that stands out is Environmental Sciences, that has become a frequently researched subfield in time period 2015-2020, showing an increasing interest in environmental problems, climate change (SDG13), and human impact on the environment.

Figure 10:
Overview of distribution SDGs (2015-2020)



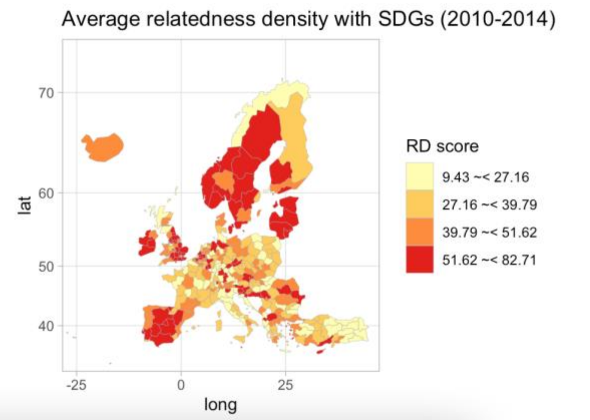
To further explore the regional distribution of SDG related research, the SDG research share for the individual SDGs was visualized for both time periods (Appendix H). Most SDGs showed a similar distribution as in figure 9, where the largest research shares were mostly concentrated in North-Western Europe and some Spanish regions. However, some maps showed a different distribution as is the case for SDG1 (Poverty), 2 (Hunger), 6 (Sanitation), 11 (Sustainable Cities and Communities), and 14 (Life below Water). SDG1 and 2, and SDG 6 and 11, showed a similar distribution, also the distribution per time period did not differ much, therefore only SDG 2, 11, and 14 are highlighted in figure 11 for the time period 2015-2020. For SDG1 and 2 a somewhat more skewed distribution across Europe can be seen, this indicates that the research done in these fields is not so much geographically bound and would thus be to a lesser extend dependent on the complexity and diversity of a region. Research on SDG 6 and 11 is somewhat more located in Northern and Eastern parts of Europe, indicating that there might be some underlying factors that influence knowledge production on these specific SDGs in these parts of Europe. Another explanation could be that Western European regions are less interested in these topics and address other scientific fields and SDGs more. Finally, SDG 14 shows that mainly coastal regions produce knowledge regarding this SDG. This is not surprising, as SDG14 focuses on Life below Water, and this is in practice more relevant to regions adjacent to bodies of water.

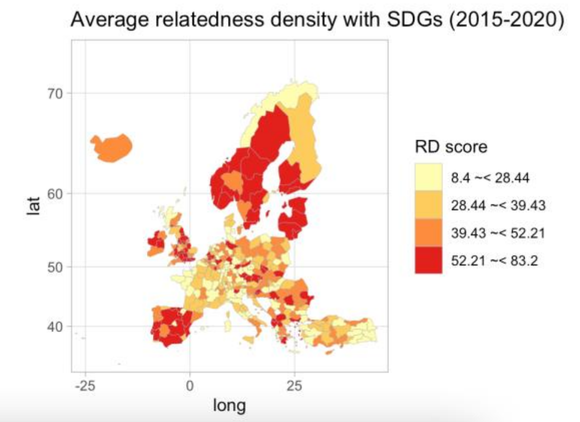
Figure 11:
Overview of SDG research share distribution for SDG 2, 11, and 14 (2015-2020)



In order for a region to be capable to diversify into SDGs it is important that they are related to each other to some extent, so that existing knowledge can be used to build upon the new knowledge related to an SDG. The relatedness density score of a region determines this and indicates the average relatedness of a region to any given SDG. The higher the score the easier it is for a region to diversify and produce knowledge on an SDG. Figure 12 shows the average relatedness score of regions to all the SDGs combined, which is similar to the average relatedness density score to the different subfields from figure 7. Regions closely related to the SDGs are on average located in Northern Europe and Spain. In order to get a deeper understanding of what this entails the relatedness density score of the regions were also visualized based on the individual SDGs (see Appendix I). From this it can be seen that most SDGs follow a similar distribution as in figure 12. However, SDG2 (Hunger) shows a rather skewed distribution across Europe, similar to the distribution of its research share. This is not surprising as it common for regions to produce knowledge on certain topics to which its existing knowledge base is related. Important to note is that for SDG4 (Education) no relatedness density score could be calculated for time period 2010-2014, as this SDG did not co-occur with any of the other SDGs. However, it showed a similar distribution as in figure 12 for period 2015-2020. Another surprising result is the fact that de relatedness density scores of SDG15 (Life on Land) are somewhat more skewed similar to the research share distribution of SDG14 (Life below Water). Indicating that coastal regions are somewhat more related to SDG15, while relatedness of SDG14 is more distributed as in figure 12. Finally, SDG16 (Peace, Justice, and Strong Institutions) showed a very low relatedness density score for almost all regions, indicating that not many regions in the period 2010-2014 had a related knowledge base. This did, however, increase over time as in the period hereafter the relatedness score approached a more similar distribution to figure 12.

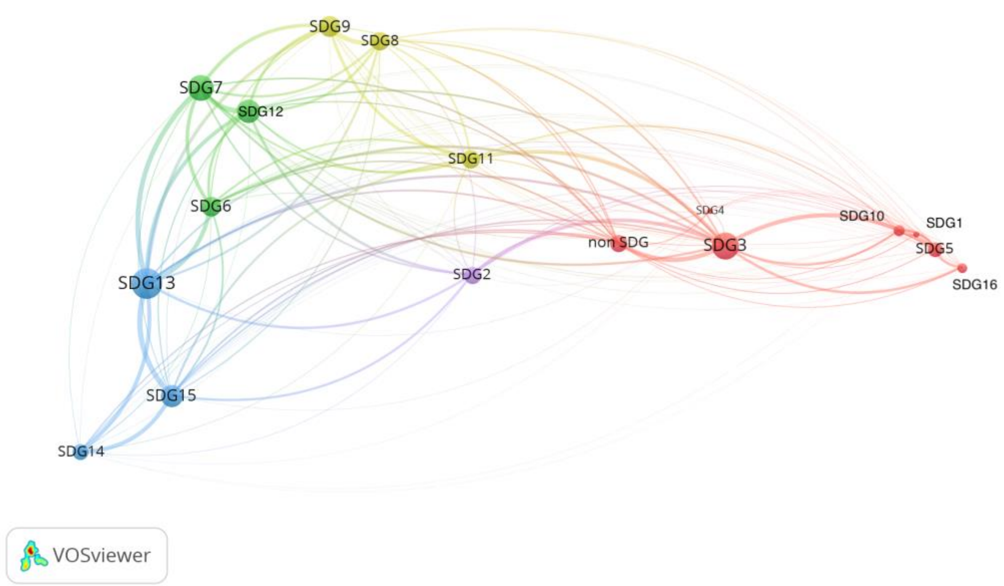
Figure 12:
Average relatedness density score with the combined SDGs





To gain deeper insights into how the SDGs co-occur with each other a scientific relatedness space was constructed (van Eck & Waltman, 2010), based on the co-occurrences of SDGs in the publications of different European regions. This field is visualized for time period 2015-2020 in figure 13. The dots represent the different SDGs, where the sizes give an indication of the number of co-occurrences. From here it can be deduced that SDG3 (Well-Being) and 13 (Climate Action), have the highest number of co-occurrences, as their dots appear the largest. Looking at the data used to construct the map, SDG13 co-occurs the most often with SDG15 (Life on Land), while SDG3 co-occurs, surprisingly, the most with non-SDG related research. Additionally, the closer the SDGs are located in the map, the more often they appear together. From the figure it can therefore be seen that SDG1 (Poverty), 5 (Equality), 10 (Reduced inequalities), and 16 (Peace, Justice, and Strong Institutions) often are attributed to the same publications. In contrast, SDG13, 14, and 15 are located at the opposite of the figure, indicating that they hardly to never co-occur with the SDGs on the other end. The scientific relatedness space for time period 2010-2014 shows a similar distribution (see Appendix K), however SDG4 and the non-SDG related research are not present in this map, as there were no co-occurrences with any of the other SDGs.

Figure 13:
SDG relatedness space showing relatedness across SDGs based on co-occurrences (2015-2020)



Finally, to analyse how the SDG research share evolved over time, the two time periods were plotted against each other. Figure 14 shows the results of this analysis, with on the x-axis the SDG research share of all combined SDGs over the period 2010-2014 and on the y-axis the SDG research share of all combined SDGs over the period 2015-2020. The reference line in the figure makes a distinguishment between regions of whom the SDG research share has increased over time (above the line) and regions that showed a decrease in SDG research share over time (below the line). In general, it can be concluded that most regions knew an increase in SDG research share, namely 96% of the regions. The regions that decreased in their research share, or at least did not improve it, are annotated in the figure, and involve regions from Austria, Greece, Turkey, the United Kingdom, Belgium, Bulgaria, and Finland. A similar pattern can be recognized for the individual SDGs, where most regions have known an increase in SDG research share (see Appendix J). Table 16 provides an overview of the percentage of regions that knew an increase in SDG related research. For almost all SDGs at least 74% of the regions has increased it SDG research share. Only for SDG 1 (Poverty) and SDG 10 (Reduced Inequalities) is the share of regions that decreased their research share, or remained equal, a bit higher, respectively 48% and 32%.

Figure 14:
SDG research share over both time periods

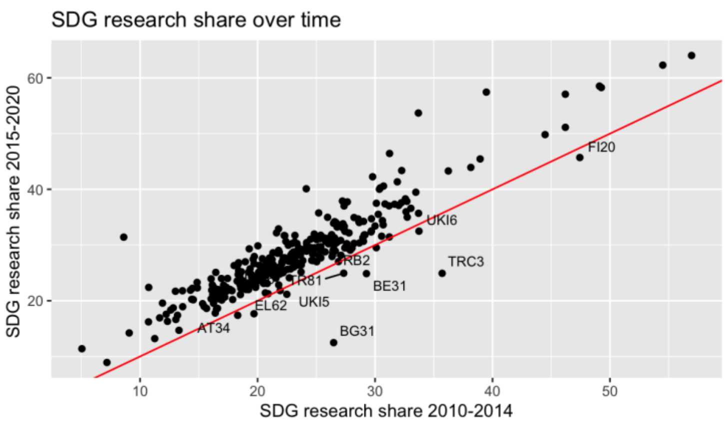


Table 16:
Percentage of regions that increased their research share over time

SDG	Increase	SDG	Increase
1	52%	**9**	86%
2	81%	**10**	68%
3	90%	**11**	83%
4	75%	**12**	89%
5	77%	**13**	89%
6	84%	**14**	74%
7	88%	**15**	80%
8	86%	**16**	77%

4.3 Regional Characteristics

This section provides more insights in the regional characteristics of the different regions, based on their country score. The scores of the indicators used to gain insights into regional characteristics were plotted against the average score of Europe, to see which countries performed worse relative to the other countries (see Appendix L). Those countries that performed worse than the average European country could then be said to experience local adverse effects, which might increase knowledge production within these areas. For indicator 1.1.1w, the proportion of the population that has access to basic water services, the minimum values are 92% (2010-2014) and 94% (2015-2020). For indicator 3.1.2, the proportion of births attended by skilled health personnel, the minimum values are 97% (2010-2014) and 95% (2015-2020). These values are considerably high, and therefore it is expected that no regions will experience significant local adverse effects for these indicators.

To give an impression of the data sample, table 17 gives an overview of the relatively lowest scoring countries per indicator. Frequently bad performing countries for both time periods are Albania, Montenegro, North-Macedonia, Romania, and Serbia. Geographically speaking these countries belong to the Balkans, except for Romania that only partly belongs to the Balkan. This area of Europe is known to face considerable challenges regarding sustainable and economic development (Sanfey & Milatovic, 2018), which could explain their performance. It is therefore, also expected, that the regions of these countries face the largest local adverse effects. Important to note is that within table 17 several countries have been displayed in bold. These countries' values are based upon imputed data for the missing values. Therefore, before the regression analysis is done an outlier analysis will be performed, which will be elaborated upon in the next section.

Table 17:
Relatively lowest scoring countries per SDG indicator

Indicator	2010-2014	2015-2020
SDG_1.1.1	Montenegro, North-Macedonia, Romania, Serbia	Montenegro, North-Macedonia, Romania, Serbia
SDG_1.4.1w	Albania, Croatia, Lithuania, Serbia	Albania, Croatia, Serbia
SDG_1.4.1s	Bulgaria, Romania	Bulgaria, Romania
SDG_2.1.1	Albania, Bulgaria, Cyprus, North-Macedonia, Slovakia	Albania, Bulgaria, Serbia, Slovakia
SDG_2.1.2		Albania, North-Macedonia, Romania
SDG_3.1.2	**Belgium**, Denmark, Turkey	Cyprus, Denmark, Romania
SDG_3.4.2	Hungary, Latvia, Lithuania	Latvia, Lithuania, Montenegro, Slovenia
SDG_7.2.1	Luxembourg, Malta, United Kingdom	Malta, Netherlands
SDG_10.2.1	Greece, North-Macedonia, Spain, Turkey	Greece, **Liechtenstein**, Montenegro, Romania
SDG_10.7.4	Albania, Croatia, Montenegro, North-Macedonia, Serbia, Turkey	Albania, Croatia, Serbia
SDG_13.1.1	**Belgium**, Switzerland	**Liechtenstein, Malta**, Sweden
SDG13.2.2	Estonia, Iceland, Luxembourg	Estonia, Iceland, Luxembourg
SDG_14.1.1		Croatia, Cyprus, Serbia
SDG_14.5.1	Croatia, Iceland, **Serbia**, Turkey	Iceland, Montenegro, Norway
SDG_15.1.2	Iceland, Montenegro, North-Macedonia, Serbia, Turkey	Iceland, Montenegro, North-Macedonia, Serbia, Turkey
SDG_15.5.1	Greece, Montenegro	Montenegro

Finally, in order to get a first impression of whether the regions from the countries that perform the worst overall do engage more in SDG related research, a deeper look was given into their research shares. Table 18 and 19 give an overview of the relative research share of each region for those SDG indicators in which they perform the worst as a country. The first columns give the NUTS_IDs of the regions and represent Albania (AL), Montenegro (ME), North Macedonia (MK), Romania (RO), and Serbia (RS). The second column *pub_tot* shows the number of publications that were attributed to the region over one of both time periods. The third column *SDG_rel* is the number of publications that could be distinguished as SDG related. Following from this, the fourth column *SDG_share* represents the SDG research share, based on the number of SDG related publications out of the total number of publications. The following columns represent the share of the SDG related research that was attributed to an SDG. For example, 0.24% of the SDG related publications of region AL02 were on SDG1. Overall, it can be seen that for almost all SDGs, with exception of SDG3 and 15, in which a region performs relatively worse, the share of research they attribute to this SDG is also quite low. As most publications are attributed to SDG3 (see Figure 10), it is not surprising to see that even the low-performing regions share a similar outcome. In addition, Montenegro attributes a considerably larger share of its publications to SDG15, compared to the other regions. The distribution of these regions is more in line with overall distribution of SDG related research. Thus, there are no clear signs that local adverse effects increase SDG related research in these areas.

Table 18:
Relative SDG share of worst performing regions (2010-2014)

NUTS_ID	pub_tot	SDG_rel	SDG_share	SDG1	SDG2	SDG10	SDG14	SDG15
AL01	19.75	1.79	9.07	0.00	0.00	0.00		
AL02	750.59	140.88	18.77	0.24	3.26	3.08		
AL03	11.89	1.02	8.61	0.00	0.00	0.00		
ME00	603.83	82.17	13.61	0.00		0.14		15.66
MK00	1546.05	224.63	14.53	0.00	1.45	0.19		4.85
RO11	7671.75	1406.95	18.34	0.00				
RO12	3610.83	712.10	19.72	0.11				
RO21	6916.08	1245.11	18.00	0.16				
RO22	1835.49	385.66	21.01	0.00				
RO31	1518.05	206.84	13.63	0.00				
RO32	17560.13	2910.26	16.57	0.23				
RO41	1728.91	284.25	16.44	0.00				
RO42	4806.04	980.69	20.41	0.00				
RS11	13635.01	2864.65	21.01	0.09		0.62	0.23	6.31
RS12	4157.91	958.66	23.06	0.26		0.64	0.16	6.69
RS21	1452.18	385.70	26.56	0.00		0.00	0.50	4.47
RS22	2786.32	577.67	20.73	0.00		0.36	0.07	2.48

Table 19:
Relative SDG share of worst performing regions (2010-2014)

NUTS_ID	pub_tot	SDG_rel	SDG_share	SDG1	SDG2	SDG3	SDG10	SDG14	SDG15
AL01	48.31	6.87	14.22	0	0		0		
AL02	932.34	215.22	23.08	1.00	3.19		2.01		
AL03	17.99	5.65	31.41	0	1.97		0		
ME00	1249.06	272.16	21.79	0	0	45.86	0.80	8.64	14.90
MK00	1973.84	398.18	20.17	0.25	1.31				4.97
RO11	10181.03	2700.45	26.52	0.01	2.03		0.33		
RO12	4240.03	1184.75	27.94	0.17	2.18		0.55		
RO21	8004.93	2106.28	26.31	0.05	1.52		0.29		
RO22	2291.94	604.61	26.38	0	1.85		0		
RO31	1626.11	307.75	18.93	0.16	4.22		0.65		
RO32	23734.34	5225.49	22.02	0.21	2.73		0.99		
RO41	2123.38	532.70	25.09	0.42	2.13		0.23		
RO42	5638.35	1553.13	27.55	0	1.35		0.23		
RS11	16248.91	4495.36	27.67	0.06	2.45	70.24	0.52	0.42	6.60
RS12	5397.69	1551.87	28.75	0	5.13	54.51	0.20	0.44	6.72
RS21	1686.90	576.73	34.19	0	1.64	76.97	0.97	0.92	4.53
RS22	2899.16	778.49	26.85	0	1.18	62.30	1.19	0.13	4.00

4.4 Outlier analysis

Before running a regression analysis, it was important to detect any outliers in the data, as they could distort the model and lead to misleading results (Manimannan et al., 2020). Outliers are values that are distant to the other values and can be seen as abnormal, therefore they should be analyzed and checked whether they can be considered extreme values or mistakes (Soetewey, 2020). A first step in checking for outliers was to review the descriptive statistics of the different variables (see Appendix M). At first sight, for the dependent and control variables it did not look like there are outliers present, as well as for the knowledge complexity index (KCI) and relatedness density (RD) scores. However, for the regional characteristics, for several variables (SDG_10.7.4 and SDG_13.1.1) it looked like the maximum score deviated significantly compared to the minimum score, mean and median. To verify this first impression, boxplots were created for the different variables to check for any outliers (see Appendix N). These showed quite a different image as all the dependent variables had outliers, as well as the control variables. Besides, most variables of the regional characteristics, apart from SDG3.1.2, 10.2.1, and 15.5.1 for both time periods, showed outliers. However, for the independent variables, the KCI scores, and most RD scores (except for the RD score of SDG2, 4, and 5 in time period 2015-2020) showed no outliers.

This number of outliers does not come as a surprise as we saw earlier from the skewness coefficients and the PP-plots that the data is not normally distributed. The number of outliers for the SDG research share ranged from 8 outliers (SDG3) to 37 outliers (SDG14) over both time periods (see Appendix N). For the control variables, Education and GDP had some outliers, respectively 6 and 2 for period 2010-2014 and 11 and 1 for period 2015-2020. Population, however, did have more outliers, 20 and 19 for both time periods respectively. Therefore, as mentioned earlier, data on the SDG research share and population was logarithmically transformed in order to approach a more normal distribution. This considerably lowered the outliers, resulting for the period 2010-2014 in no outliers for the research share of SDG1 and 16, the range of outliers between 1 (SDG5) and 21 (SDG8), and for the population variable there was just a single outlier. For period 2015-2020, this resulted in no outliers for the research share of SDG 1 and 10, the range of outliers between 1(SDG5 and 16) and 22 (SDG8), and just a single outlier for the population control variable. It was unlikely that the outliers of several regions that still existed could be attributed to a measurement error. In case of the SDG research share variables, this could be due to a measurement error, by using the STRINGs thesaurus and using the WoS as a database, thereby excluding non-English publications and underrepresenting scientific fields that cover social sciences and humanities ((Mingers & Leydesdorff, 2015). Nevertheless, this research kept the outliers of the SDG research share, RD scores, Education, and the population, as they were real values (non-imputed) and treated them as extreme cases. Therefore, they were not be excluded from the data sample.

In case of the other variables, GDP and the regional characteristics, the outliers were checked in order to make sure the outliers were not caused by imputed data. For the regional characteristic variables, SDG_1.4.1 w/s, SDG_7.2.1, SDG_15.1.2, and SDG_15.5.1 did not contain any missing values, therefore the outliers could not be caused by imputed data. The variables thus showed legitimate observations and were therefore not excluded from the research. Additionally, table 20 gives an overview of the variables that had outliers which came from imputed data. This means that the other regional characteristic variables had no outliers that could be attributed to imputed data, and therefore were included in the sample as special cases. From table 20 it can be seen that a total of 15 observations caused outliers in the variables due to a possible measurement error in imputing the data. For variable SDG_10.7.4 (number of refugees) the outlier was attributed to Luxembourg, although no data was available for this period, data was available for the period 2015-2020. Therefore, as there was no considerable increase in the refugee population of Luxembourg (the World Bank, n.d.), the outlier was corrected for with the same value as for the period 2015-2020. For variable SDG_13.1.1 (deaths due to disasters) period 2010-2014, 9 out of 11 outliers were attributed to Belgium. In order to correct for these outliers, additional data was collected from the Belgium Federal Planning Bureau (2021) on natural disaster victims, which is in line with indicator 13.1.1. of the SDGs. This data was used to correct for the outliers and imputed in the data sample, as these were considered to be legitimate observations. The remaining outliers of the first time period were attributed to Lithuania, and from period 2015-2020 to Iceland, Liechtenstein, and Malta. In order to correct for these outliers and the outliers in the GDP the mode value was used, that is the most frequently occurring value in the variable dataset. This method was deemed appropriate as it decreases the influence the outlier has on the regression analysis, without excluding it from the sample.

Table 20:
Overview outliers caused by imputed data

2010-2014		2015-2020	
Variable	**Outliers**	**Variable**	**Outliers**
GDP	2 of 6	**GDP**	4 of 11
SDG_10.7.4	1 of 66		
SDG_13.1.1	11 of 25	**SDG_13.1.1**	3 of 24

4.5 Regression analysis

This section elaborates on the results of the regression analyses. The analyses were conducted for all SDGs in order to find an explanation for differences between regions in knowledge production regarding 16 SDGs. The analyses covered two non-overlapping time periods, 2010-2014 before the SDGs were introduced and 2015-2020 after the SDGs were introduced, to incorporate the different stages of knowledge production. As some variables were logarithmically transformed, such as the SDG share, Population, and some variables in the regional characteristics, they need to be interpreted differently. Especially since the dependent variable has been transformed the regression models no longer show the absolute change in an independent variable compared to the absolute change in SDG research share. Consequently, with a logarithmically transformed variable, the absolute variation in the logarithm equals the relative variation in the original variable (Rodríguez-Barranco et al., 2017). In the regression models the statistical significance is indicated by asterisks. As a rule of thumb, a *p* value under 0.05 can be regarded as an indicator of statistical significance.

4.5.1 Regression models 1 and 2

First, the relationship between the SDG research share with the knowledge complexity and relatedness density of regions was tested in Model 1 and 2, of which the results are listed in table 21 and 22. Starting with the KCI score it can be seen that it has a positive relationship with the SDG research share, except for SDG6 and SDG11. Besides, the coefficient of the KCI score is significantly positive for all SDGs in the analysis, except for SDG8 and 12. This suggests for the positive relationships, that the larger the knowledge complexity of a region is i.e., its capabilities to develop complex knowledge, the more a region is capable to develop complex knowledge regarding the SDGs expressed in SDG research share. In addition, the coefficient of SDG6 is significantly negative in model 1 and in model 2 the coefficient for knowledge complexity of both SDG6 and 11 are significantly negative. The increased significance for SDG11 might be due to an increase in the SDG related research for this SDG. These findings indicate, in general, that the complexity of a region's knowledge base is a driver for the knowledge production on SDGs (Heimeriks et al., 2019). Therefore, the assumption that regions with a more complex knowledge base are more likely to engage in SDG related research is valid in the case of the combined SDGs, SDG4, 5, 9, and 16. This means that hypothesis 1 can be accepted for these SDGs, for SDG6 in model 1 and SDG 6 and 11 in model 2, the hypothesis is rejected as these imply a negative relationship. For SDG9, 11, and 12 in model 1 and SDG8 and 12 in model 2, no statistically significant relationship could be found, so hypothesis 1 can't be accepted or rejected for these SDGs.

Second, for the relatedness density scores between regions and SDGs it can be seen that all results are statistically significant. Besides, only the relatedness density score for the combined SDGs shows a negative relationship. This might be due to the fact that SDG3 also has a significantly negative relationship between relatedness density and research share. As seen earlier, SDG3 contributes for approximately 60% to all the publications that are related to an SDG. Therefore, the results of the combined SDGs will be very dependent on the results of SDG3, upon which will be further elaborated in section 4.5.2. For models 2 up to and including model 9 in table 21 and 22, it can be said that the more related the knowledge base of a region is to an SDG, the more capabilities a region has to produce complex knowledge regarding this SDG. These findings indicate that besides knowledge complexity, scientific relatedness is also a driver for knowledge production on the SDGs and they confirm that regions tend to produce knowledge regarding the SDGs that are related to their existing knowledge base (Balland & Boschma, 2021; Li, 2020). Consequently, the initial hypotheses H2 can be entirely accepted for SDG4, 5, 6, 8, 9, 11, 12 and 16 and is rejected for the SDG research share of the combined SDGs.

Overall, for the two models surprising is to see that GDP has a regression coefficient of approximately zero in most of the cases, indicating that is does not have an effect on the SDG research share. This stands in contrast with earlier research, that indicated that the economic development is correlated with complex knowledge production (Hidalgo & Hausmann, 2009). As the SDGs aim to incorporate all worldwide pressing issues, relevant to all regions, this might explain why GDP is of less importance to SDG related research. As seen earlier there are many regions with a very low publication output, however, the publications they produce are related to SDGs giving them a relatively high SDG research share. This might mediate the effects of economic development on the SDG research share. Looking at the population, SDG16 has the strongest statistically significant positive relationship. Besides, SDG5 and 6 also shows a significant positive relationship with population. In contrary the combined SDGs and SDG 8 showed a significant negative relationship with population. Finally, education shows almost all negative relationships, from which only that of SDG4 is statistically significant. This is surprising as it was expected that tertiary education was of key importance for sustainable development (Salmi, 2017). A possible explanation could be that as the SDGs involve all layers of society, educational level does not affect research in these areas as much, as it is crucial for all regions to invest in knowledge production on these topics. Consequently, for regions with a lower educational level a relatively larger share of the population with tertiary education will be steered towards SDG research.

Finally, in terms of explanatory power, the R^2 and the adjusted R^2 give an indication of the number of data points that fall within the line of the regression equation. Both values are given as a percentage and the difference between the two is that the R^2 assumes all variables affect the dependent variable, while the adjusted R^2 takes the number of variables into consideration and corrects for those variables that do not affect the dependent variable (Harel, 2009). The closer the value is to 1, the more perfectly the fit of the model. From this it can be seen that SDG5 and SDG16 have the best fit, which indicates that the SDG research share of these SDGs is most strongly shaped by the independent variables. On the contrary, SDG11 has the worst fit of both models with a value below 0.1, although this increased over time. This indicates that for SDG11 the data is the most scattered and there are large differences between observations. Consequently, a general trend within the data is captured, but many data points fall around the trendline instead of following it. Besides, the lower the R^2 and adjusted R^2, the less well the explained the SDG research share is by the chosen variables, indicating that other variables might explain the dependent variable better.

Table 21:
Regression results model 1

	Dependent variable:								
	SDG research share (log)								
	all SDGs	SDG4	SDG5	SDG6	SDG8	SDG9	SDG11	SDG12	SDG16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KCI score	0.002***	0.009***	0.006***	-0.004***	0.001	0.004***	-0.0001	0.0004	0.013***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RD regions-SDG	-0.001***		0.003***	0.006***	0.005***	0.005***	0.004***	0.006***	0.003***
	(0.0004)		(0.001)	(0.001)	(0.0005)	(0.0004)	(0.001)	(0.001)	(0.001)
Population (log)	-0.025***	0.047*	0.051***	0.079***	-0.048***	-0.008	0.001	0.025	0.231***
	(0.008)	(0.025)	(0.015)	(0.021)	(0.018)	(0.017)	(0.019)	(0.017)	(0.027)
GDP	-0.00000	-0.00001**	0.00000	-0.00000**	0.00000***	0.00000**	-0.00000	-0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	-0.001	-0.006**	-0.001	0.001	-0.003	-0.002	-0.001	-0.0004	0.002
	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Constant	1.718***	-1.115***	-1.500***	-1.187***	0.185	-0.528**	-0.234	-0.834***	-5.028***
	(0.116)	(0.361)	(0.213)	(0.304)	(0.261)	(0.247)	(0.272)	(0.248)	(0.388)
Observations	333	333	333	333	333	333	333	333	333
R^2	0.150	0.148	0.465	0.293	0.272	0.359	0.073	0.310	0.516
Adjusted R^2	0.137	0.138	0.457	0.282	0.261	0.349	0.058	0.299	0.508
Residual Std. Error	0.119 (df = 327)	0.369 (df = 328)	0.208 (df = 327)	0.307 (df = 327)	0.267 (df = 327)	0.253 (df = 327)	0.279 (df = 327)	0.253 (df = 327)	0.389 (df = 327)
F Statistic	11.567*** (df = 5; 327)	14.277*** (df = 4; 328)	56.884*** (df = 5; 327)	27.045*** (df = 5; 327)	24.462*** (df = 5; 327)	36.637*** (df = 5; 327)	5.121*** (df = 5; 327)	29.362*** (df = 5; 327)	69.672*** (df = 5; 327)
Note:									*p<0.1; **p<0.05; ***p<0.01

***Table 22**:*
Regression results model 2

	Dependent variable:								
	SDG research share (log)								
	all SDGs	SDG4	SDG5	SDG6	SDG8	SDG9	SDG11	SDG12	SDG16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KCI score	0.001***	0.005***	0.006***	-0.004***	0.001	0.003***	-0.001**	-0.001	0.010***
	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.001)	(0.001)
RD regions-SDG	-0.001**	0.005***	0.003***	0.005***	0.006***	0.006***	0.003***	0.007***	0.003***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0005)	(0.0005)	(0.001)	(0.001)
Population (log)	-0.027***	-0.056**	0.018	0.022	-0.051***	0.003	-0.006	0.016	0.164***
	(0.007)	(0.022)	(0.014)	(0.019)	(0.017)	(0.018)	(0.009)	(0.016)	(0.026)
GDP	-0.00000**	-0.00000**	-0.00000	-0.00000***	-0.00000	0.00000	-0.00000**	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	-0.001	-0.004*	0.0003	0.002	0.001	-0.001	0.001	-0.001	0.003
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)
Constant	1.860***	0.373	-0.904***	-0.260	0.388	-0.492*	0.305**	-0.576**	-3.793***
	(0.102)	(0.309)	(0.203)	(0.273)	(0.246)	(0.257)	(0.130)	(0.232)	(0.374)
Observations	333	333	333	333	333	333	333	333	333
R^2	0.137	0.252	0.480	0.264	0.312	0.354	0.160	0.364	0.459
Adjusted R^2	0.124	0.241	0.472	0.253	0.301	0.345	0.147	0.355	0.451
Residual Std. Error (df = 327)	0.104	0.313	0.201	0.278	0.252	0.261	0.133	0.236	0.382
F Statistic (df = 5; 327)	10.366***	22.090***	60.402***	23.482***	29.642***	35.908***	12.488***	37.503***	55.439***
Note:									*p<0.1; **p<0.05; ***p<0.01

4.5.2 Regression models 3 and 4

The relationship between the SDG research share with the knowledge complexity, relatedness density, and regional characteristic was tested for in Model 3 and 4 for a limited number of SDGs. The results of this analysis are listed in table 23 and 24. Starting with the KCI score it can be seen that it has, for most SDGs, a positive relationship with the SDG research share. Exceptions are SDG7 and 15 from model 3 and SDG2 and 7 from model 4, these regression coefficients are, however, not statistically significant. Besides, the coefficient of the KCI score is significantly positive for SDG1, 3, 10, and 14 in model 3, where SDG 13 also became significantly positive in model 4. This increase in statistical significance might be due to an increase in research on the SDG. These findings indicate, in general, that the complexity of a region's knowledge base is a driver for the knowledge production on SDGs (Heimeriks et al., 2019). Therefore, the assumption that regions with a more complex knowledge base are more likely to engage in SDG related research is valid in the case of SDG1, 3, 10, 13 (model 4), and 14. This means that hypothesis 1 can be accepted for these cases. For SDG2, 7, 13 and 15 in model 3 and SDG 2,7, and 15 in model 4 no statistically significant relationship could be found, so hypothesis 1 can't be accepted or rejected for these SDGs. Overall, this means that for the period 2010-2014 8 out of 16 SDGs proved to have a statistically significant positive relationship between the KCI score and the SDG research share, while for period 2015-2020 this was the case for 9 out of 16 SDGs.

Second, for the relatedness density scores between regions and SDGs it can be seen that almost all results are statistically significant in both models. However, as indicated earlier, only SDG3 has a significantly negative relationship between the relatedness density and research share of this SDG. This indicates that the relatedness density of a region to SDG3 does not affect the SDG related research share of this region on SDG3. This might be due to the fact that the knowledge that is required for SDG3 is less complex, which makes it easier to diversify to this SDG. Another explanation could be that due to the large share of SDG related publications that could be attributed to SDG3 it is easier for regions to diversify into this SDG as more knowledge is available. In addition, earlier the research showed that there is already a lot of existing knowledge for SDG3 as shown in the most represented subcategory fields in the publications of the WoS (see Appendix G). For model 1 up to and including model 8, except for model 3, in table 23 and 24, it can thus be said that the more related the knowledge base of a region is to an SDG, the more capabilities a region has to produce complex knowledge regarding the SDGs. These findings indicate that besides knowledge complexity, scientific relatedness is also a driver for knowledge production on the SDGs and confirm that regions tend to produce knowledge regarding the SDGs that are related to their existing knowledge base (Balland & Boschma, 2021; Li, 2020). Consequently, the initial hypotheses H2 can be entirely accepted for SDG1, 2, 7, 10, 13, 14, and 15 and is rejected for SDG3. Overall, this means that for both time periods, 15 out of 16 SDGs proved to have a statistically significant positive relationship between the relatedness density score between a region and an SDG and the SDG research share.

In addition, for the regional characteristics of the different SDGs and the SDG research share it can be seen that for most selected SDGs the regional characteristics have no significant influence on the SDG research share. In model 3, the regional characteristics that are statistically significant, namely characteristic two for SDG1, both characteristics for SDG3, and characteristic one for SDG10, show a negative relationship with the SDG research share. However, the negative relationship of SDG1 should be reversed, as was discussed in the methodology due to left-skewed data. There is thus a positive relationship between the second regional characteristics of SDG1 and the SDG research share. Another important aspect to consider is the indicator the variable represents. Hypothesis 3 can only be accepted if a region has a high SDG research share, while experiencing local adverse effects. First, the second characteristic of SDG1, the proportion of the population with access to basic sanitation services, shows a positive relationship, indicating that the SDG share increases as the access to sanitation increases. This is not in line with the hypothesis as the SDG research share is not high when access to sanitation is limited. Therefore, the hypothesis is rejected for this characteristic of the SDG.

Second, the first characteristic of SDG3 represents the proportion of births attended by skilled health personnel. This shows a negative relationship to the SDG but is in line with hypothesis 3 as research share decreases if the attendance of skilled personnel increases. The hypothesis is thus accepted for this variable. For the other two indicators, the Suicide Mortality Rate (3.4.2) and the Proportion of people living below 50 per cent of median income (10.2.1), the negative relationship does have negative consequences as the SDG research share decreases when these indicators increase and therefore the hypothesis is rejected for these variables. This is surprising as initially it was thought that regions that would experience local adverse effects would increase research in these areas to overcome these effects. A potential explanation for this is that societal demands only explain a regions research trajectory to a limited extend and research has shown that there is often a misalignment between the priorities and needs of regions (Ciarli & Ràfols, 2019).

For model 4, both regional characteristic two for SDG1 and one for SDG3 lost their statistical significance. This might be due to an improvement in local circumstances. In contrary to model 3, only the second characteristic of SDG3 and SDG10 show a significant negative relationship to the SDG research share, while SDG7 and SDG14 increased significance and now show a positive relationship to SDG research share. However, again, for SDG14 the second regional characteristic should be reversed due to its left skew, indicating that this shows a negative relationship to the SDG research share. Also, the indicators should be considered again. The negative relationships of SDG3 and the first characteristic of SDG10 are not in line with hypothesis 3, as earlier discussed. Additionally, the second characteristic of SDG10, namely the number of refugees, also shows a negative relationship, meaning that SDG related research decreases when the number of refugees increases. Consequently, hypothesis 3 is rejected for this variable as the adverse effects do not lead to an increase in research share. Following, the now negative relationship of the second characteristic of SDG14 can be considered to have a positive effect as SDG related research only decreases if protected marine key biodiversity areas increase. Hypothesis 3 is thus accepted for this variable. For the positive relationship of SDG7, the proportion of people with access to electricity, the hypothesis is rejected as the SDG research share is not high when access to electricity is limited. Finally, for the first characteristic of SDG14, beach litter, the hypothesis is accepted since the SDG research share is higher when a region experiences more beach litter. Overall hypotheses 3 could thus only be accepted for the first characteristic of SDG3 in model 3, and for both characteristics of SDG14 in model 4. Thus, for these SDGs it can be said that regions engage more in SDG related research if they experience relatively more adverse regional effects related to these SDGs. However, for the other statistically significant variables the hypothesis did not hold and was rejected. Besides, for the SDGs that showed no statistical significance the hypothesis could not be rejected nor accepted.

Overall, the population showed to have a positive significant effect on SDG1, 7, 10 and 14 in model 3. Confirming that complex knowledge production concentrates more in larger cities (Balland et al., 2020; Boschma et al., 2014). In contrary, model 4 shows that the population is only significantly positively related to the SDG research share for SDG 1 and 10, while a significant negative relationship exists between the population and the research share of SDG13. This might be due to the adoption of the SDGs, as this also provided smaller regions with guidelines on where to steer research (Sachs, 2012), making population a less important driver of SDG related publications. GDP shows the same surprising results as in the previous two models, with only very small regression coefficients and several coefficients that were zero. In model 3, SDG 1 and 10 show a somewhat significant positive relationship to the GDP, while in both models SDG2 shows a significant negative relationship. In model 4, SDG1 no longer showed a positive significant relationship, while SDG10 and 13 did to some extent. This indicates that GDP might not be a considerable driver of research related to the SDGs. This is in line with the findings that Outermost Regions, who often have a low economic output (European Commission, 2017), still have a high SDG research share. For education no significant relationships could be discovered in model 3. For model 4 education only showed a significant positive relationship for SDG1.

Therefore, it can be assumed that education is also not an important driver for SDG related research. Finally, looking at the R^2 and adjusted R^2, SDG1 and SDG7 show to be best fitted to the regression equation. Meaning that their SDG research share is most strongly shaped by the independent variables compared to the other SDGs. On the contrary SDG2 shows the worst fit for both model 3 and 4, with less than 10% of the datapoints falling on the trendline, indicating that other variables might explain this SDG better.

Table 23:

Results regression analysis model 3

	Dependent variable:							
	SDG research share (log)							
	SDG1	SDG2	SDG3	SDG7	SDG10	SDG13	SDG14	SDG15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KCI score	0.006**	0.002*	0.002***	-0.002*	0.004***	0.001	0.008***	-0.002*
	(0.002)	(0.001)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
RD regions-SDG	0.004***	0.006***	-0.001***	0.006***	0.002***	0.007***	0.006***	0.007***
	(0.001)	(0.001)	(0.0002)	(0.001)	(0.0004)	(0.001)	(0.001)	(0.001)
Regional characteristic 1	0.027		-1.854***		-0.793**	-0.029		-0.008
	(0.089)		(0.495)		(0.381)	(0.044)		(0.070)
Regional characteristic 2	-0.225***		-0.002**	0.051	-0.006	-0.010	0.960	0.655
	(0.066)		(0.001)	(0.065)	(0.012)	(0.008)	(0.692)	(0.468)
Population (log)	0.278***	0.009	0.007	0.053***	0.036***	-0.026	0.108**	-0.042*
	(0.045)	(0.020)	(0.006)	(0.020)	(0.013)	(0.021)	(0.043)	(0.024)
GDP	0.00002***	-0.00001***	-0.00000	-0.00000	0.00000*	0.00001***	-0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	0.005	0.002	0.001	0.0002	0.0005	0.001	0.004	0.001
	(0.005)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.005)	(0.003)
Constant	-7.300***	-0.712**	3.119***	-0.806***	-1.392***	-0.087	-2.697***	-0.075
	(0.694)	(0.289)	(0.506)	(0.284)	(0.194)	(0.325)	(0.631)	(0.512)
Observations	333	333	333	333	333	333	333	333
R^2	0.359	0.081	0.233	0.289	0.456	0.247	0.197	0.242
Adjusted R^2	0.345	0.067	0.217	0.276	0.444	0.230	0.182	0.226
Residual Std. Error	0.652 (df = 325)	0.295 (df = 327)	0.086 (df = 325)	0.289 (df = 326)	0.190 (df = 325)	0.314 (df = 325)	0.627 (df = 326)	0.346 (df = 325)
F Statistic	25.972*** (df = 7; 325)	5.780*** (df = 5; 327)	14.136*** (df = 7; 325)	22.090*** (df = 6; 326)	38.892*** (df = 7; 325)	15.196*** (df = 7; 325)	13.315*** (df = 6; 326)	14.823*** (df = 7; 325)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$							
	Some variables in the Regional Characteristics were log or sqrt transformed, see table 6							

Table 24:

Results regression analysis model 4

	Dependent variable:							
	SDG research share (log)							
	SDG1	SDG2	SDG3	SDG7	SDG10	SDG13	SDG14	SDG15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KCI score	0.002**	-0.001	0.001***	-0.001*	0.004***	0.002**	0.008***	0.0001
	(0.001)	(0.001)	(0.0003)	(0.0005)	(0.001)	(0.001)	(0.002)	(0.001)
RD regions-SDG	0.004***	0.006***	-0.001***	0.004***	0.002***	0.007***	0.006***	0.008***
	(0.001)	(0.001)	(0.0002)	(0.0004)	(0.0004)	(0.001)	(0.001)	(0.001)
Regional characteristic 1	0.026		0.013		-0.850**	-0.016*	0.208***	-0.039
	(0.033)		(0.022)		(0.361)	(0.009)	(0.065)	(0.064)
Regional characteristic 2	-0.051*	-0.177*	-0.003**	0.112***	-0.026**	0.007	3.530***	0.306
	(0.030)	(0.097)	(0.001)	(0.043)	(0.011)	(0.008)	(0.900)	(0.396)
Population (log)	0.140***	0.004	0.010	0.012	0.045***	-0.041**	0.056	-0.032
	(0.022)	(0.020)	(0.006)	(0.011)	(0.013)	(0.019)	(0.039)	(0.022)
GDP	0.00000	-0.00001***	-0.00000*	0.00000	0.00000***	0.00001***	-0.00000	-0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	0.005**	0.004*	0.001*	0.001	0.0001	-0.003	0.005	-0.003
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)	(0.002)
Constant	-3.985***	-0.614**	1.331***	0.224	-1.409***	0.289	-3.313***	0.228
	(0.325)	(0.281)	(0.107)	(0.159)	(0.191)	(0.290)	(0.707)	(0.455)
Observations	333	333	333	333	333	333	333	333
R^2	0.363	0.098	0.140	0.298	0.467	0.273	0.233	0.280
Adjusted R^2	0.349	0.081	0.122	0.285	0.456	0.257	0.217	0.264
Residual Std. Error	0.316 (df = 325)	0.281 (df = 326)	0.090 (df = 325)	0.162 (df = 326)	0.187 (df = 325)	0.280 (df = 325)	0.564 (df = 325)	0.321 (df = 325)
F Statistic	26.455*** (df = 7; 325)	5.874*** (df = 6; 326)	7.570*** (df = 7; 325)	23.079*** (df = 6; 326)	40.712*** (df = 7; 325)	17.416*** (df = 7; 325)	14.124*** (df = 7; 325)	18.047*** (df = 7; 325)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$							
	Some variables in the Regional Characteristics were log or sqrt transformed, see table 7							

Finally, a broader look can be given at the regional patterns of complex knowledge among the different SDGs. This is done by taking a transformative lens perspective, as discussed earlier in the theory section. The adjusted R^2 of all the SDGs for both time periods are compared to give better insights into which SDGs are best explained by the regional patterns of complex knowledge. An overview of the adjusted R^2 scores is given in table 25, where the type of SDG is indicated between brackets. From the transformative lens perspective, it can be said that the nature of the goals is not equal, some SDGs are more technical, while others are more socially related. This influences the research that is done on these topics. Expected was that SDGs that belong to the 'Socio-Technical Systems and Application Areas' (ST) domain are more susceptible to scientific research and innovation and therefore more influenced by different patterns of complex knowledge production. However, table 25 surprisingly shows that SDG16, the Framework Conditions type (FC), is explained very well by the independent variables. Besides, looking at the averages of the explained variance for the different types of SDGs it shows that for period 2010-2014 and 2015-2020, the adjusted R^2 for Transversal Direction (TD) SDGs is respectively 0.300 and 0.324 and for ST SDGs 0.216 and 0.234. This shows, that in contrary to what was expected, ST SDGs are the least explained by the independent variables. No considerable explanation could be found on why ST SDGs are the least explained by the independent variables. Nevertheless, important to keep in mind is that North-Western regions in Europe often have larger funds attributed to research (Mićić, 2017), thereby leaving more room for other research fields such as the social sciences. This might influence the complex knowledge production on FC and TD SDGs and in doing so also the regional patterns of complex knowledge production. Besides, as mentioned earlier, social sciences and humanities are considered the most complex scientific subfields, thereby increasing the complexity score of regions that research these subfields. Consequently, SDGs related to these scientific subfields are probably more affected by the knowledge complexity variable. Overall, it can thus be said that research on Socio-Technical SDGs is less affected by complexity, scientific relatedness, and regional characteristics than research on other SDGs. Therefore, hypothesis 4 is rejected.

Table 25:
Overview of the Adjusted R^2 for both time periods and all the SDGs

2010-2014				**2015-2020**			
SDG	**Adj. R^2**	**SDG**	**Adj. R^2**	**SDG**	**Adj. R^2**	**SDG**	**Adj. R^2**
1 (TD)	0.345	**9 (ST)**	0.349	**1 (TD)**	0.349	**9 (ST)**	0.345
2 (TD)	0.067	**10 (TD)**	0.444	**2 (TD)**	0.081	**10 (TD)**	0.456
3 (ST)	0.217	**11 (ST)**	0.058	**3 (ST)**	0.122	**11 (ST)**	0.147
4 (ST)	0.138	**12 (TD)**	0.299	**4 (ST)**	0.241	**12 (TD)**	0.355
5 (TD)	0.457	**13 (TD)**	0.230	**5 (TD)**	0.472	**13 (TD)**	0.257
6 (ST)	0.282	**14 (ST)**	0.182	**6 (ST)**	0.253	**14 (ST)**	0.217
7 (ST)	0.276	**15 (ST)**	0.226	**7 (ST)**	0.285	**15 (ST)**	0.264
8 (TD)	0.261	**16 (FC)**	0.508	**8 (TD)**	0.301	**16 (FC)**	0.451

4.5.3 Robustness check

To increase the robustness of the results, several additional regression models were estimated. On the one hand, the same regression models were estimated but without the imputed data (hereafter named; control model). Therefore, four control models were created, based on the original data with missing values in several regional characteristics, GDP, and education. An overview of the regression results can be found in Appendix O. On the other hand, extra robustness checks were done on the database with imputed data, to check the influence of the control variables and the regional characteristic variables on the regression models. Therefore, the SDG research share of all SDGs was estimated only for the control variables, to see how they explain the variance in the SDG share. Besides for model 3 and 4 additional regression analyses were performed without the regional characteristics, to see how this influences the variance (see Appendix P for the regression models). This was done in order to make sure the acceptance and rejection of the hypotheses are not influenced by the imputation of data or by the control variables alone.

Regression models without imputed data

First, the regression results of model 1. Looking at the KCI score, overall, the regression coefficients of both models are in the same order of magnitude. Also, de significance remained the same for most SDGs. However, in the control model SDG6 lost the significant negative relationship to the SDG research share, while it did show a significant positive relationship to the research share of SDG8. This was not observed in the original model. No significant changes were found for the relatedness density score of regions to the SDGs. All regression coefficients are roughly the same, and significance remained for all SDGs in the model. Looking at the control variables it could be said that population in the control model has a slightly larger impact on the SDG research share than in the original model. This might be due to the fact that there are fewer observations, 277 compared to 333 in the full model. In addition, the statistical significance of the model stayed roughly the same, in exception the combined SDGs, for which the population was no longer significant in the control model. For Education and GDP no remarkable differences existed between the models. The regression coefficient of GDP is in both models approximately zero and for Education in both models no statistically significant relationship could be identified. Finally, the model fits are roughly the same, looking at the overall explained variance it can be seen that the full model fits slightly better than the control model.

Second, the comparison between the control and full model of model 2 from the regression analysis. Looking at the KCI score, overall, the regression results are similar to that of the first model. The regression coefficients and significance are roughly the same. However, again, in the control model SDG5 lost its significance, while the KCI variable for SDG8 became significant. Comparing the regression coefficients of the relatedness density variable, two things stand out. First, the relatedness density of the combined SDGs lost its significance in the control model and second, the regression coefficient of SDG6 increased considerably in the control model. Furthermore, while the regression coefficients of all the other SDGs remained in the same order of magnitude, that of the relatedness density of SDG6 was multiplied a hundredfold. Looking closer at the data this could be explained by the fact that, contrary to the full model, in the control model the relatedness density of SDG1, 6, and 10 was logarithmically transformed, a similar observation is thus expected for SDG1 and 10 in control model 4. In order to correct for this, the coefficients need to be divided by 100, which then brings it back in the same order of magnitude as the other SDGs. For the control variables, the regression coefficients were roughly the same. Important to note is that for the control variables, population became statistically significant for SDG5 and 9, contrary to the full model and that the GDP shows a slightly increased significant negative relationship to the research share of the combined SDGs. Finally, similar as to model 1 the full model has, overall, a slightly better fit to the regression equation than the control model.

Following, the comparison of model 3. Looking at the KCI score and the relatedness density scores, no significant changes in the regression coefficients where found, except for SDG3 and the relatedness density score of the combined SDGs. SDG3 shows a considerable larger positive relationship to the KCI score and a larger negative relationship to the relatedness density. As for these variables no imputed data was used, this effect is attributed to the number of observations, since the control model only incorporates 66% of the total observations. Going on to the regional characteristics, the main difference is that for the first characteristics, the regression coefficients of the control model are larger, while in this model no statistical significance could be identified. For the second set of characteristics no changes in significance where observed, but again for SDG3 there was a considerable increase in the negative relationship to the research share. This difference might have to do with the outliers, that are now more influential without the imputed data. For the control variables a few minor changes in the regression coefficients were observed. For population the coefficient increased overall, although a considerable increase could be observed for SDG1, SDG5 (although not significant), and SDG14. As population did not contain any missing values, this increase is attributed to the decreased number of observations. Besides, for GDP and education no significant differences in the regression coefficient were observed, but SDG3 for GDP and SDG1 and 10 for education did become statistically significant. Overall, both model's fits are again very comparable, but in this case the control model has a slightly better fit.

Finally, the control and full model of the fourth regression analysis is compared. In terms of the KCI variable, no relevant changes have been observed in the regression coefficient. In the control model, however, SDG1 and SDG6 lost their positive significant relationship to the research share. For the relatedness density, no changes in significance appeared. However, the regression coefficients of SDG1 and SDG10 did increase considerably, which again is attributed to being logarithmically transformed. Besides, SDG3 also showed an increase in the estimates, this might be due the limited number of observations. Looking at the regional characteristics, for the first characteristic the estimate of the regression coefficients is slightly higher in the full model for the statistically significant observations, besides SDG10 is no longer significant for this variable in the control model. The second characteristic variables show somewhat higher coefficients in the control model, however both SDG7 and 10 lose their significance. For the control variables the population coefficient is also higher in the control model, however no considerable other differences could be observed between the control and full model. Finally, similar to all models the fit of both models is very comparable, however the full model shows a higher overall fit.

Regression models control variables

In order to check for the impact of the control variables and the regional characteristics on the SDG research share, a closer look is given to the differences in variance that occur from the different models. First a model is compared, where only the control variables are included in the regression model, against the model where for all SDGs the KCI, RD, and control variables are included. This way it is attempted to explain the influence of KCI and RD on the variance and thus the fit of the model. Table 26 and 27 give an overview of how much of the variance is explained by the control variables for both time periods. This means that the rest of the variance can be attributed to the knowledge complexity and relatedness density variables. Overall, it can be seen that for most SDGs, except for SDG1, 10, and 16 for period 2010-2014, and the combined SDGs for period 2015-2020, the largest part of the variance is explained by the knowledge complexity and relatedness density. Besides, SDG2, SDG11, and the SDGs combined have a rather small part of the variables explaining the SDG research share, the variables combined explain approximately 10% of the change in research share. Indicating that other variables might explain the change better. Important to note is that in table 26 no share is given for SDG4, and in table 27 no share is given for SDG3. This is because these variables had a negative adjusted R^2 in the model that only included control variables. This means that the part that the control variables play in explaining the SDG research share is negligible.

Table 26:
Overview of the variance explained by the control variables (2010-2014)

SDG	Share	SDG	Share	SDG	Share	SDG	Share
All	35.7% (0.049)	**5**	32.2% (0.147)	**9**	15.8% (0.044)	**13**	16.6% (0.038)
1	84% (0.273)	**6**	39.0% (0.11)	**10**	51.7% (0.055)	**14**	10% (0.018)
2	14.9% (0.01)	**7**	14% (0.04)	**11**	0.5% (0.0003)	**15**	2.2% (0.005)
3	1.2% (0.002)	**8**	16.9% (0.04)	**12**	12.4% (0.037)	**16**	50.4% (0.256)

Table 27:
Overview of the variance explained by the control variables (2015-2020)

SDG	Share	SDG	Share	SDG	Share	SDG	Share
All	58% (0.072)	**5**	32.0% (0.151)	**9**	2.3% (0.008)	**13**	11.8% (0.03)
1	67% (0.234)	**6**	31.6% (0.08)	**10**	54.9% (0.239)	**14**	9.2% (0.016)
2	29.7% (0.022)	**7**	1.1% (0.008)	**11**	18.4% (0.027)	**15**	5.8% (0.018)
4	15.7% (0.151)	**8**	4.3% (0.013)	**12**	5.1% (0.018)	**16**	49.9% (0.225)

Second, the model where for all the SDGs the knowledge complexity, relatedness density, and the control variables are included (full model), is compared with the model where for a limited number of SDGs regional characteristics are included (RC model). This way it is attempted to explain the influence of the regional characteristic (RC) variables on the SDG research share. Table 28 gives an overview of how much of the variance is explained by the RC variables for both time periods. Overall, it can be seen that for most SDGs, except SDG 3 of the first time period and SDG 14 of the second time period, the RC variables don't explain that much of the variance. Besides, looking at the change in the variance of the two models and how much this increased by adding the RC variables, shown in brackets in table 28, it is rather small. For several SDGs adding the RC variables resulted in a less than 1% increase in the overall explanation of the SDG research share. For SDG7 in the first period and SDG15 in the second period no values were given at all, as the explained variance decreased by adding the RC variables, indicating that they did not explain the research share at all. In general, it can thus be said that by adding the RC variables the regression equation of the SDG research share was not explained considerably.

Table 28:
Overview of the variance explained by the regional characteristic variables

2010-2014				2015-2020			
SDG	**Share**	**SDG**	**Share**	**SDG**	**Share**	**SDG**	**Share**
1	5.8% (0.02)	**10**	1.1% (0.005)	**1**	0.9% (0.003)	**10**	6.5% (0.03)
2	0% (0)	**13**	0.4% (0.001)	**2**	8.6% (0.007)	**13**	1.2% (0.003)
3	20.7% (0.045)	**14**	1.1% (0.002)	**3**	7.4% (0.009)	**14**	19.8% (0.043)
7	-	**15**	0.4% (0.001)	**7**	4.2% (0.012)	**15**	-

Table 29 Overview of the variance explained by the regional characteristic variables

After performing these robustness checks no considerable differences were found between the control models and the full models that could influence the acceptance and rejection of the hypotheses. This means that the size of the effects does not change that much that the imputed data poses a problem to the dataset. However, some variables that were not statistically significant in the full models of the regression analyses were significant in the control models of the robustness checks, and vice versa. This means that some of the SDGs are not that robust for the effect they have on the SDG research share. This will be addressed further in the discussion section. Overall, it can be said that the positive and negative relationships of most of the variables is robust and that the results are roughly the same for the models where data was not imputed. Besides, looking at the change in the variance gives the results of the regression analyses extra robustness by showing the positive and negative relationships of the different variables, how this explains the SDG research share and to what extent.

5. Discussion

This section elaborates on the theoretical implications of this research, where the theory is linked with the insights from the results section. Besides, it discusses the theoretical and practical contributions to the existing knowledge base and society. Furthermore, the limitations with regards to the reliability and validity of this research are discussed. Finally, several recommendations are given for possible future research trajectories.

5.1 Theoretical implications

This study further explored the patterns of complex knowledge production regarding the SDGs for European regions. Theory was built around the concepts from the literature of Evolutionary Economic Geography, where knowledge complexity and scientific relatedness alongside with several regional characteristics were seen as key contributors to complex knowledge production on SDGs. A distinction was made in this research between the period leading up to the adoption of the goals, 2010-2014, and the period hereafter, 2015-2020. Further exploring the knowledge development in this area was seen as a valuable contribution to literature, as it gives an indication of the current state of the knowledge base of regions and thereby could identify leading and lagging regions on the track to sustainability and the 2030 Agenda.

Starting with the results of the descriptive analysis of regional knowledge production, results showed that the number of scientific publications were unevenly distributed over European regions. Publications were discovered to have a 'spiky' distribution, whereas a few regions produced considerably more scientific publications than other regions. Besides, these high performing regions were characterized by consisting metropolitan areas, such as London, Paris, and Madrid. These findings are consistent with earlier studies on path- and place-dependent mechanisms in Evolutionary Economic Geography (Balland & Rigby, 2017; Heimeriks et al., 2019; Nomaler et al., 2014). This literature is also found to be consistent with the results of the number of SDG related publications in Europe. Although this research is interested in research share, it was important to check whether the exact number of publications was also unevenly distributed. Again, it was seen that the highest performing regions distinguished themselves by metropolitan areas such as, London, Paris, Milan, and Barcelona. This was not surprising as these regions also had the highest scientific output in general. Looking at the research share, it was found that exclaves, overseas departments, and islands had the highest SDG share in their publications. This was surprising but could be most likely attributed to the fact that these regions had a low scientific output, and the limited number of publications they produced were related to SDGs due to their obligations to European Treaties (European Commission, 2017).

In addition, knowledge production is dependent on the complexity of the knowledge and the ability of a region to accumulate this knowledge. It is therefore said that the higher the complexity of a region's knowledge base is, the easier it is for a region to acquire and produce complex knowledge (Heimeriks & Balland, 2016). By mapping the knowledge complexity index (KCI) of regions it was shown that mainly regions in the North-West of Europe had the most complex and diverse knowledge base, which is in line with findings of Heimeriks et al. (2019). The findings of this study also indicated the presence of spatial spillovers, as the production of complex knowledge and the complexity of a region's knowledge base was spatially concentrated and often stretched over neighboring regions, as also became apparent in research by Pintar and Scherngell (2021). In addition to this, the findings suggested that geographical proximity is an important driver for complex knowledge production, as close interactions are required for the effective communication and distribution of complex knowledge (Kim & Anand, 2018). Consequently, as the results suggested from the distribution of the SDG related research share, regions agglomerate as the complexity of a regions knowledge base increases. These agglomeration effects became especially apparent by looking at the most complex regions in Europe, that were mainly located in the United Kingdom. Overall, the knowledge complexity of a regions was seen as an important driver to SDG-related research as most SDGs showed a positive significant relationship to the knowledge complexity of a region.

Another driver for knowledge production is the scientific relatedness, that positively influenced the SDG related research for all SDGs, with the exception of SDG3. Thus, regions with a knowledge base closer related to the SDGs do engage more in SDG related research. This indicates that path-dependent mechanisms that lead actors to build on existing knowledge they have required in the past, lead to scientific specialization of European regions on the SDGs. This is in line with previous studies by Cohen and Levinthal (1990) on absorptive capacity, that suggest that regions need a degree of prior related knowledge to be able to assimilate and exploit new knowledge. Besides, it adds to the literature by Hidalgo et al. (2018) and Heimeriks and Boschma (2014), that regions that have a similar and complementary knowledge base to the SDGs are more capable to assimilate new knowledge on the SDGs. Furthermore, this confirms that the existing knowledge base of a regions to the SDGs increases the likeliness to diversify in related SDGs and discourages the knowledge production on unrelated SDGs (Heimeriks et al., 2019). This is especially relevant for the SDGs, that are highly complex due to their inter- and transdisciplinarity and build upon knowledge from a diverse range of scientific research fields (Arroyave et al., 2021).

In contrast to previous findings is SDG3, that showed a negative relationship between the scientific relatedness and SDG research share. Earlier mentioned possible reasons for this relationship could be due to a less complex knowledge base that is required for this SDG or due to the large share of publications that is already attributed to this SDG. Research by Zhang et al. (2015) builds upon these arguments and gives two explanations of why scientific relatedness has a negative relationship with scientific output. On the one hand, it might be that the scientific value of SDG3 is exhausted. This means that when the costs of research exceed the benefits, regions might be inclined to diversify further into related areas. On the other hand, SDG3 might show a negative relationship due to ongoing globalization. Globalization promotes international collaboration and knowledge dissemination, which makes it easier for regions with a low relatedness to SDG3 to still diversify into this area and produce knowledge on it. Overall, scientific relatedness can still be considered as an important theoretical concept for SDG related research as it also provides the opportunity, once knowledge is acquired on an SDG, to diversify into related SDGs as shown in the relatedness space of the SDGs.

In case of the regional characteristics, hardly any significant relationships could be determined and the relationships that were discovered varied in being either negative or positive. The regional characteristics units varied per SDG, so it was not surprising that this is reflected in the different type of relationships. Only taking the SDGs into consideration that were significant in either of both time periods, it could be seen that the expectation that regional adverse effects lead to an increase in SDG research share was only right for the first characteristic of SDG3 in the first time period and for both characteristics of SDG14 in the second time period. The characteristic of SDG3 is concerned with the proportion of births attended by skilled health personnel. A closer look at the data already showed that most regions scored very high on this characteristic, as the worst performing regions had a value of 95%. It is therefore not really possible to speak of regional adverse effects for this characteristic, meaning that this result should not be considered of large importance to the research. The characteristics of SDG14 are related to the average proportion of Marine Key Biodiversity Areas covered by protected areas and beach litter. Earlier it was already confirmed that coastal regions perform considerably more research in this SDGs than land-locked regions. Consequently, when these regions experience adverse effects, it is more likely they increase research and protect important marine and beach areas as it affects them directly. In contrary, the other statistically significant regional characteristics, had a negative relationship to the SDG research share, meaning that although they underperformed in these areas, research was not increased. This is in line with findings by Ciarli and Ràfols (2019) that found a misalignment between research priorities and societal needs. The authors saw a potential for the SDGs to set relevant directions and reverse the misalignments, but the findings of this research showed that these are still present. Overall, regional characteristics in the form of local adverse effects, based on the SDG indicators, cannot be considered to be of considerable effect when explaining SDG related research. Nevertheless, these findings are important for setting future research trajectories as societal needs appear to be neglected and the misalignment is still present.

5.2 Limitations

During this research several methodological limitations emerged that were taken into consideration during the analysis. The first limitation arose from the use of the Web of Science (WoS) database, that biased the general data sample. By retrieving publications through this database, a considerable number of publications was left out. WoS only incorporates scientific publications, thereby leaving out more practical documents such as patents. Including these documents might provide better insights into the practical capabilities of a region. Furthermore, as indicated earlier the WoS is known to strongly represent natural sciences, while publications in the fields of humanities and social sciences, as well as books are sparsely represented (Mingers & Leydesdorff, 2015). In addition, publications in foreign languages are not included in the database but could be valuable to the research as local knowledge is often expressed in a region's native language and can be considered highly relevant. Consequently, the exclusion of these types of publications might lead to an incorrect estimate of the SDG research share among regions as the data sample does not represent the entire knowledge base of a region. Nevertheless, expanding the research beyond the use of WoS and developing key words for publications in foreign languages to identify SDG related research would prove to be very time consuming and not feasible for this study.

Another effect of the use of this database can be seen from the calculations on the knowledge complexity index (KCI). Since social sciences and humanities are underrepresented in the sample, these proved to be the most complex scientific subfields when looking at the KCI. This is not surprising, if we take the considerable influence of ubiquity on the KCI into account. If certain subfields are underrepresented, it is logical that relatively less regions produce knowledge on this topic, thereby lowering ubiquity and increasing the knowledge complexity. This stresses a limitation of the use of the KCI. Furthermore, as earlier mentioned, the KCI is an extended version of the model by Hidalgo and Hausmann (2009), that includes intangible assets. As the original model was designed to fit tangible products, the question arises whether the KCI explains the knowledge complexity of a region as well as the original model explains economic complexity. Further research is required to test for the robustness of the KCI and should prove whether the KCI is a direct reflection of the complexity of a region's knowledge base, or whether other determinants should be included besides the diversity and ubiquity, such as for example citations.

Another methodological limitation was the use of the thesaurus and the corresponding keywords to retrieve the SDG-related publications. In general, the use of key words encounters several problems such as being very literal. On the one hand, synonyms are not retrieved from publications if not specified, on the other hand words with multiple meanings often make keywords return false positives (i.e., irrelevant results). Besides, keywords are very dependent on the researcher's perspective and the aim of the research or project, as they are manually selected to fit the requirements of the study. Earlier research by Romero-Goyeneche et al. (2021) and Rafols et al. (2021), already showed that, when retrieving SDG related publications, using different approaches yield very different results. The use of keywords is thus of considerable influence on the data sample that is retrieved. This research made use of the STRINGS thesaurus, that is mostly based on policy agendas with a focus on individual SDGs. This thesaurus is known to attribute a substantial number of publications to SDG3, while according to Romero-Goyeneche et al. (2021) this should be distributed more evenly across SDG1 and

2. Besides, as this thesaurus was not developed by researchers incorporating the transformative lens perspective (Ramirez et al., 2019), it might have influenced the assumption that socio-technological SDGs are best explained by the independent variables. This was, however, taken into consideration during the analysis. Besides, there is no best practice method yet on mapping the SDGs and keeping these limitations in mind during the analysis allowed for a robust study.

Finally, several other limitations should be acknowledged for this research. First, although the indicators for regional characteristics were carefully chosen, they do not fully comprehend the SDGs they represent. For the selected SDGs only two indicators were selected, from which some due to computational issues were excluded from the regression analyses. Furthermore, for each SDG the variables differed in terms of what they measured, therefore the results could not be compared or generalized. Besides, these variables were provided for on the country level, meaning no differences between regions from the same country could be distinguished. Second, in the robustness check some variables lost their significance, this means that although the size of the effect (i.e., the regression coefficient) is robust, the SDGs themselves were not. This could be attributed to the influence of the imputed data and both the regional characteristic and control variables. However, the size of the effects was not considered to be problematic for the research. Finally, as mentioned earlier the SDGs are highly complex goals, that spread over several disciplines, interact with each other, and cannot be solved easily. Therefore, using a quantitative approach with a linear regression analysis, might not be the best solution to explain research on SDGs. Nevertheless, by simplifying the SDGs into a few variables, although they are more likely to be explained by many, this research added to the understanding of the goals and made research on SDGs more tangible.

5.3 Contributions

This study contributes to the existing knowledge base of the literature on Evolutionary Economic Geography. The concepts of this research area are used to understand and analyse the possibilities and constraints of regions in producing knowledge on the SDGs. The findings of this research demonstrate that scientific relatedness and knowledge complexity are important drivers for the production of knowledge on the SDGs Consequently, this research shows that regional knowledge development regarding the SDGs is subject to path- and place-dependency mechanisms, which provides further empirical evidence on the theory of Evolutionary Economic Geography. Besides confirming this existing theory, this study extended the theory by, instead of taking a comprehensive perspective, taking a more specific perspective by only targeting the SDGs. For this purpose, this study implemented the previous work from the STRINGS project to estimate the knowledge base of European regions related to the SDGs. This approach allowed the combination of existing knowledge to distinguish SDG related research without the extensive use of keywords and create an overview of regions leading and lagging in research on the SDGs based on visualization maps. It thus provides a first attempt to explain SDG related research for 16 SDGs based on an extensive database to ensure validity and reliability. Besides, this study does an initial attempt to operationalize regional characteristics and study their effects. Although more research and empirical evidence is required, this research demonstrates the concerning misalignment between research priorities and societal needs.

Besides scientific contributions, this study also provides several societal contributions. There is not a single solution in achieving the SDGs, as they require the collaboration of various groups and institutions across a variety of fields. By exploring the effects that lead to an increased SDG research share, regions are given opportunities to increase research. The findings from this study contribute by stressing the importance of, amongst others, the geographical proximity and knowledge complexity. By exchanging knowledge and collaborating with better performing neighboring regions, the complexity of regions and their SDG research share could be increased. Furthermore, the important role of scientific relatedness allows governments and policymakers to focus on specific SDGs that are underrepresented in their knowledge base but related to their existing knowledge base. Thus, overall, this research contributed to reaching the sustainability goals by providing a more tangible perspective in how to increase SDG research, thereby giving guidance in directing future trajectories for knowledge-based sustainable growth.

5.4 Recommendations for further research

Taking all this into consideration, this research opens up opportunities for future research. This study focused mainly on regional characteristics based on the SDG indicators, where no clear positive relationship could be identified. However, the research showed that other characteristics do have an influence on the research share of SDGs, such as regions bordering the coast having a higher research share in SDG14 (Life Below Water). For other SDGs no such relationship was identified, but this finding does call for more extensive research into geographical regional characteristics, that might influence the research on SDGs. If other considerable effects are to be found this could greatly promote collaboration between specialized regions and foster knowledge exchange. Another variable that is worth to further investigate is the contribution of non-SDG related research on the knowledge production relating the SDGs. From the analysis it could be seen that regions with a high scientific output only had a limited SDG research share. A possible explanation for this is that they have sufficient funds to also explore other scientific areas of research. Interesting would be then to see whether this other research influences the research on SDGs and if so in what way; does it promote the research on SDGs or does certain research poses threats to the SDG related research and their objectives?

Other interesting opportunities for further research involve on the one hand, the expansion of the data sample. This study built on data from the WoS, which as earlier mentioned comes with several limitations. Interesting would be to add publications from other data sources, to see how this influences the results and to approach a more genuine representation of a region's knowledge base. By including scientific publications cited in patents, a more practical and applied side of the knowledge base of a region could be highlighted. Besides, by adding data from the Dimensions (n.d) or Overton (n.d.) database, also policy documents, think tank publications, clinical trials and working papers can be added to the data sample, thereby creating a more comprehensive approach to estimate a region's knowledge base. On the other hand, instead of expanding the data sample, it might also prove interesting to narrow down the data sample, by focusing on specific regions. By taking a more qualitative approach, knowledge on lagging and leading regions could be further deepened. By conducting interviews and in-depth analyses on factors promoting and constraining SDG related research, the capabilities of regions and the effects of amongst others formal and informal institutions could be researched. Results from these types of studies could than again be expanded across other regions to see whether the effects can be generalized, and the SDG research share better explained.

Finally, this research showed that the explanation of the SDG research share by the different variables differs per SDG. Although this research attempted to find overarching effects for the research share of the different SDGs, it might be good to consider the SDGs individually. The SDGs differ greatly in terms of targets and objectives, and it can therefore be assumed that the factors that explain them might not be equal for all the SDGs. A deeper analysis can be done on a specific SDG instead of targeting all of them, thereby leaving more room to test for different effects. Additionally, from the transformative lens approach it becomes apparent that the SDGs don't operate in isolation but interact with and reinforce each other. Therefore, although looking at the SDGs as an individual unit, future research should also focus on the interactions between them. If research is to discover the complex interactions between the SDGs, this could be used to further promote the research share. Furthermore, if there is more insight in how the SDGs support and perhaps overlap each other in certain areas, regions could more specifically target SDGs to effectively promote and foster transformative change.

6. Conclusion

Society is increasingly becoming aware of the pressure the exponential growth of the human population has created on the earth (Sachs, 2012). Therefore, the Sustainable Development Goals (SDGs) were introduced in 2015 as a global agenda to guide the world on a sustainable trajectory. Although several efforts have been made, progress is slow, and scholars urge scientists from different scientific fields to share knowledge and collaborate more to fill knowledge gaps in critical areas (McCollum et al., 2018; Messerli et al., 2019). In order to promote structural transformation, it is of key importance to gain insights in the current knowledge base of regions to identify adjacent possibilities for diversification and understand the underlying dynamics of knowledge production with regards to the SDGs. To find factors that promote or constrain the capability of regions to produce complex knowledge on the SDGs, this research aims to answer the following research question and sub questions:

What explains differences among regions in complex knowledge production on SDG related research?

a. What is the relation between the complexity of a region's knowledge base and its ability to produce SDG related research?
b. What is the relation between the relatedness of a region's knowledge base to the SDGs and its ability to produce SDG related research?
c. How do region-specific characteristics influence the complex knowledge production on SDGs?

To answer these questions concepts from the Evolutionary Economic Geography literature are used, namely the complexity of a region's knowledge base, based on the knowledge complexity index, and the relatedness density of a region's knowledge base to the SDGs. Besides, regional characteristics are incorporated using the SDG indicators developed by the UN and a transformative lens perspective is taken to subdivide the SDGs into three types. The data sample consists of scientific publications retrieved from the Web of Science through the CWTS wos_2113 database from before the introduction of the SDGs, 2010-2014, and after the introduction, 2015-2020. SDG related research is then distinguished using data from the STRINGS project. In doing so, SDG3, Health and Well-being, is found to be by far the SDG with the largest number of publications, mainly due to the STRINGS thesaurus being somewhat biased towards SDG3. Other initial findings show that the most complex regions are located in the North-West of Europe and Spain, a similar pattern is found for the relatedness density to the SDGs. In order to find concrete relationships between the SDG related research and the independent variables, several regression models are estimated based on the two time periods. As the regional characteristic variables are selected for a limited number of SDGs, these are distinguished leading to two models for the different time periods, one with the inclusion of the regional characteristic variables and one without.

Based on the results of this analysis the sub questions can be answered, from which an answer to the main research question can be deduced. For the first sub question, the results show, in general, a positive relationship between the complexity of a region's knowledge base and the SDG research share of that region. This means that for half of the SDGs, with the exception of SDG2, 7, 8, 12, 13, and 15 for which no statistically significant relationship is found and SDG6 and 11 that show a negative relationship, the ability of a region to produce knowledge on these SDGs is dependent on the degree of knowledge complexity of that region. A similar but clearer trend is found for the second sub question. The findings show that the relationship between the scientific relatedness of a region to an SDG is significant and positive for all SDGs, with the exception of SDG3 that shows a negative significant relationship. This indicates that if a region's knowledge base is related to the knowledge required for an SDG it increases the ability of that region to diversify into SDG related research. This is, however, thus not the case for SDG3 where the scientific relatedness of a region to SDG3 does not increase the research related to this topic. Possible explanations that are found were due to the exhaustion of the scientific value and globalization, that decreases the cognitive and geographic proximity advantages.

Finally, the last sub question on the influence of region-specific characteristics on the SDG related research share. Overall, only for characteristics of SDG1, 3, 7, 10, and 14 a statistically significant relationship is found, leaving SDG2, 13, and 15 out of this conclusion. The hypothesis could only be accepted for the first characteristic of SDG3 and both of SDG14, meaning that the remaining characteristics were rejected. This indicates that, contrary to what was expected, for most of the researched SDGs region-specific characteristic do not influence the knowledge production on these SDGs. This finding is quite alarming as it highlights that although certain regions underperform based on the SDG indicators, research is not steered towards the objectives of the SDGs. Large differences thus exist between the SDGs and since it was not possible to find any overarching characteristics for the specific SDGs, the regional characteristic variables represent different indicators and sometimes different units. It is therefore not possible to generalize these findings across the SDGs and give one clear and distinctive answer to the last sub question. It does, however, provides valuable insights in the current state of regions and allows for further research to deepen knowledge in this area.

Coming back to the research question, the differences among regions in complex knowledge production on SDG related research can be partly explained by the complexity and scientific relatedness of the knowledge base of a region and to a lesser extent by the regional characteristics of a region. However, findings show that differences exist between the various SDGs and that there is not a single explanation that fits all the SDGs, instead they are all predicted in a different manner. Besides, not only the complexity and relatedness of the knowledge base of a region is important in producing knowledge on the SDGs, also the geographic location of the regions is of importance as knowledge production is characterized by being path- and place-dependent. This also becomes apparent when looking at SDG14, Life below Water, where the scientific relatedness is the highest for regions bordering the ocean compared to land-locked regions. This indicates that there are other variables beyond this research that might prove an explanation for the regional differences in SDG related research. Moreover, the limited changes that are observed between the different time periods show that the introduction of the SDGs did not considerably change the research trajectories of the regions. A simple explanation is that regions did not choose to diversify into unrelated fields regarding the SDGs, but instead kept building on the existing body of knowledge they had.

From this research several recommendations can be made to shape future trajectories regarding knowledge production related to the SDGs. This study provides evidence that for most SDGs the scientific relatedness positively influences knowledge production on SDGs. This opens up opportunities for policymakers, governments, and research institutes to diversify into SDGs that are not only related to their knowledge base, but direct research into SDGs that are closely related to SDGs that fall within their current knowledge base. Besides, by using the knowledge complexity index as a tool, the strengths and capabilities of a regions scientific profile can be better assessed. By looking at the ubiquity and diversity of the knowledge, gaps can be identified and opportunities for smart specialization arise. This approach gives stakeholders concrete handholds to identify leading and lagging regions and allows them to focus on specific SDGs that are underrepresented in their current environment but have some degree of relatedness to a region. Finally, this research provides evidence for the advantages of geographic proximity. Therefore, it is highly recommended to increase inter-regional collaboration. By exchanging knowledge with neighboring regions tacit information is easier accumulated and it could help diminish the difficulties that stem from the inter- and transdisciplinarity of the goals. In short, by identifying the strengths and opportunities of regions and increasing collaboration, stakeholders can steer for more specialized and focused trajectories thereby accelerating the achievement of the goals.

The achievement of the SDGs should have a priority on the agendas of all the regions. The rate at which the earth is declining, is too high to keep ignoring it. Therefore, the United Nations have called for help and support of scientific institutions and their research. In general, there is a need for clear legislation and inventive new technologies that can contribute to reaching the goals. This research provides an initial attempt to explain how regions can increase SDG related research in order to give handholds and to further expand the current knowledge base on SDGs. Important is that regions continue to build upon their strengths and expand research into areas that are related to their current knowledge base and the SDGs. By prioritizing these related fields and increasing collaboration, SDG related research can be increased. Nevertheless, this study also shows that the SDG related research cannot entirely be explained by mechanisms of path- and place dependency. It is therefore of key importance that research on other factors that explain a regions knowledge base related to the SDGs continues, as scientific research is the starting point of the global sustainability transition.

Reference list

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374. https://doi.org/10.1002/wics.84

Allison, P. (2015, March 5). Imputation by Predictive Mean Matching: Promise & Peril. *Statistical Horizons*.

Arroyave, F., Romero-Goyeneche, O. Y., Gore, M., Heimeriks, G., Jenkins, J., & Petersen, A. (2021). On the social and cognitive dimensions of wicked environmental problems characterized by conceptual and solution uncertainty. *Advances in Complex Systems*.

Autant-Bernard, C., Fadairo, M., & Massard, N. (2013). Knowledge diffusion and innovation policies within the European regions: Challenges based on recent empirical evidence. *Research Policy*, *42*(1), 196–210. https://doi.org/10.1016/j.respol.2012.07.009

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Balland, P. A. (2017). Economic Geography in R: Introduction to the EconGeo Package. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2962146

Balland, P.-A., & Boschma, R. (2021). Complementary interregional linkages and Smart Specialisation: an empirical study on European regions. *Regional Studies*, *55*(6), 1059–1070. https://doi.org/10.1080/00343404.2020.1861240

Balland, P.-A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Regional Studies*, *53*(9), 1252–1268. https://doi.org/10.1080/00343404.2018.1437900

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*, *4*(3), 248–254. https://doi.org/10.1038/s41562-019-0803-3

Balland, P.-A., & Rigby, D. (2017). The Geography of Complex Knowledge. *Economic Geography*, *93*(1), 1–23. https://doi.org/10.1080/00130095.2016.1205947

Barbieri, N., Perruchas, F., & Consoli, D. (2020). Specialization, Diversification, and Environmental Technology Life Cycle. *Economic Geography*, *96*(2), 161–186. https://doi.org/10.1080/00130095.2020.1721279

Bertero, M. (2009). Science attracts Spain. *EMBO Reports*, *10*(8), 817–819. https://doi.org/10.1038/embor.2009.174

Boschma, R. (2004). Competitiveness of Regions from an Evolutionary Perspective. *Regional Studies*, *38*(9), 1001–1014. https://doi.org/10.1080/0034340042000292601

Boschma, R., Heimeriks, G., & Balland, P.-A. (2014). Scientific knowledge dynamics and relatedness in biotech cities. *Research Policy*, *43*(1), 107–114. https://doi.org/10.1016/j.respol.2013.07.009

Boschma, R., & Martin, R. (2010). The aims and scope of evolutionary economic geography. In *The Handbook of Evolutionary Economic Geography*. https://doi.org/10.4337/9781849806497.00007

Bretschger, L. (1999). Knowledge diffusion and the development of regionsKnowledge diffusion and the development of regions. *The Annals of Regional Science*, *33*(3), 251–268. https://doi.org/10.1007/s001680050104

Bryman, A. (2008). *Social Research Methods* (3rd ed.). Oxford University Press.

Callen, T. (2020, February 24). *Gross Domestic Product: An Economy's All*. International Monetary Fund. https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm

Chandra, P., & Dong, A. (2018). The relation between knowledge accumulation and technical value in interdisciplinary technologies. *Technological Forecasting and Social Change*, *128*, 235–244. https://doi.org/10.1016/j.techfore.2017.12.006

Ciarli, T., & Ràfols, I. (2019). The relation between research priorities and societal demands: The case of rice. *Research Policy*, *48*(4), 949–967. https://doi.org/10.1016/j.respol.2018.10.027

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, *35*(1), 128–152. https://doi.org/10.2307/2393553

Dimensions. (n.d.). *The world's largest linked research information dataset*. Retrieved July 15, 2022, from https://www.dimensions.ai

European Commission. (n.d.-a). *Eurostat - European statistics*. Retrieved December 1, 2021, from https://ec.europa.eu/info/departments/eurostat-european-statistics_en

European Commission. (n.d.-b). *Sustainable Development Goals | International Partnerships*. Retrieved December 1, 2021, from https://ec.europa.eu/international-partnerships/sustainable-development-goals_en

European Commission. (2017). *The Outermost Regions; European lands in the world*. https://ec.europa.eu/regional_policy/sources/policy/themes/outermost-regions/pdf/rup_2017/rup_eu_lands_world_en.pdf

European Union. (2020). *Statistical regions in the European Union and partner countries — NUTS and statistical regions 2021*.

Eurostat. (2022). *Regional gross domestic product (PPS per inhabitant) by NUTS 2 regions [Table]*. https://ec.europa.eu/eurostat/databrowser/product/view/NAMA_10R_2GDP

Eyisi, D. (2016). The Usefulness of Qualitative and Quantitative Approaches and Methods in Researching Problem-Solving Ability in Science Education Curriculum. *Journal of Education and Practice*, *7*(15), 91–100.

Federal Planning Bureau. (2021, June 10). *Natural disaster victims*. Indicators.Be . https://indicators.be/en/i/G13_NDV/Natural_disaster_victims

Fritsch, M., & Slavtchev, V. (2005). The Role of Regional Knowledge for Innovation. *European Regional Science Association (ERSA)*, 23–27. http://hdl.handle.net/10419/117753www.econstor.eu

Fuso Nerini, F., Tomei, J., To, L. S., Bisaga, I., Parikh, P., Black, M., Borrion, A., Spataru, C., Castán Broto, V., Anandarajah, G., Milligan, B., & Mulugetta, Y. (2018). Mapping synergies and trade-offs between energy and the Sustainable Development Goals. *Nature Energy*, *3*(1), 10–15. https://doi.org/10.1038/s41560-017-0036-5

Granberg-Rademacker, J. S. (2007). A Comparison of Three Approaches to Handling Incomplete State-Level Data. *State Politics & Policy Quarterly*, *7*(3), 325–338. https://doi.org/10.1177/153244000700700306

Hall, P., & Soskice, D. (2001). An Introduction to Varieties of Capitalism. In *Varieties of Capitalism* (pp. 1–68). Oxford University Press. https://doi.org/10.1093/0199247757.003.0001

Harel, O. (2009). The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, *36*(10), 1109–1118. https://doi.org/10.1080/02664760802553000

Heimeriks, G., & Balland, P.-A. (2016). How smart is specialisation? An analysis of specialisation patterns in knowledge production. *Science and Public Policy*, *43*, 562–574. https://doi.org/10.1093/scipol/scv061

Heimeriks, G., & Boschma, R. (2014). The path- and place-dependent nature of scientific knowledge production in biotech 1986-2008. *Journal of Economic Geography*, *14*(2), 339–364. https://doi.org/10.1093/jeg/lbs052

Heimeriks, G., Li, D., Lamers, W., Meijer, I., & Yegros, A. (2019). Scientific knowledge production in European regions: patterns of growth, diversity and complexity. *European Planning Studies*, *27*(11), 2123–2143. https://doi.org/10.1080/09654313.2019.1645814

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics*, *3*(2), 92–113. https://doi.org/10.1038/s42254-020-00275-1

Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The Product Space Conditions - the Development of Nations. *Science*, *317*(5837), 482–487. https://doi.org/10.1126/science.1144581

Hidalgo, C., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S., & Zhu, S. (2018). *The Principle of Relatedness* (pp. 451–457). https://doi.org/10.1007/978-3-319-96661-8_46

Hidalgo, C., & Hausmann, R. (2009). The Building Blocks of Economic Complexity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10570–10575. https://doi.org/10.1073/pnas.0900943106

Kim, S., & Anand, J. (Jay). (2018). Knowledge complexity and the performance of inter-unit knowledge replication structures. *Strategic Management Journal*, *39*(7), 1959–1989. https://doi.org/10.1002/SMJ.2899

Kolk, A. (2013). Mainstreaming sustainable coffee. *Sustainable Development*, *21*(5), 324–337. https://doi.org/10.1002/sd.507

Kostoska, O., & Kocarev, L. (2019). A Novel ICT Framework for Sustainable Development Goals. *Sustainability*, *11*(7), 1961. https://doi.org/10.3390/su11071961

LaMorte, W. (2016, May 31). *The Multiple Linear Regression Equation*. Boston University School of Public Health. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html

Laursen, K., Masciarelli, F., & Prencipe, A. (2007). *Regions Matter: How Regional Characteristics Affect External Knowledge Acquisition and Innovation* (No. 07–20). www.druid.dk

Leal Filho, W., Azeiteiro, U., Alves, F., Pace, P., Mifsud, M., Brandli, L., Caeiro, S. S., & Disterheft, A. (2018). Reinvigorating the sustainable development research agenda: the role of the sustainable development goals (SDG). *International Journal of Sustainable Development & World Ecology*, *25*(2), 131–142. https://doi.org/10.1080/13504509.2017.1342103

Levin, K. A. (2006). Study design III: Cross-sectional studies. *Evidence-Based Dentistry*, *7*(1), 24–25. https://doi.org/10.1038/sj.ebd.6400375

Leydesdorff, L., & Bensman, S. (2006). Classification and powerlaws: The logarithmic transformation. *Journal of the American Society for Information Science and Technology*, *57*(11), 1470–1486. https://doi.org/10.1002/asi.20467

Li, D. (2020). *Place-dependence of renewable energy technologies. Connecting the local and global scale.* [Utrecht University]. http://dspace.library.uu.nl/handle/1874/396682

Li, D., Heimeriks, G., & Alkemade, F. (2021). Knowledge flows in global renewable energy innovation systems: the role of technological and geographical distance. *Technology Analysis & Strategic Management*, 1–15. https://doi.org/10.1080/09537325.2021.1903416

Lönnqvist, A., Käpylä, J., Salonius, H., & Yigitcanlar, T. (2014). Knowledge That Matters: Identifying Regional Knowledge Assets of the Tampere Region. *European Planning Studies*, *22*(10), 2011–2029. https://doi.org/10.1080/09654313.2013.814621

Makri, M., Hitt, M. A., & Lane, P. J. (2010). Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal*, *31*, 602–628. https://doi.org/10.1002/smj.829

Manimannan, G., Salomi, M., Lakshmi Priya, R., & Saranraj. (2020). Detecting Outliers using R Package in Fitting Data with Linear and Nonlinear Regression Models. *International Journal of Scientific and Innovative Mathematical Research*, *8*(4). https://doi.org/10.20431/2347-3142.0804001

Martin, R., & Sunley, P. (2006). Path dependence and regional economic evolution. *Journal of Economic Geography*, *6*(4), 395–437. https://doi.org/10.1093/jeg/lbl012

McCollum, D. L., Echeverri, L. G., Busch, S., Pachauri, S., Parkinson, S., Rogelj, J., Krey, V., Minx, J. C., Nilsson, M., Stevance, A.-S., & Riahi, K. (2018). Connecting the sustainable development goals by their energy inter-linkages. *Environmental Research Letters*, *13*(3), 033006. https://doi.org/10.1088/1748-9326/aaafe3

Messerli, P., Kim, E. M., Lutz, W., Moatti, J.-P., Richardson, K., Saidam, M., Smith, D., Eloundou-Enyegue, P., Foli, E., Glassman, A., Licona, G. H., Murniningtyas, E., Staniškis, J. K., van Ypersele, J.-P., & Furman, E. (2019). Expansion of sustainability science needed for the SDGs. *Nature Sustainability*, *2*(10), 892–894. https://doi.org/10.1038/s41893-019-0394-z

Mićić, L. (2017). Digital Transformation and Its Influence on GDP. *Economics*, *5*(2), 135–147. https://doi.org/10.1515/eoik-2017-0028

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19. https://doi.org/10.1016/j.ejor.2015.04.002

Mitchell, R. E., & Leach, B. (2019). Knowledge coproduction in environmental impact assessment: Lessons from the mining industry in Panama. *Environmental Policy and Governance*, *29*(2), 87–96. https://doi.org/10.1002/eet.1832

Moore, A. W., Anderson, B., Das, K., & Wong, W.-K. (2006). Combining Multiple Signals for Biosurveillance. In *Handbook of Biosurveillance* (pp. 235–242). Elsevier. https://doi.org/10.1016/B978-012369378-5/50017-X

Nilsson, M., Chisholm, E., Griggs, D., Howden-Chapman, P., McCollum, D., Messerli, P., Neumann, B., Stevance, A.-S., Visbeck, M., & Stafford-Smith, M. (2018). Mapping interactions between the sustainable development goals: lessons learned and ways forward. *Sustainability Science*, *13*(6), 1489–1503. https://doi.org/10.1007/s11625-018-0604-z

Nomaler, Ö., Frenken, K., & Heimeriks, G. (2014). On Scaling of Scientific Knowledge Production in U.S. Metropolitan Areas. *PLoS ONE*, *9*(10), e110805. https://doi.org/10.1371/journal.pone.0110805

OECD. (2021). *Population with tertiary education*. https://data.oecd.org/eduatt/population-with-tertiary-education.htm

Overton. (n.d.). *Track your policy impact*. Retrieved July 15, 2022, from https://www.overton.io

Perruchas, F., Consoli, D., & Barbieri, N. (2020). Specialisation, diversification and the ladder of green technology development. *Research Policy*, *49*(3), 103922. https://doi.org/10.1016/j.respol.2020.103922

Pettigrew, L. M., de Maeseneer, J., Anderson, M.-I. P., Essuman, A., Kidd, M. R., & Haines, A. (2015). Primary health care and the Sustainable Development Goals. *The Lancet*, *386*(10009), 2119–2121. https://doi.org/10.1016/S0140-6736(15)00949-6

Pintar, N., & Scherngell, T. (2021). The complex nature of regional knowledge production: Evidence on European regions. *Research Policy*, 104170. https://doi.org/10.1016/J.RESPOL.2020.104170

Rafols, I., Noyons, E., Confraria, H., & Ciarli, T. (2021, May 10). *Visualising plural mappings of science for Sustainable Development Goals (SDGs)*.

Ramirez, M., Romero, O., Schot, J., & Arroyave, F. (2019). *Mobilizing the Transformative Power of the Research System for Achieving the Sustainable Development Goals*. https://doi.org/10.2139/ssrn.3497623

Reichert, S. (2006). *The Rise of Knowledge Regions: Emerging Opportunities and Challenges for Universities*. www.eua.be

Rodrigues, G., & Breach, A. (2021). *What levelling up really means. Changing the geography of knowledge in Britain*.

Rodríguez-Barranco, M., Tobías, A., Redondo, D., Molina-Portillo, E., & Sánchez, M. J. (2017). Standardizing effect size from linear regression models with log-transformed variables for meta-analysis. *BMC Medical Research Methodology*, *17*(1), 44. https://doi.org/10.1186/s12874-017-0322-8

Rodríguez-Pose, A. (2013). Do Institutions Matter for Regional Development? *Regional Studies*, *47*(7), 1034–1047. https://doi.org/10.1080/00343404.2012.748978

Romero Goyeneche, O. Y., Ramirez, M., Schot, J., & Arroyave, F. (2022). Mobilizing the transformative power of research for achieving the Sustainable Development Goals. *Research Policy*, *51*(10), 104589. https://doi.org/10.1016/j.respol.2022.104589

Romero-Goyeneche, O. Y., Vis, J., Vila i de Villasante, E., Arroyave, F., Heimeriks, G., de Laat, J., & Schot, J. (2021). *Addressing the Sustainable Development Goals at Utrecht University*. https://www.uu.nl/sites/default/files/SDG%20Project%20Report%207%20October%202021.pdf

Sachs, J. D. (2012). From Millennium Development Goals to Sustainable Development Goals. *The Lancet*, *379*(9832), 2206–2211. https://doi.org/10.1016/S0140-6736(12)60685-0

Salmi, J. (2017). *The Tertiary Education Imperative*. SensePublishers. https://doi.org/10.1007/978-94-6351-128-5

Salvia, A. L., Leal Filho, W., Brandli, L. L., & Griebeler, J. S. (2019). Assessing research trends related to Sustainable Development Goals: local and global issues. *Journal of Cleaner Production*, *208*, 841–849. https://doi.org/10.1016/j.jclepro.2018.09.242

Sanfey, P., & Milatovic, J. (2018). *The Western Balkans in transition: diagnosing the constraints on the path to a sustainable market economy*.

Schikowitz, A. (2020). Creating relevant knowledge in transdisciplinary research projects - Coping with inherent tensions. *Journal of Responsible Innovation*, *7*(2), 217–237. https://doi.org/10.1080/23299460.2019.1653154

Schot, J., Boni, A., Ramirez, M., & Steward, F. (2018). *Addressing SDGs through Transformative Innovation Policy*. www.tipconsortium.net

Schot, J., & Kanger, L. (2018). Deep transitions: Emergence, acceleration, stabilization and directionality. *Research Policy*, *47*(6), 1045–1059. https://doi.org/10.1016/j.respol.2018.03.009

Soetewey, A. (2020, November 8). *Outliers detection in R*. Stats and R. https://statsandr.com/blog/outliers-detection-in-r/

Soini, K., Jurgilevich, A., Pietikäinen, J., & Korhonen-Kurki, K. (2018). Universities responding to the call for sustainability: A typology of sustainability centres. *Journal of Cleaner Production*, *170*, 1423–1432. https://doi.org/10.1016/j.jclepro.2017.08.228

the World Bank. (n.d.). *Refugee population by country or territory of asylum - Luxembourg*. Retrieved May 12, 2022, from https://data.worldbank.org/indicator/SM.POP.REFG?locations=LU

Tranmer, M., Murphy, J., Elliot, M., & Pampaka, M. (2020). *Multiple Linear Regression (2nd Edition)*. https://hummedia.manchester.ac.uk/institutes/cmist/a rchive-publications/working-papers/2020/2020-1- multiple-linear-regression.pdf

UN General Assembly. (2015, October 21). *Transforming our world : the 2030 Agenda for Sustainable Development*. A/RES/70/1. https://www.refworld.org/docid/57b6e3e44.html

UN General Assembly. (2022). *Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development*. https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202022%20refinement_Eng.pdf

UN-iLibrary. (n.d.). *Global SDG Indicators Database* . Retrieved December 15, 2021, from https://www.un-ilibrary.org/economic-and-social-development/data/global-sdg-indicators-database_099907c5-en

UNIS. (n.d.). *Student life*. Retrieved May 3, 2022, from https://www.unis.no/studies/student-life/

United Nations. (n.d.). *Decade of Action*. Retrieved November 23, 2021, from https://www.un.org/sustainabledevelopment/decade-of-action/

United Nations. (2019a). THE FUTURE IS NOW | Science for achieving sustainable development. *Global Sustainable Development Report 2019*. https://sustainabledevelopment.un.org/content/documents/24797GSDR_report_2019.pdf

United Nations. (2019b). *World Population Prospects 2019*.

Uniyal, S., Paliwal, R., Kaphaliya, B., & Sharma, R. K. (2020). Human Overpopulation. In *Megacities and Rapid Urbanization* (pp. 20–30). IGI Global. https://doi.org/10.4018/978-1-5225-9276-1.ch002

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Wade, C. (2018, August 21). *Transforming Skewed Data. A Survey of Friendly Functions*. Towards Data Science. https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16

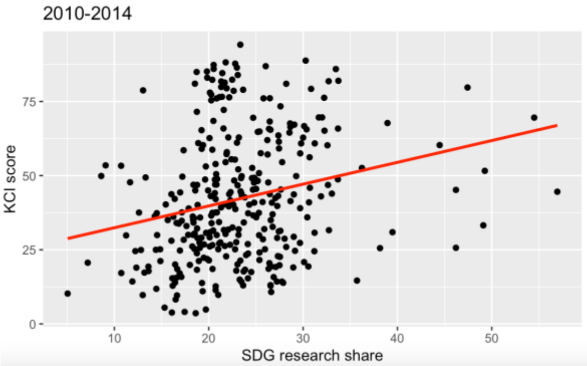
Zhang, J., Yan, Y., & Guan, J. (2015). Scientific relatedness in solar energy: a comparative study between the USA and China. *Scientometrics*, *102*(2), 1595–1613. https://doi.org/10.1007/s11192-014-1487-0

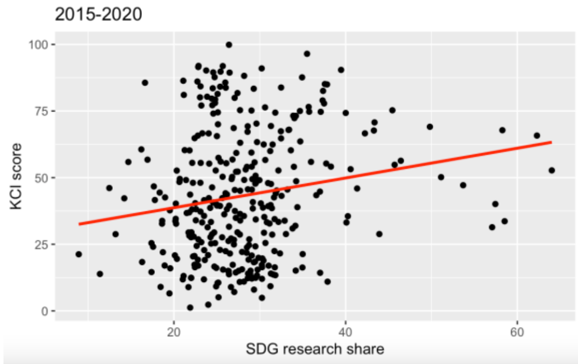
Appendix

Appendix A: Scatterplots

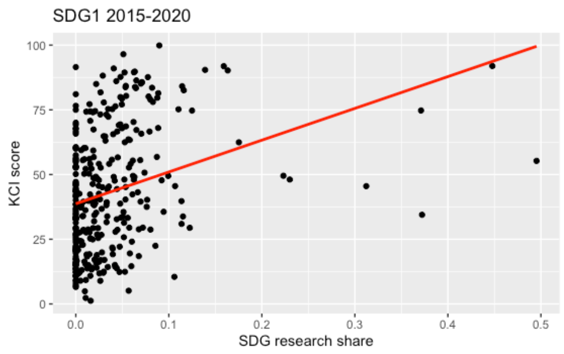
A1: SDG research share and Knowledge complexity (KCI)

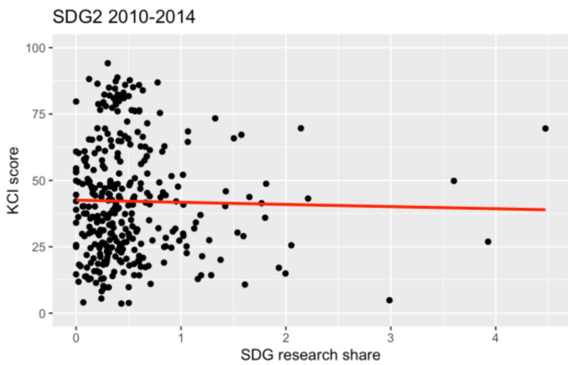
All SDGs combined:

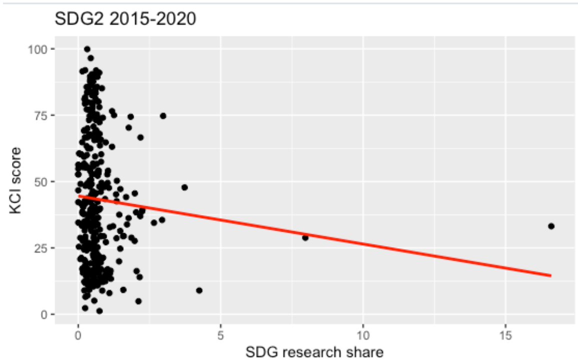


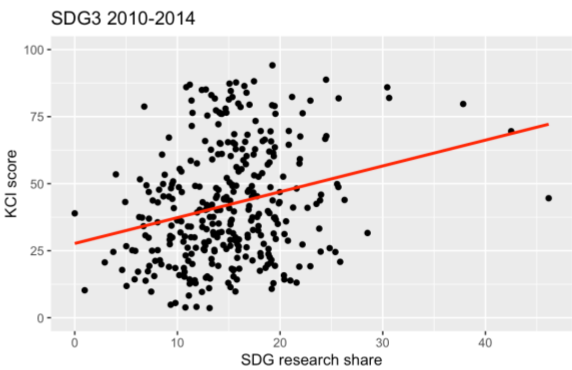


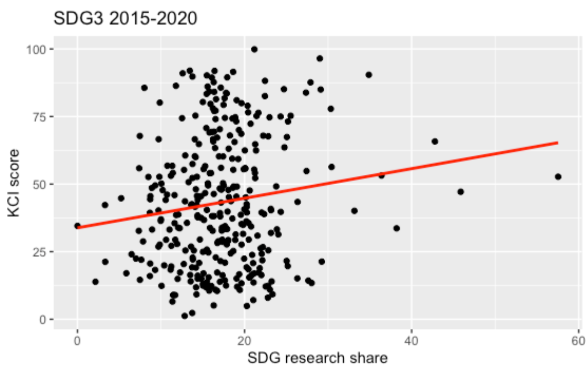
Individual SDGs:



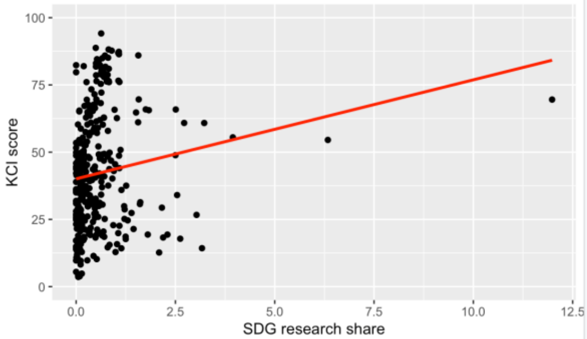




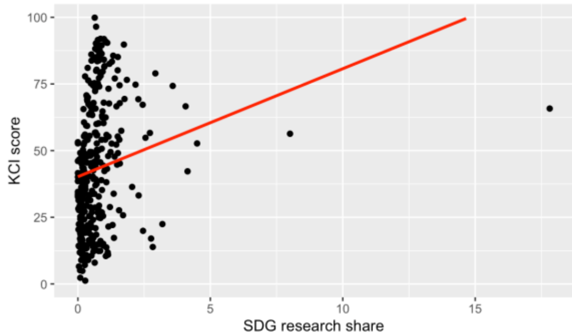




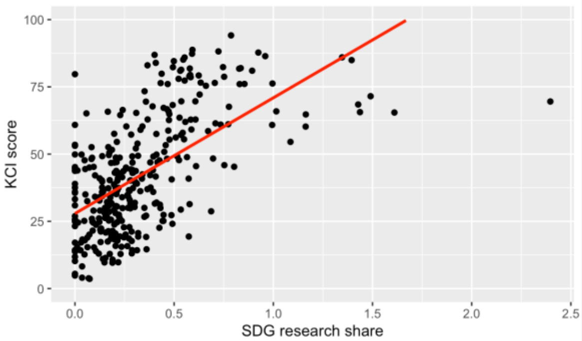
SDG4 2010-2014



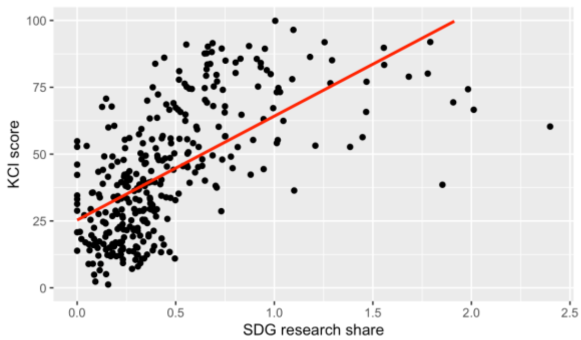
SDG4 2015-2020



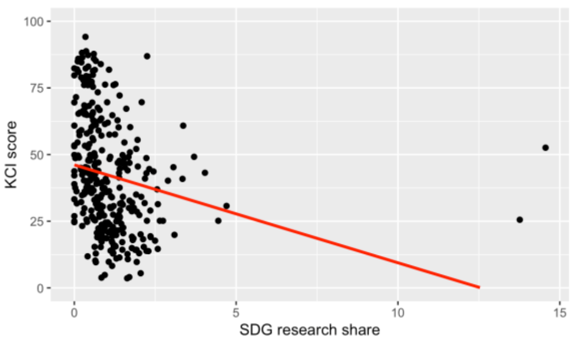
SDG5 2010-2014



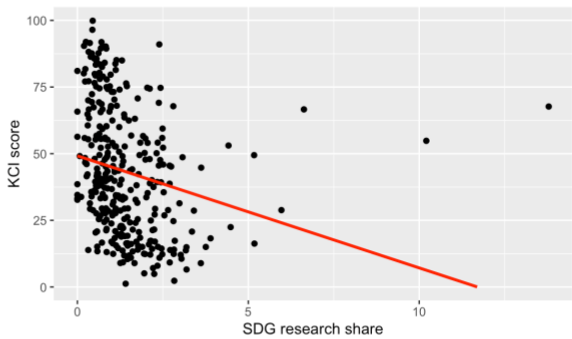
SDG5 2015-2020



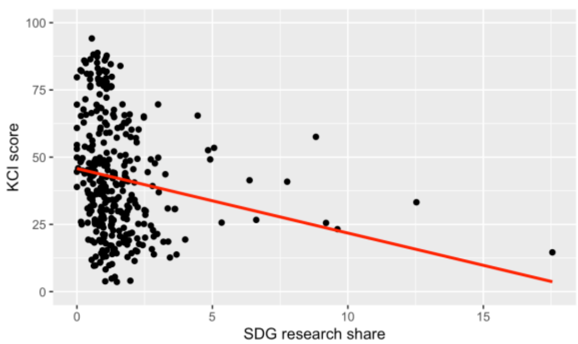
SDG6 2010-2014



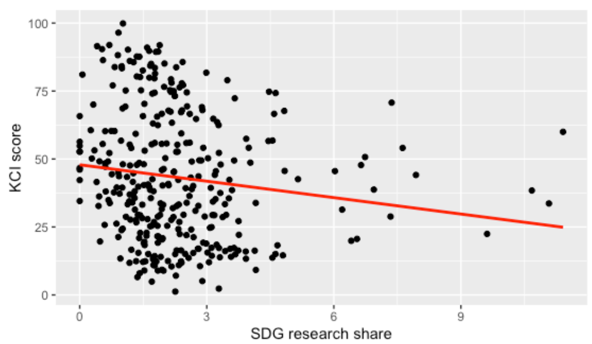
SDG6 2015-2020

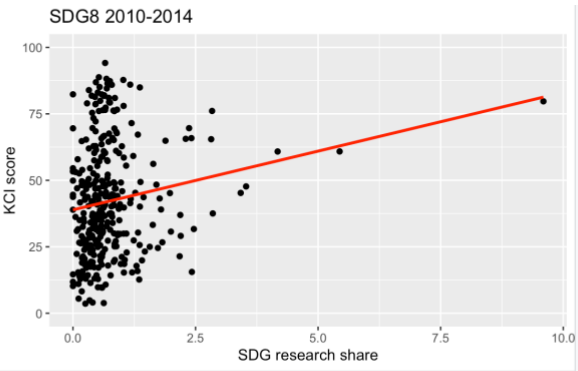


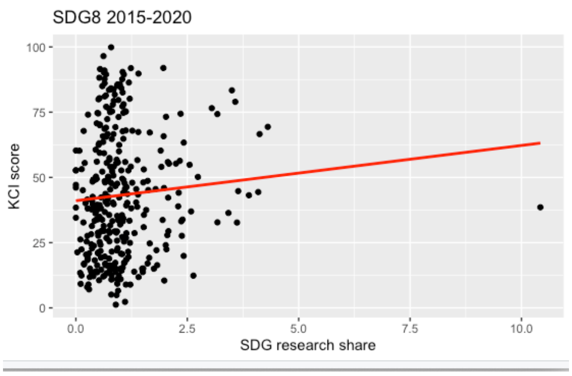
SDG7 2010-2014

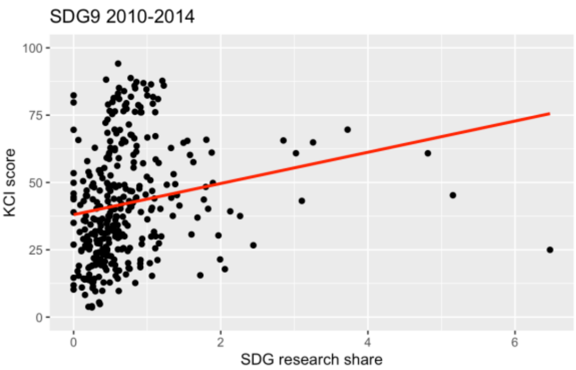


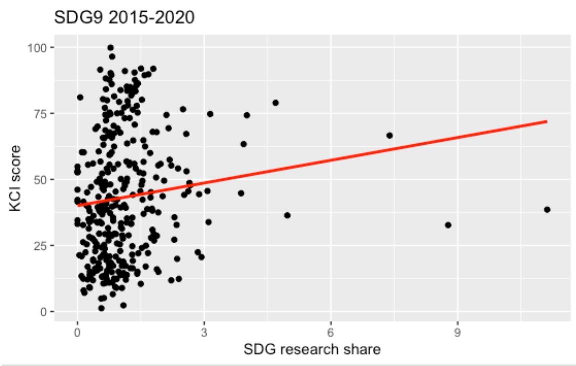
SDG7 2015-2020

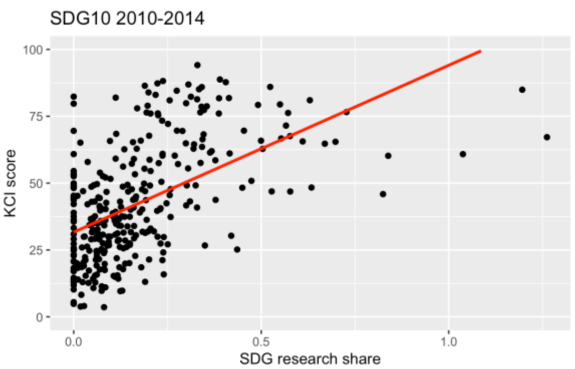


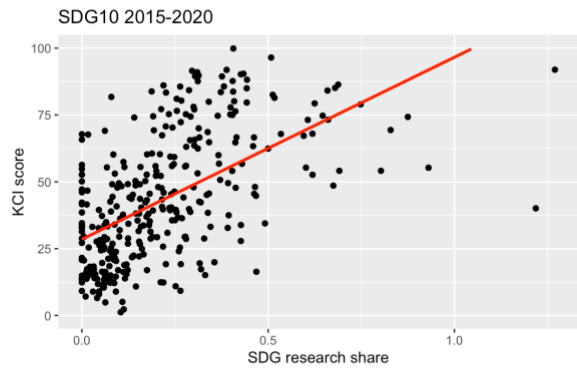

SDG8 2010-2014
KCI score
SDG research share

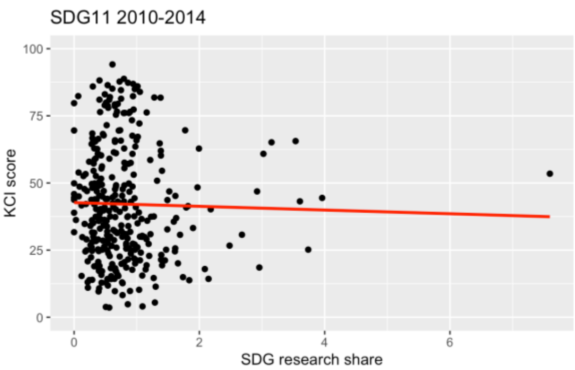

SDG8 2015-2020
KCI score
SDG research share

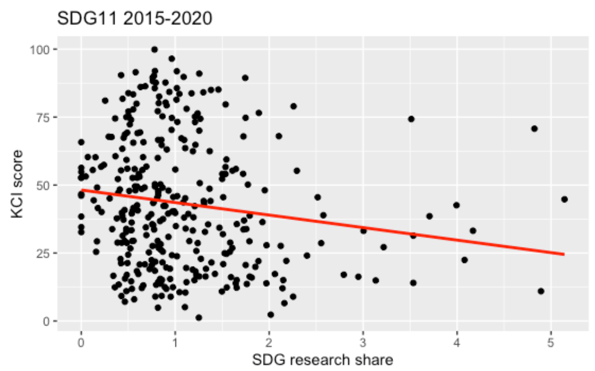

SDG9 2010-2014
KCI score
SDG research share


SDG9 2015-2020
KCI score
SDG research share

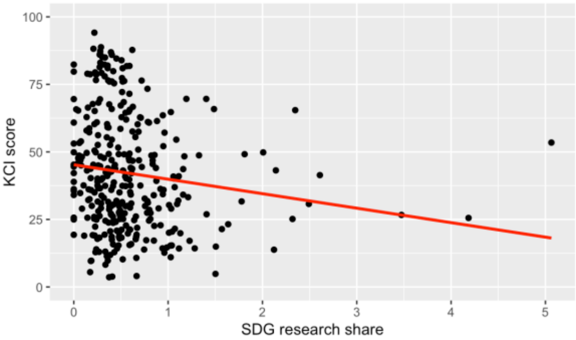

SDG10 2010-2014
KCI score
SDG research share


SDG10 2015-2020
KCI score
SDG research share

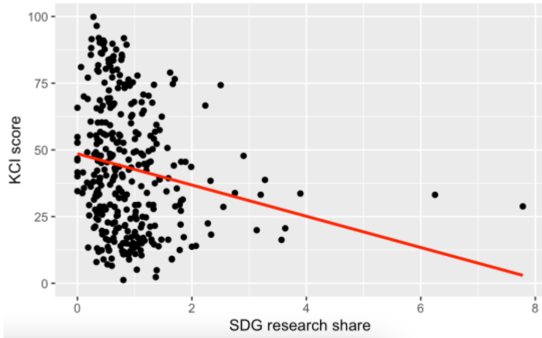

SDG11 2010-2014
KCI score
SDG research share


SDG11 2015-2020
KCI score
SDG research share

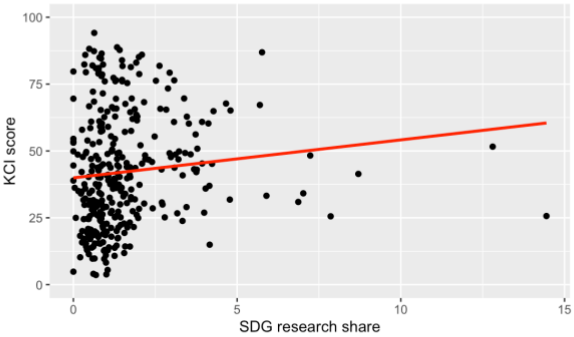
SDG12 2010-2014



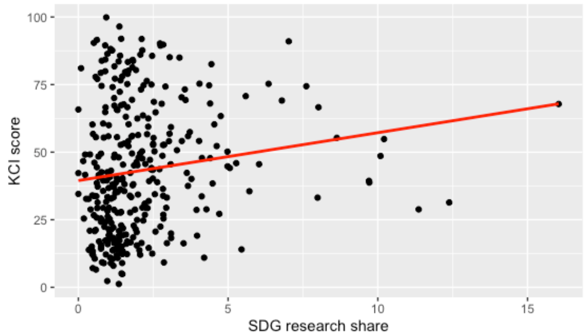
SDG12 2015-2020



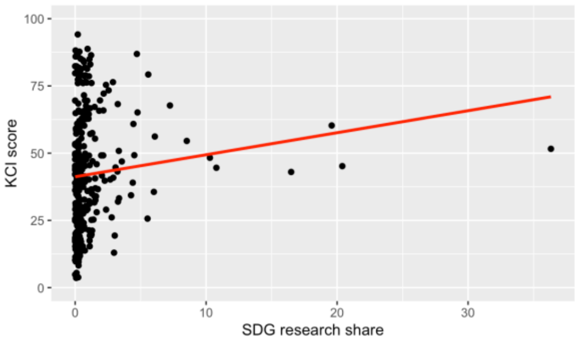
SDG13 2010-2014



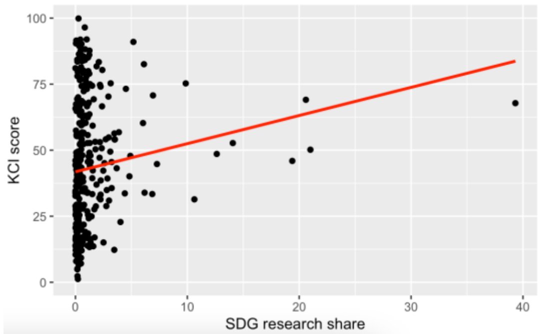
SDG13 2015-2020



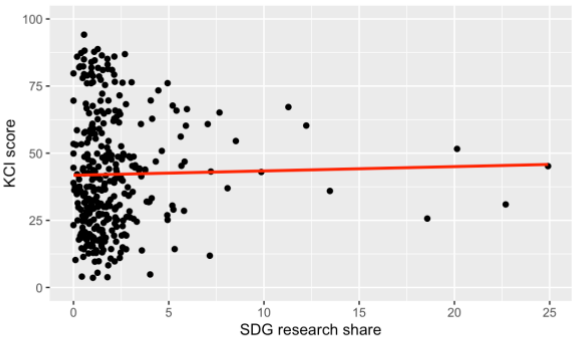
SDG14 2010-2014



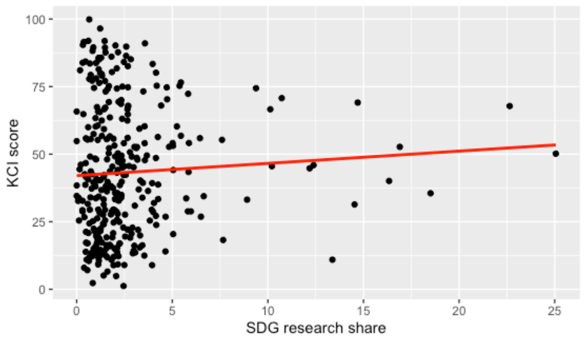
SDG14 2015-2020

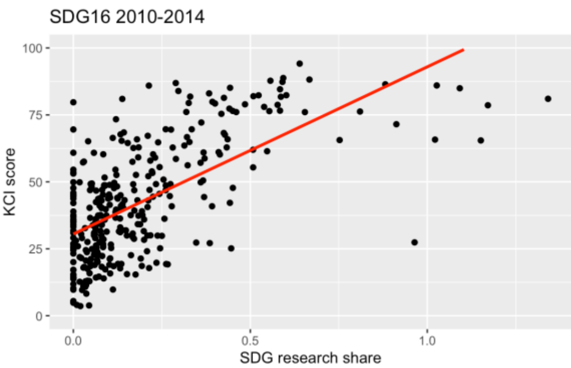


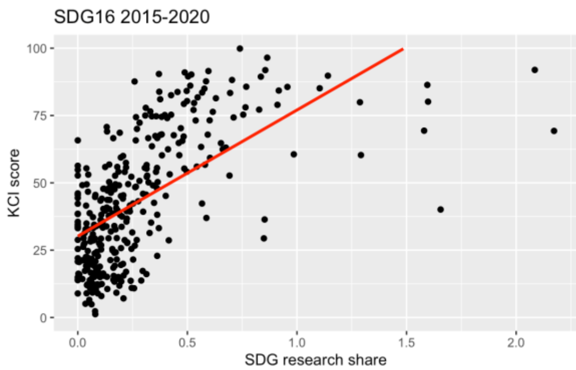
SDG15 2010-2014



SDG15 2015-2020

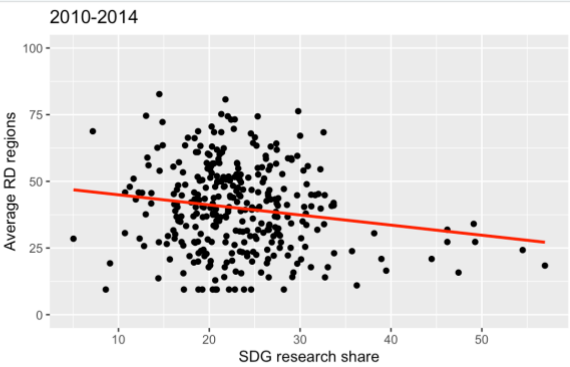


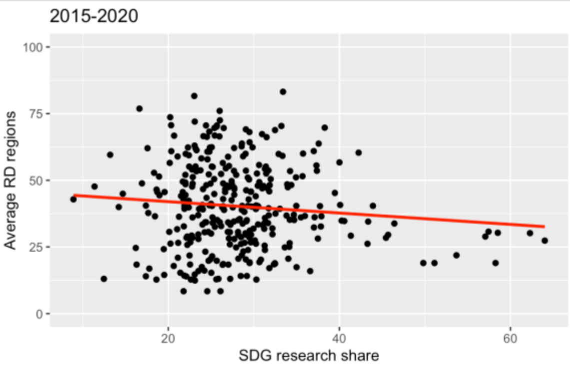




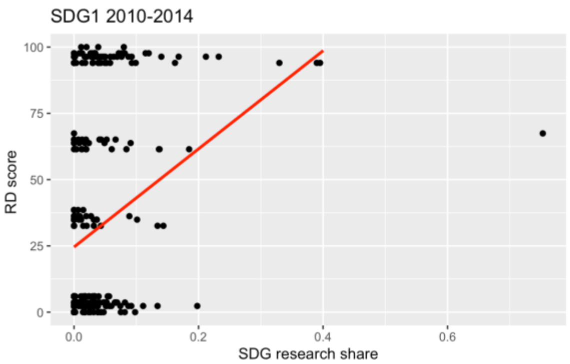
A2: SDG research share and scientific relatedness density (RD)

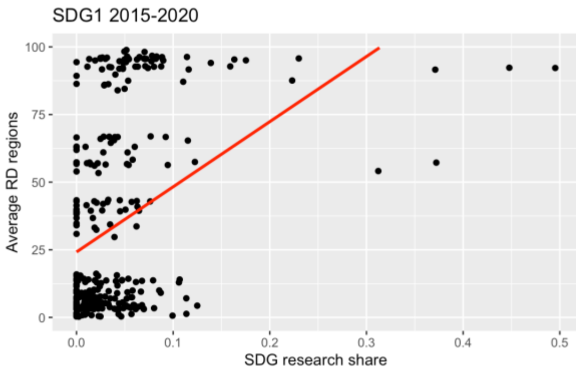
All SDGs combined:

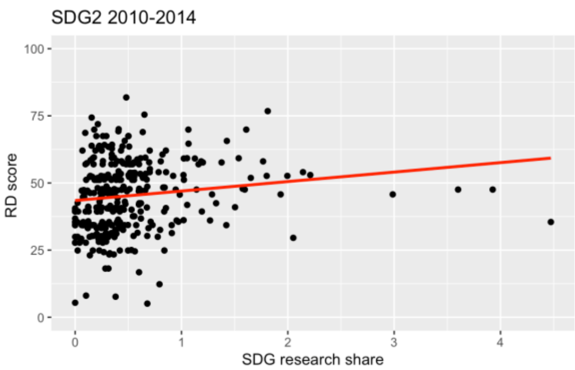


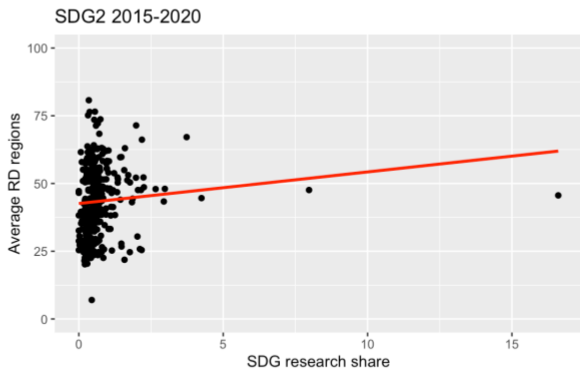


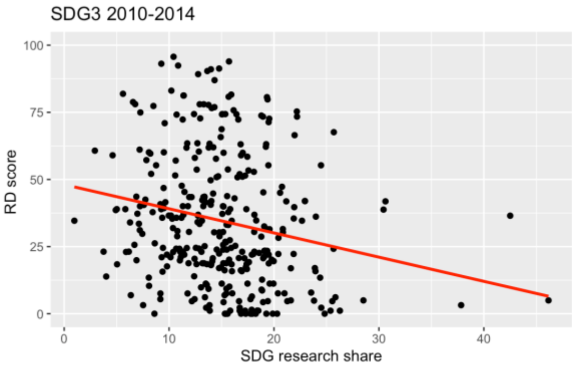
Individual SDGs:

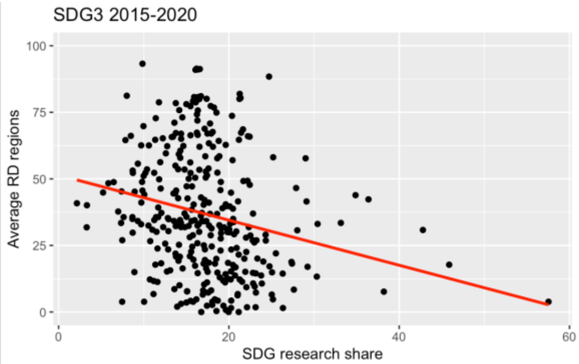


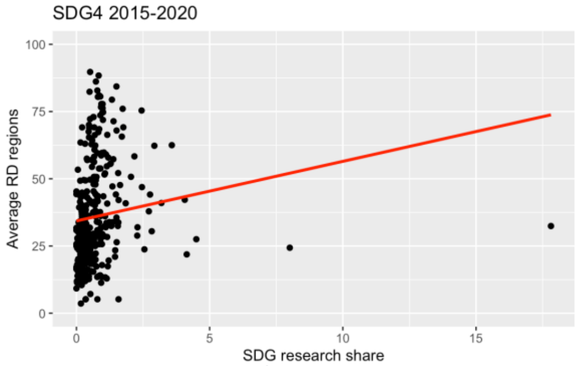




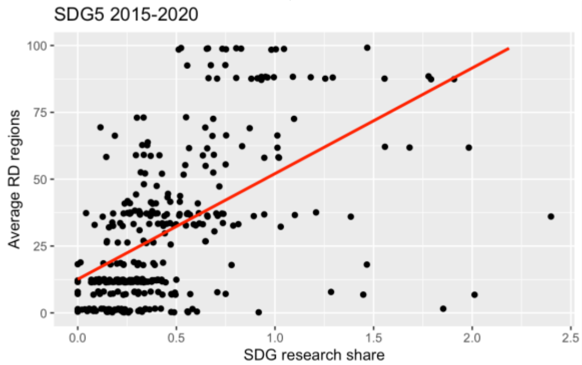


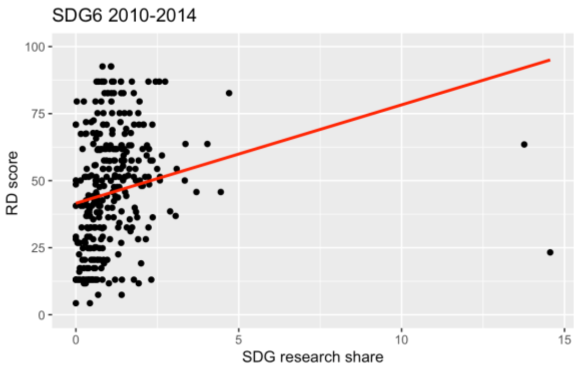

SDG3 2010-2014
RD score
SDG research share

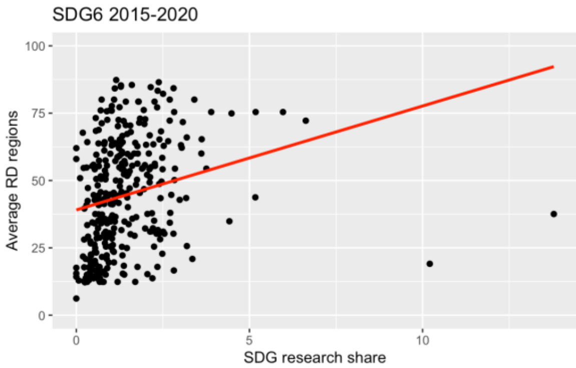

SDG3 2015-2020
Average RD regions
SDG research share


SDG4 2015-2020
Average RD regions
SDG research share

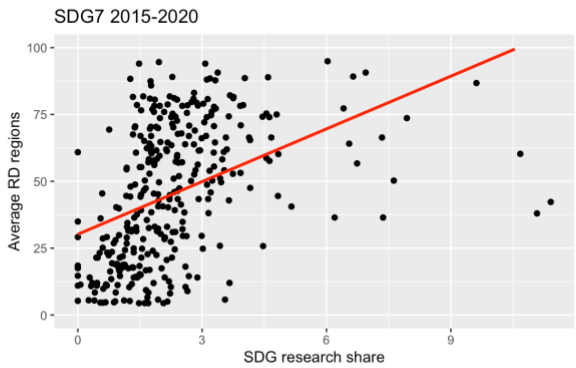

SDG5 2010-2014
RD score
SDG research share

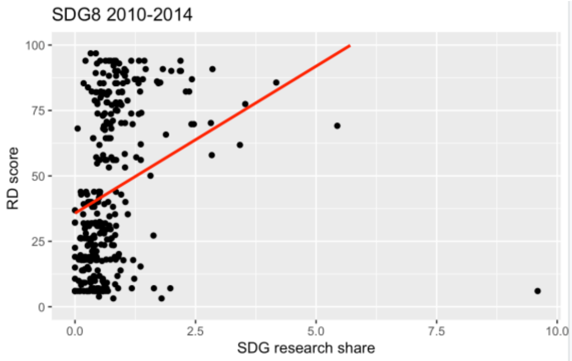

SDG5 2015-2020
Average RD regions
SDG research share

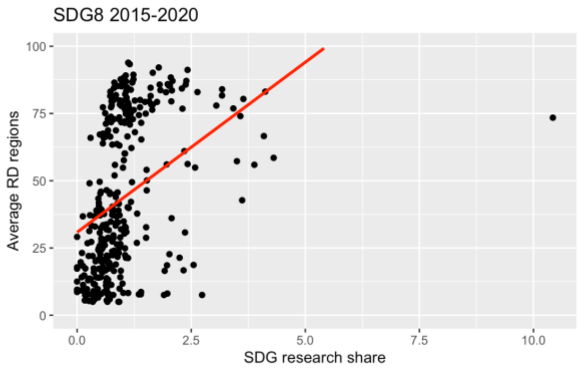

SDG6 2010-2014
RD score
SDG research share

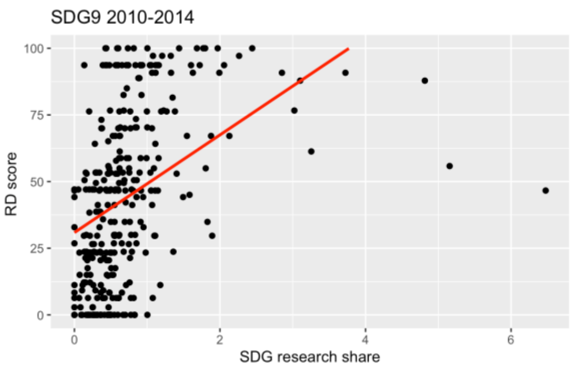

SDG6 2015-2020
Average RD regions
SDG research share

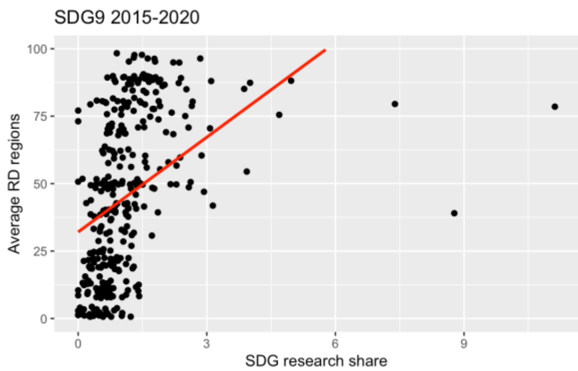

SDG7 2010-2014
RD score
SDG research share
0
25
50
75
100
0
5
10
15


SDG7 2015-2020
Average RD regions
SDG research share
0
25
50
75
100
0
3
6
9

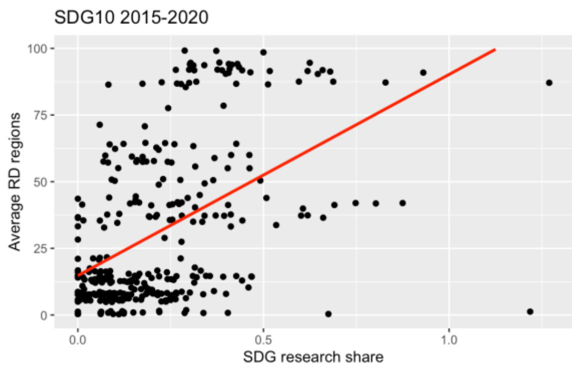

SDG8 2010-2014
RD score
SDG research share
0
25
50
75
100
0.0
2.5
5.0
7.5
10.0

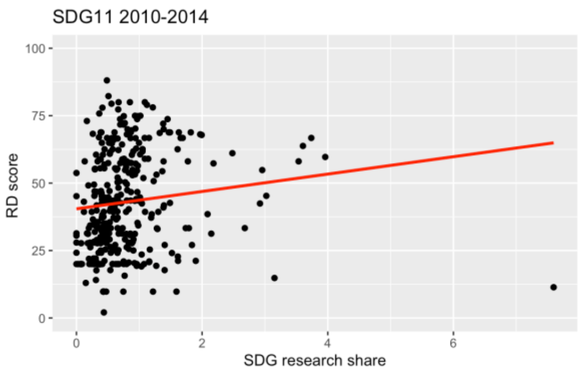

SDG8 2015-2020
Average RD regions
SDG research share
0
25
50
75
100
0.0
2.5
5.0
7.5
10.0

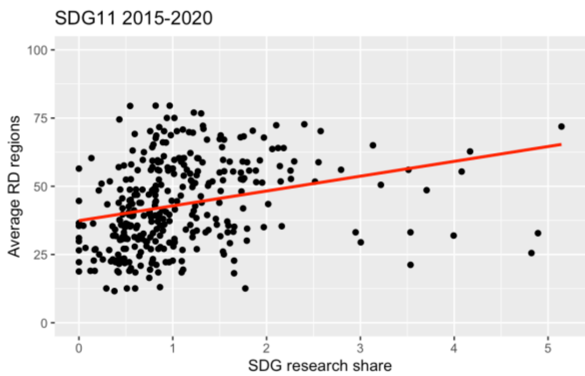

SDG9 2010-2014
RD score
SDG research share
0
25
50
75
100
0
2
4
6

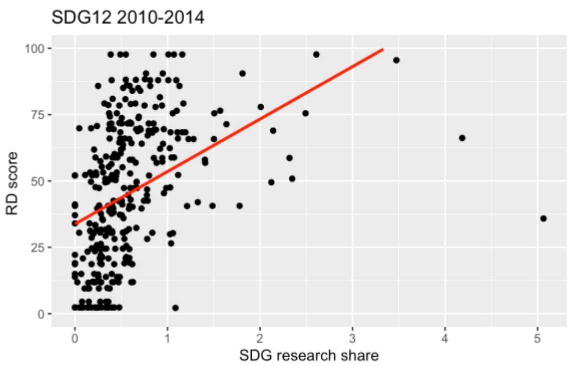

SDG9 2015-2020
Average RD regions
SDG research share
0
25
50
75
100
0
3
6
9

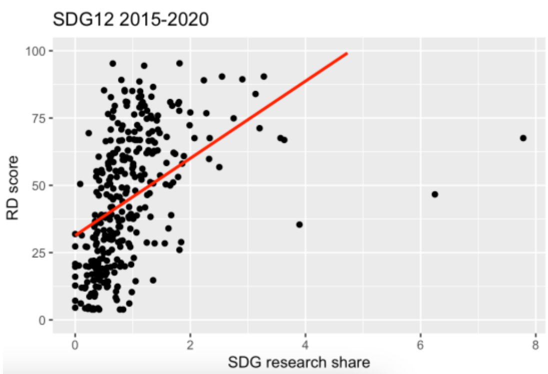

SDG10 2010-2014
RD score
SDG research share
0
25
50
75
100
0.0
0.5
1.0

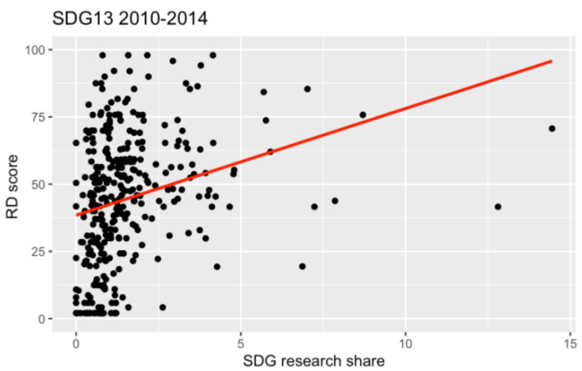

SDG10 2015-2020
Average RD regions
SDG research share
0
25
50
75
100
0.0
0.5
1.0

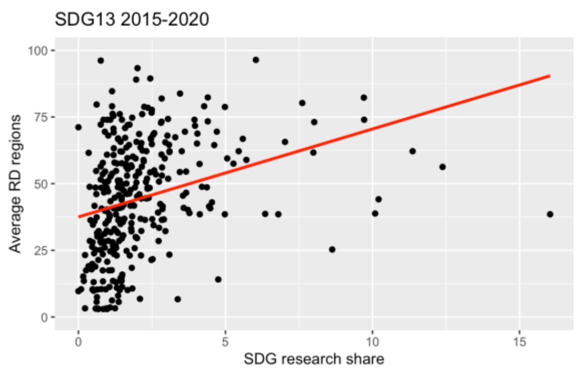

SDG11 2010-2014
RD score
SDG research share
0 25 50 75 100
0 2 4 6

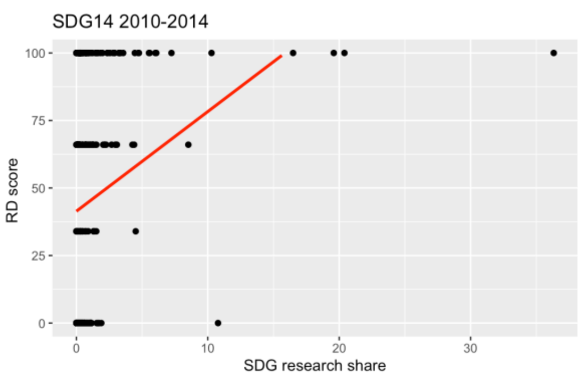

SDG11 2015-2020
Average RD regions
SDG research share
0 25 50 75 100
0 1 2 3 4 5

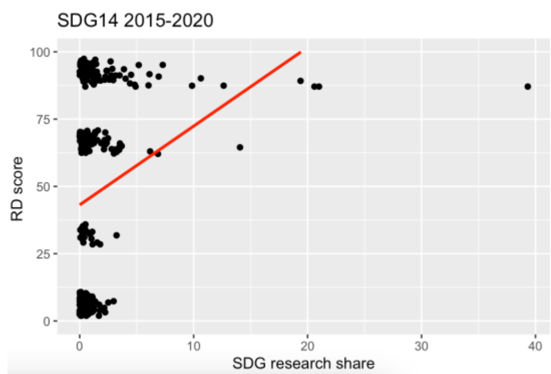

SDG12 2010-2014
RD score
SDG research share
0 25 50 75 100
0 1 2 3 4 5

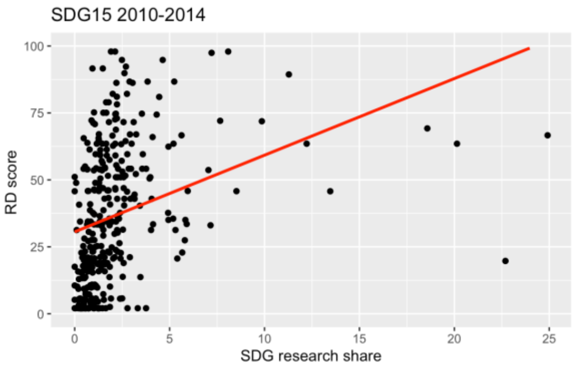

SDG12 2015-2020
RD score
SDG research share
0 25 50 75 100
0 2 4 6 8

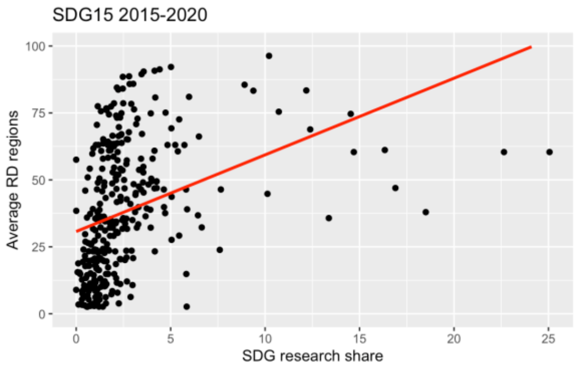

SDG13 2010-2014
RD score
SDG research share
0 25 50 75 100
0 5 10 15


SDG13 2015-2020
Average RD regions
SDG research share
0 25 50 75 100
0 5 10 15

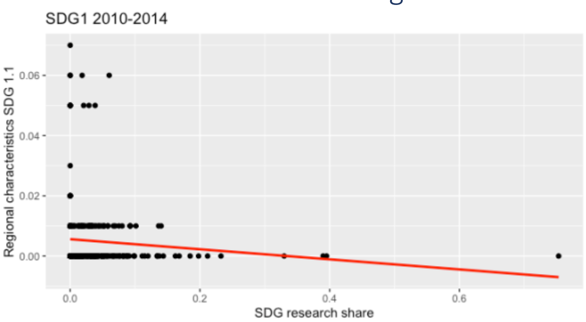

SDG14 2010-2014
RD score
SDG research share
0 25 50 75 100
0 10 20 30

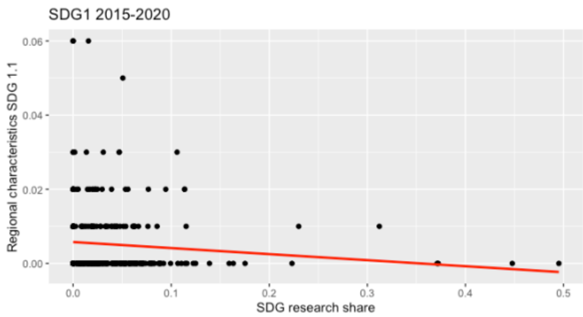

SDG14 2015-2020
RD score
SDG research share
0 25 50 75 100
0 10 20 30 40

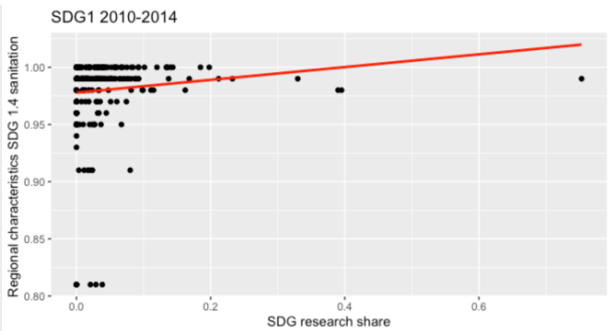


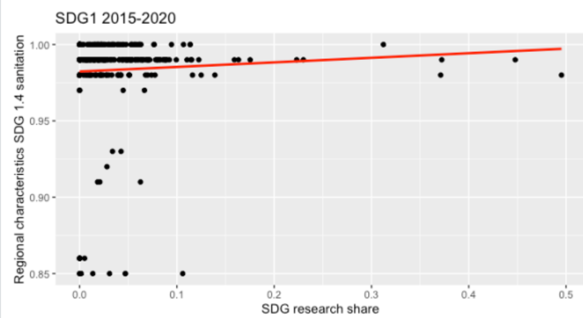


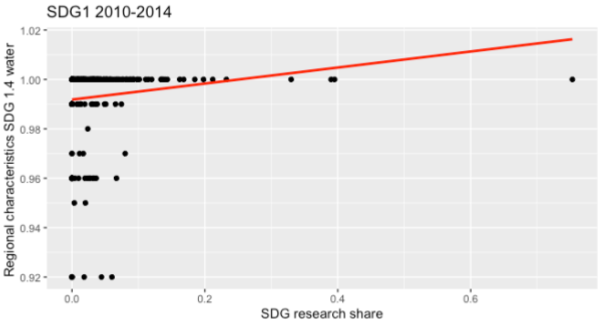
A3: SDG research share and regional characteristics (RC)

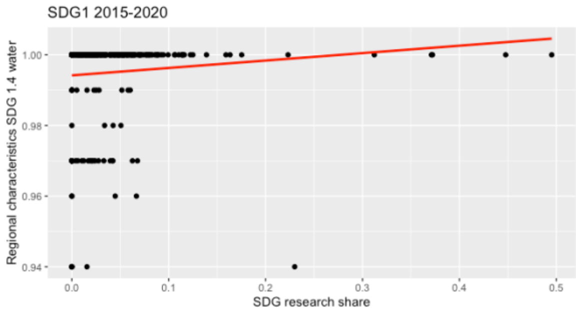


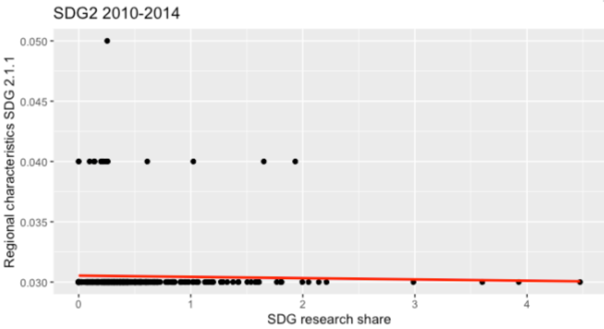


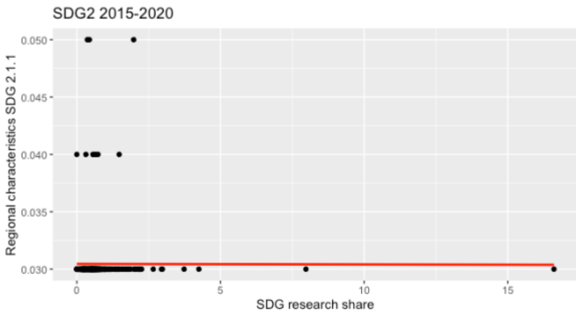


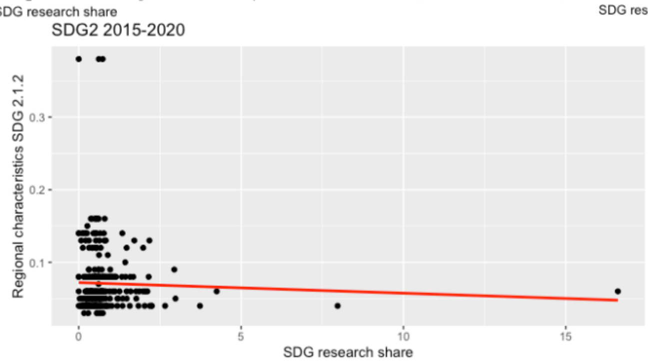


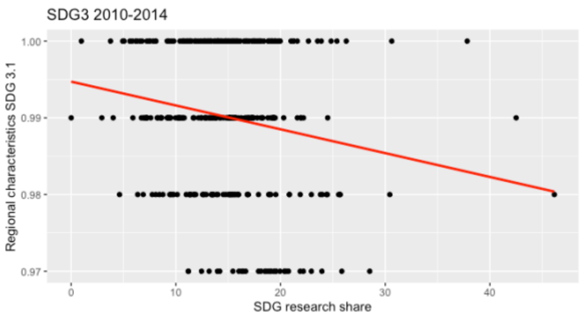

SDG1 2010-2014
Regional characteristics SDG 1.4 water
SDG research share

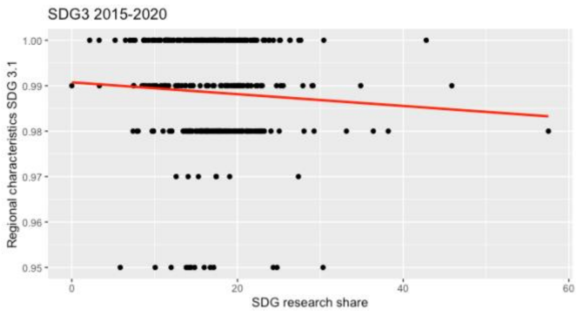

SDG1 2015-2020
Regional characteristics SDG 1.4 water
SDG research share

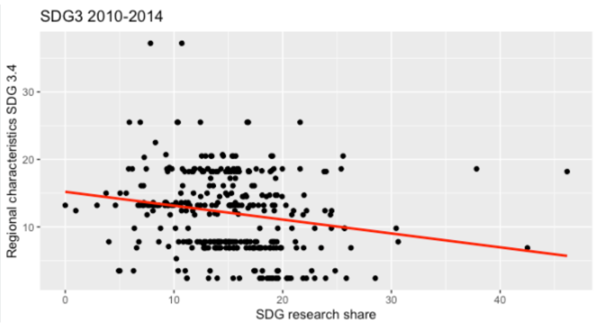

SDG2 2010-2014
Regional characteristics SDG 2.1.1
SDG research share


SDG2 2015-2020
Regional characteristics SDG 2.1.1
SDG research share

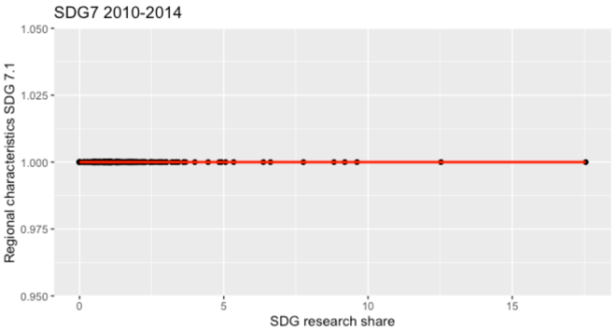

SDG2 2015-2020
Regional characteristics SDG 2.1.2
SDG research share

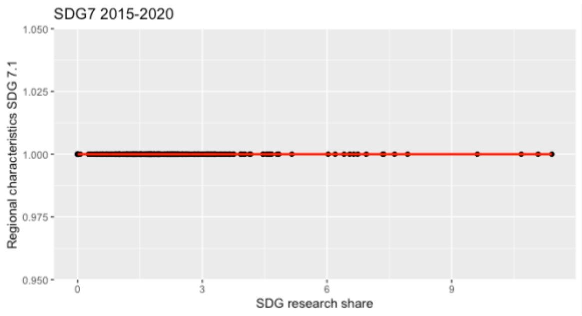

SDG3 2010-2014
Regional characteristics SDG 3.1
SDG research share

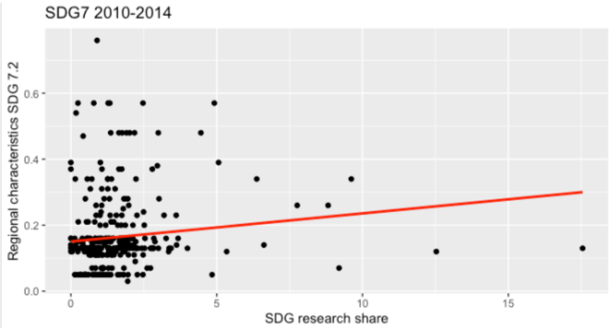

SDG3 2015-2020
Regional characteristics SDG 3.1
SDG research share

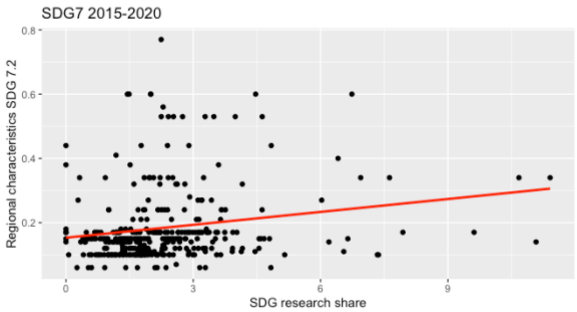

SDG3 2010-2014
Regional characteristics SDG 3.4
SDG research share


SDG3 2015-2020
Regional characteristics SDG 3.2
SDG research share

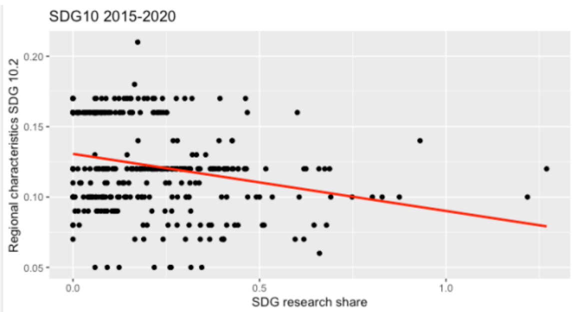

SDG7 2010-2014
Regional characteristics SDG 7.1
1.050
1.025
1.000
0.975
0.950
0
5
10
15
SDG research share

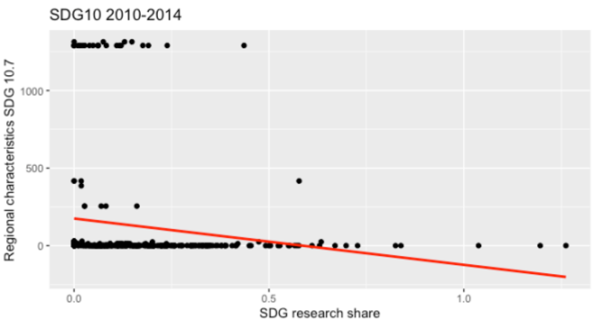

SDG7 2015-2020
Regional characteristics SDG 7.1
1.050
1.025
1.000
0.975
0.950
0
3
6
9
SDG research share

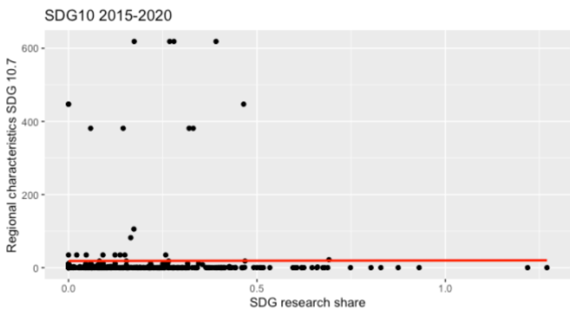

SDG7 2010-2014
Regional characteristics SDG 7.2
0.6
0.4
0.2
0.0
0
5
10
15
SDG research share

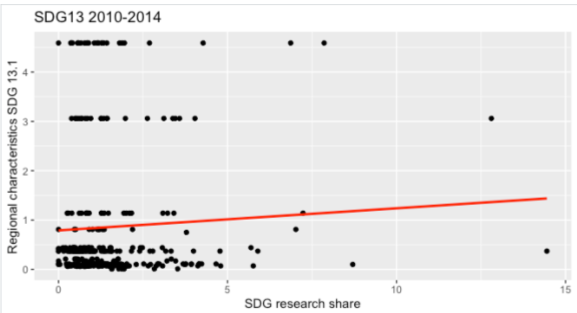

SDG7 2015-2020
Regional characteristics SDG 7.2
0.8
0.6
0.4
0.2
0
3
6
9
SDG research share

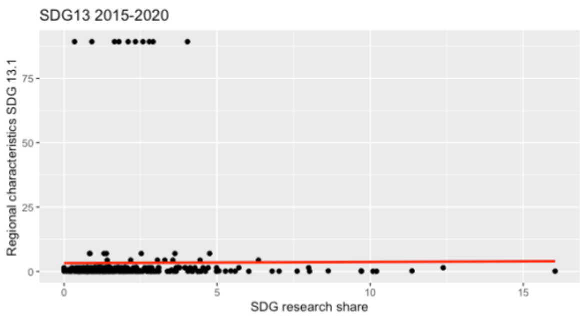

SDG10 2010-2014
Regional characteristics SDG 10.2
0.20
0.15
0.10
0.0
0.5
1.0
SDG research share

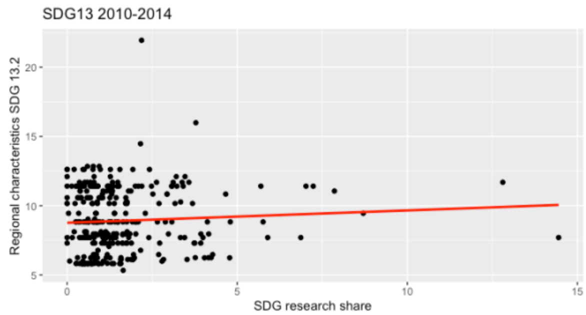

SDG10 2015-2020
Regional characteristics SDG 10.2
0.20
0.15
0.10
0.05
0.0
0.5
1.0
SDG research share

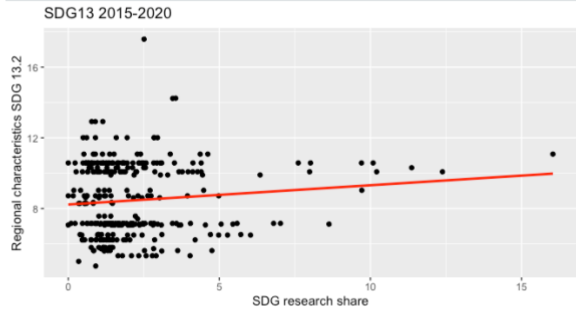

SDG10 2010-2014
Regional characteristics SDG 10.7
1000
500
0
0.0
0.5
1.0
SDG research share

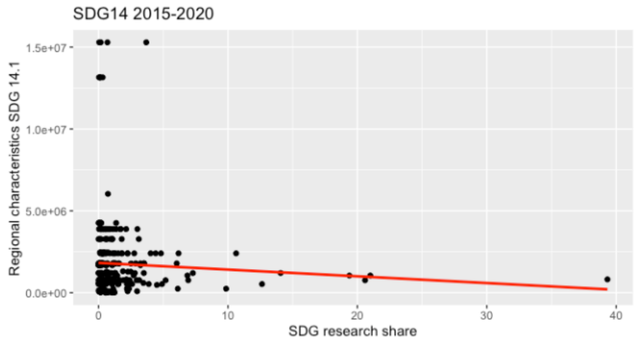

SDG10 2015-2020
Regional characteristics SDG 10.7
600
400
200
0
0.0
0.5
1.0
SDG research share

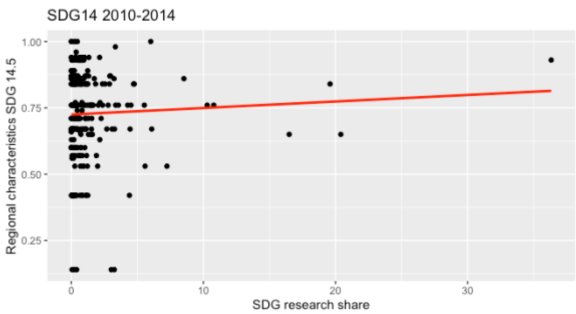

SDG13 2010-2014
Regional characteristics SDG 13.1
4
3
2
1
0
0
5
10
15
SDG research share

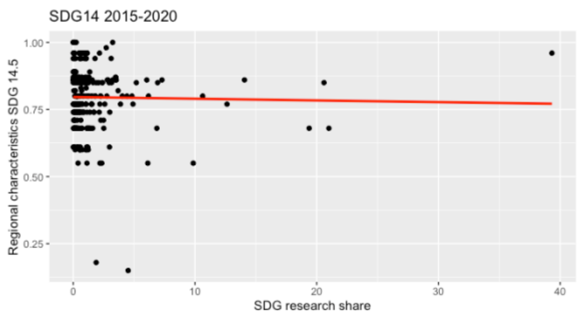

SDG13 2015-2020
Regional characteristics SDG 13.1
75
50
25
0
0
5
10
15
SDG research share

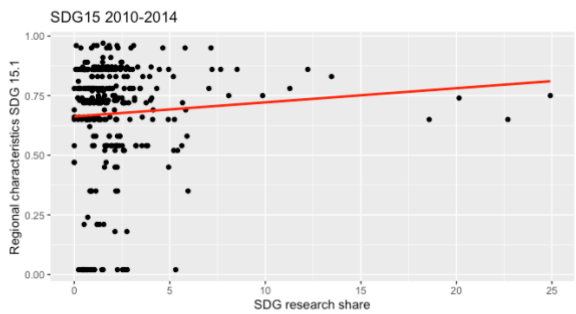

SDG13 2010-2014
Regional characteristics SDG 13.2
SDG research share

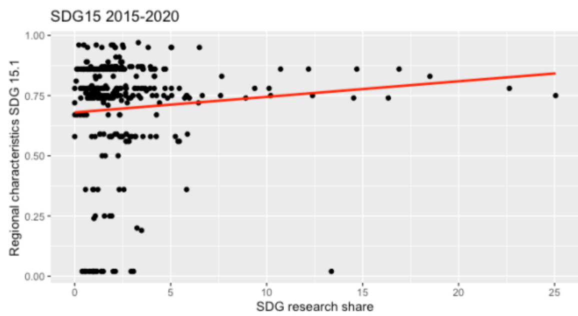

SDG13 2015-2020
Regional characteristics SDG 13.2
SDG research share

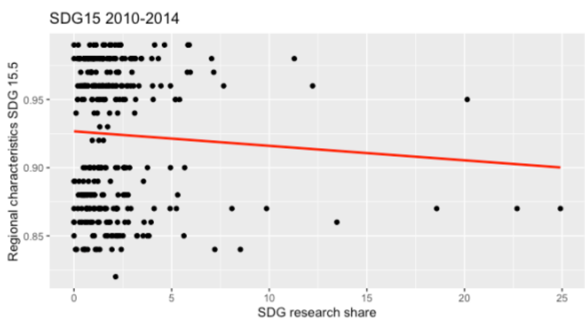

SDG14 2015-2020
Regional characteristics SDG 14.1
SDG research share

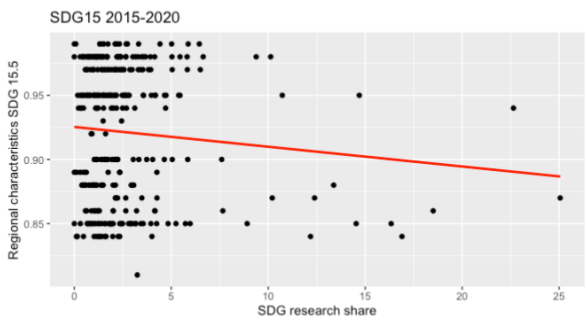

SDG14 2010-2014
Regional characteristics SDG 14.5
SDG research share


SDG14 2015-2020
Regional characteristics SDG 14.5
SDG research share


SDG15 2010-2014
Regional characteristics SDG 15.1
SDG research share


SDG15 2015-2020
Regional characteristics SDG 15.1
SDG research share


SDG15 2010-2014
Regional characteristics SDG 15.5
SDG research share


SDG15 2015-2020
Regional characteristics SDG 15.5
SDG research share

Appendix B: Correlation matrices

All SDGs included

2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	**1**	1	0.257	-0.120	-0.106	0.023	-0.003
KCI_1	**2**	0.257	1	0.252	-0.172	0.502	0.542
RD_SDG	**3**	-0.120	0.252	1	-0.032	0.152	0.274
Population	**4**	-0.106	-0.172	-0.032	1	0.035	-0.032
GDP	**5**	0.023	0.502	0.152	0.035	1	0.632
Education	**6**	-0.003	0.542	0.274	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.181	-0.068	-0.129	-0.070	-0.032
KCI_2	2	0.181	1	0.308	-0.180	0.528	0.579
RD_SDG	3	-0.068	0.308	1	-0.071	0.134	0.241
Population	4	-0.129	-0.180	-0.071	1	0.048	-0.028
GDP	5	-0.070	0.528	0.134	0.048	1	0.692
Education	6	-0.032	0.579	0.241	-0.028	0.692	1

SDG 1

2010-2014

Variable	#	1	2	3	4	5	6	7	8	9
SDG_share	**1**	1	0.255	0.314	-0.103	0.127	0.112	-0.030	0.237	0.179
KCI	**2**	0.255	1	0.572	-0.282	0.335	0.280	-0.172	0.502	0.542
RD_SDG	**3**	0.314	0.572	1	-0.185	0.093	0.104	-0.023	0.358	0.434
Loc_1	**4**	-0.103	-0.282	-0.185	1	-0.299	-0.628	0.061	-0.275	-0.275
Loc_2	**5**	0.127	0.335	0.093	-0.299	1	0.350	-0.077	0.383	0.362
Loc_3	**6**	0.112	0.280	0.104	-0.628	0.350	1	-0.099	0.346	0.282
Population	**7**	-0.030	-0.172	-0.023	0.061	-0.077	-0.099	1	0.035	-0.032
GDP	**8**	0.237	0.502	0.358	-0.275	0.383	0.346	0.035	1	0.632
Education	**9**	0.179	0.542	0.434	-0.275	0.362	0.282	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8	9
SDG_share	**1**	1	0.299	0.401	-0.095	0.097	0.059	0.007	0.258	0.256
KCI	**2**	0.299	1	0.573	-0.309	0.297	0.172	-0.180	0.528	0.579
RD_SDG	**3**	0.401	0.573	1	-0.161	0.081	0.063	-0.075	0.342	0.417
Loc_1	**4**	-0.095	-0.309	-0.161	1	-0.295	-0.389	0.039	-0.265	-0.325
Loc_2	**5**	0.097	0.297	0.081	-0.295	1	0.170	-0.073	0.335	0.326
Loc_3	**6**	0.059	0.172	0.063	-0.389	0.170	1	0.007	0.175	0.144
Population	**7**	0.007	-0.180	-0.075	0.039	-0.073	0.007	1	0.048	-0.028
GDP	**8**	0.258	0.528	0.342	-0.265	0.335	0.175	0.048	1	0.692
Education	**9**	0.256	0.579	0.417	-0.325	0.326	0.144	-0.028	0.692	1

SDG2

2010-2014

Variable	#	1	2	3	4	5	6	7
SDG_share	1	1	-0.020	0.148	-0.019	-0.109	-0.141	-0.037
KCI	2	-0.020	1	-0.174	-0.114	-0.172	0.502	0.542
RD_SDG	3	0.148	-0.174	1	-0.002	0.134	0.014	0.042
Loc_1	4	-0.019	-0.114	-0.002	1	-0.084	-0.202	-0.091
Population	5	-0.109	-0.172	0.134	-0.084	1	0.035	-0.032
GDP	6	-0.141	0.502	0.014	-0.202	0.035	1	0.632
Education	7	-0.037	0.542	0.042	-0.091	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	-0.084	0.106	-0.002	-0.038	-0.085	-0.105	-0.058
KCI	2	-0.084	1	0.044	-0.133	-0.262	-0.180	0.528	0.579
RD_SDG	3	0.106	0.044	1	0.075	0.054	0.084	0.053	0.085
Loc_1	4	-0.002	-0.133	0.075	1	0.281	-0.051	-0.116	-0.111
Loc_2	5	-0.038	-0.262	0.054	0.281	1	-0.067	-0.436	-0.323
Population	6	-0.085	-0.180	0.084	-0.051	-0.067	1	0.048	-0.028
GDP	7	-0.105	0.528	0.053	-0.116	-0.436	0.048	1	0.692
Education	8	-0.058	0.579	0.085	-0.111	-0.323	-0.028	0.692	1

SDG3

2010-2014

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.255	-0.193	-0.176	-0.195	0.035	0.087	0.054
KCI	2	0.255	1	0.435	0.166	-0.200	-0.172	0.502	0.542
RD_SDG	3	-0.193	0.435	1	0.155	-0.058	-0.072	0.180	0.331
Loc_1	4	-0.176	0.166	0.155	1	0.057	-0.131	0.059	0.192
Loc_2	5	-0.195	-0.200	-0.058	0.057	1	-0.122	0.075	0.180
Population	6	0.035	-0.172	-0.072	-0.131	-0.122	1	0.035	-0.032
GDP	7	0.087	0.502	0.180	0.059	0.075	0.035	1	0.632
Education	8	0.054	0.542	0.331	0.192	0.180	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.143	-0.209	-0.068	-0.109	0.061	-0.009	0.025
KCI	2	0.143	1	0.465	-0.153	0.020	-0.180	0.528	0.579
RD_SDG	3	-0.209	0.465	1	-0.039	0.009	-0.107	0.212	0.358
Loc_1	4	-0.068	-0.153	-0.039	1	0.078	-0.039	-0.127	-0.091
Loc_2	5	-0.109	0.020	0.009	0.078	1	-0.179	0.180	0.278
Population	6	0.061	-0.180	-0.107	-0.039	-0.179	1	0.048	-0.028
GDP	7	-0.009	0.528	0.212	-0.127	0.180	0.048	1	0.692
Education	8	0.025	0.579	0.358	-0.091	0.278	-0.028	0.692	1

SDG4

2010-2014

Variable	#	1	2	3	4	5
SDG_share	1	1	0.159	-0.091	-0.107	-0.075
KCI	2	0.159	1	-0.172	0.502	0.542
Population	3	-0.091	-0.172	1	0.035	-0.032
GDP	4	-0.107	0.502	0.035	1	0.632
Education	5	-0.075	0.542	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.208	0.142	-0.163	-0.047	0.004
KCI	2	0.208	1	0.504	-0.180	0.528	0.579
RD_SDG	3	0.142	0.504	1	0.018	0.309	0.390
Population	4	-0.163	-0.180	0.018	1	0.048	-0.028
GDP	5	-0.047	0.528	0.309	0.048	1	0.692
Education	6	0.004	0.579	0.390	-0.028	0.692	1

SDG5

2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.607	0.554	-0.078	0.259	0.304
KCI	2	0.607	1	0.599	-0.172	0.502	0.542
RD_SDG	3	0.554	0.599	1	0.048	0.399	0.446
Population	4	-0.078	-0.172	0.048	1	0.035	-0.032
GDP	5	0.259	0.502	0.399	0.035	1	0.632
Education	6	0.304	0.542	0.446	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.629	0.544	-0.128	0.236	0.336
KCI	2	0.629	1	0.600	-0.180	0.528	0.579
RD_SDG	3	0.544	0.600	1	0.047	0.389	0.443
Population	4	-0.128	-0.180	0.047	1	0.048	-0.028
GDP	5	0.236	0.528	0.389	0.048	1	0.692
Education	6	0.336	0.579	0.443	-0.028	0.692	1

SDG6

2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	1	1	-0.220	0.214	-0.021	-0.230	-0.138
KCI	2	-0.220	1	-0.199	-0.172	0.502	0.542
RD_SDG	3	0.214	-0.199	1	-0.036	-0.051	-0.001
Population	4	-0.021	-0.172	-0.036	1	0.035	-0.032
GDP	5	-0.230	0.502	-0.051	0.035	1	0.632
Education	6	-0.138	0.542	-0.001	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	-0.226	0.244	-0.043	-0.238	-0.182
KCI	2	-0.226	1	-0.112	-0.180	0.528	0.579
RD_SDG	3	0.244	-0.112	1	-0.040	-0.029	-0.006
Population	4	-0.043	-0.180	-0.040	1	0.048	-0.028
GDP	5	-0.238	0.528	-0.029	0.048	1	0.692
Education	6	-0.182	0.579	-0.006	-0.028	0.692	1

SDG7

2010-2014

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	-0.186	0.302	NA	0.123	-0.071	-0.179	-0.162
KCI	2	-0.186	1	-0.240	NA	-0.031	-0.172	0.502	0.542
RD_SDG	3	0.302	-0.240	1	NA	0.111	-0.008	-0.161	-0.076
Loc_1	4	NA	NA	NA	1	NA	NA	NA	NA
Loc_2	5	0.123	-0.031	0.111	NA	1	-0.146	0.011	-0.098
Population	6	-0.071	-0.172	-0.008	NA	-0.146	1	0.035	-0.032
GDP	7	-0.179	0.502	-0.161	NA	0.011	0.035	1	0.632
Education	8	-0.162	0.542	-0.076	NA	-0.098	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	-0.140	0.428	NA	0.190	-0.071	-0.087	-0.081
KCI	2	-0.140	1	-0.230	NA	0.106	-0.180	0.528	0.579
RD_SDG	3	0.428	-0.230	1	NA	0.101	0.002	-0.194	-0.135
Loc_1	4	NA	NA	NA	1	NA	NA	NA	NA
Loc_2	5	0.190	0.106	0.101	NA	1	-0.173	0.023	0.106
Population	6	-0.071	-0.180	0.002	NA	-0.173	1	0.048	-0.028
GDP	7	-0.087	0.528	-0.194	NA	0.023	0.048	1	0.692
Education	8	-0.081	0.579	-0.135	NA	0.106	-0.028	0.692	1

SDG8
2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.166	0.300	-0.152	0.088	-0.005
KCI	2	0.166	1	0.083	-0.172	0.502	0.542
RD_SDG	3	0.300	0.083	1	-0.013	0.045	0.090
Population	4	-0.152	-0.172	-0.013	1	0.035	-0.032
GDP	5	0.088	0.502	0.045	0.035	1	0.632
Education	6	-0.005	0.542	0.090	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.081	0.410	-0.167	-0.128	-0.062
KCI	2	0.081	1	0.032	-0.180	0.528	0.579
RD_SDG	3	0.410	0.032	1	-0.027	-0.067	-0.019
Population	4	-0.167	-0.180	-0.027	1	0.048	-0.028
GDP	5	-0.128	0.528	-0.067	0.048	1	0.692
Education	6	-0.062	0.579	-0.019	-0.028	0.692	1

SDG9
2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.192	0.415	-0.123	0.114	0.067
KCI	2	0.192	1	-0.056	-0.172	0.502	0.542
RD_SDG	3	0.415	-0.056	1	-0.010	0.007	0.017
Population	4	-0.123	-0.172	-0.010	1	0.035	-0.032
GDP	5	0.114	0.502	0.007	0.035	1	0.632
Education	6	0.067	0.542	0.017	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	0.130	0.423	-0.137	-0.041	0.032
KCI	2	0.130	1	-0.086	-0.180	0.528	0.579
RD_SDG	3	0.423	-0.086	1	-0.012	-0.094	-0.025
Population	4	-0.137	-0.180	-0.012	1	0.048	-0.028
GDP	5	-0.041	0.528	-0.094	0.048	1	0.692
Education	6	0.032	0.579	-0.025	-0.028	0.692	1

SDG10

2010-2014

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.543	0.496	-0.239	-0.152	-0.086	0.335	0.348
KCI	2	0.543	1	0.480	-0.222	-0.307	-0.172	0.502	0.542
RD_SDG	3	0.496	0.480	1	-0.082	-0.111	0.077	0.338	0.398
Loc_1	4	-0.239	0.222	-0.082	1	0.416	0.168	-0.376	-0.310
Loc_2	5	-0.152	-0.307	-0.111	0.416	1	0.073	-0.350	-0.395
Population	6	-0.086	-0.172	0.077	0.168	0.073	1	0.035	-0.032
GDP	7	0.335	0.502	0.338	-0.376	-0.350	0.035	1	0.632
Education	8	0.348	0.542	0.398	-0.310	-0.395	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.561	0.498	-0.252	0.003	-0.090	0.378	0.392
KCI	2	0.561	1	0.523	-0.240	-0.104	-0.180	0.528	0.579
RD_SDG	3	0.498	0.523	1	-0.063	0.005	0.016	0.331	0.385
Loc_1	4	-0.252	-0.240	-0.063	1	0.065	0.189	-0.271	-0.334
Loc_2	5	0.003	-0.104	0.005	0.065	1	-0.079	-0.194	-0.128
Population	6	-0.090	-0.180	0.016	0.189	-0.079	1	0.048	-0.028
GDP	7	0.378	0.528	0.331	-0.271	-0.194	0.048	1	0.692
Education	8	0.392	0.579	0.385	-0.334	-0.128	-0.028	0.692	1

SDG11

2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	1	1	-0.023	0.127	-0.113	-0.152	-0.109
KCI	2	-0.023	1	0.095	-0.172	0.502	0.542
RD_SDG	3	0.127	0.095	1	0.080	0.108	0.149
Population	4	-0.113	-0.172	0.080	1	0.035	-0.032
GDP	5	-0.152	0.502	0.108	0.035	1	0.632
Education	6	-0.109	0.542	0.149	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	-0.155	0.276	-0.118	-0.195	-0.128
KCI	2	-0.155	1	0.056	-0.180	0.528	0.579
RD_SDG	3	0.276	0.056	1	0.053	-0.033	0.032
Population	4	-0.118	-0.180	0.053	1	0.048	-0.028
GDP	5	-0.195	0.528	-0.033	0.048	1	0.692
Education	6	-0.128	0.579	0.032	-0.028	0.692	1

SDG12

2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	1	1	-0.139	0.417	-0.111	-0.193	-0.100
KCI	2	-0.139	1	-0.240	-0.172	0.502	0.542
RD_SDG	3	0.417	-0.240	1	-0.014	-0.150	-0.059
Population	4	-0.111	-0.172	-0.014	1	0.035	-0.032
GDP	5	-0.193	0.502	-0.150	0.035	1	0.632
Education	6	-0.100	0.542	-0.059	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6
SDG_share	1	1	-0.193	0.464	-0.092	-0.174	-0.120
KCI	2	-0.193	1	-0.208	-0.180	0.528	0.579
RD_SDG	3	0.464	-0.208	1	-0.005	-0.143	-0.069
Population	4	-0.092	-0.180	-0.005	1	0.048	-0.028
GDP	5	-0.174	0.528	-0.143	0.048	1	0.692
Education	6	-0.120	0.579	-0.069	-0.028	0.692	1

SDG13

2010-2014

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.108	0.272	0.054	0.063	-0.197	0.106	0.065
KCI	2	0.108	1	-0.224	-0.030	0.219	-0.172	0.502	0.542
RD_SDG	3	0.272	-0.224	1	-0.070	-0.084	-0.066	-0.243	-0.099
Loc_1	4	0.054	-0.030	-0.070	1	0.133	-0.114	0.133	0.132
Loc_2	5	0.063	0.219	-0.084	0.133	1	-0.160	0.329	0.275
Population	6	-0.197	-0.172	-0.066	-0.114	-0.160	1	0.035	-0.032
GDP	7	0.106	0.502	-0.243	0.133	0.329	0.035	1	0.632
Education	8	0.065	0.542	-0.099	0.132	0.275	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.147	0.307	0.006	0.105	-0.199	0.091	0.049
KCI	2	0.147	1	-0.168	0.164	0.020	-0.180	0.528	0.579
RD_SDG	3	0.307	-0.168	1	-0.021	-0.039	-0.091	-0.231	-0.107
Loc_1	4	0.006	0.164	-0.021	1	-0.288	-0.092	0.073	0.145
Loc_2	5	0.105	0.020	-0.039	-0.288	1	-0.065	0.188	0.132
Population	6	-0.199	-0.180	-0.091	-0.092	-0.065	1	0.048	-0.028
GDP	7	0.091	0.528	-0.231	0.073	0.188	0.048	1	0.692
Education	8	0.049	0.579	-0.107	0.145	0.132	-0.028	0.692	1

SDG 14

2010-2014

Variable	#	1	2	3	4	5	6	7
SDG_share	1	1	0.112	0.258	0.043	-0.144	-0.018	-0.009
KCI	2	0.112	1	0.044	0.218	-0.172	0.502	0.542
RD_SDG	3	0.258	0.044	1	0.082	-0.188	0.023	0.104
Loc_2	4	0.043	0.218	0.082	1	-0.088	0.135	0.360
Population	5	-0.144	-0.172	-0.188	-0.088	1	0.035	-0.032
GDP	6	-0.018	0.502	0.023	0.135	0.035	1	0.632
Education	7	-0.009	0.542	0.104	0.360	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.146	0.257	-0.051	-0.019	-0.155	0.026	-0.011
KCI	2	0.146	1	0.104	-0.246	0.065	-0.180	0.528	0.579
RD_SDG	3	0.257	0.104	1	-0.087	0.007	-0.247	0.021	0.071
Loc_1	4	-0.051	-0.246	-0.087	1	0.027	0.054	-0.197	-0.194
Loc_2	5	-0.019	0.065	0.007	0.027	1	-0.040	0.023	0.154
Population	6	-0.155	-0.180	-0.247	0.054	-0.040	1	0.048	-0.028
GDP	7	0.026	0.528	0.021	-0.197	0.023	0.048	1	0.692
Education	8	-0.011	0.579	0.071	-0.194	0.154	-0.028	0.692	1

SDG 15

2010-2014

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.021	0.319	0.069	-0.059	-0.212	-0.073	-0.064
KCI	2	0.021	1	0.103	0.220	0.330	-0.172	0.502	0.542
RD_SDG	3	0.319	0.103	1	0.086	-0.095	-0.103	-0.026	0.073
Loc_1	4	0.069	0.220	0.086	1	0.361	-0.163	0.197	0.345
Loc_2	5	-0.059	0.330	-0.095	0.361	1	-0.133	0.421	0.431
Population	6	-0.212	-0.172	-0.103	-0.163	-0.133	1	0.035	-0.032
GDP	7	-0.073	0.502	-0.026	0.197	0.421	0.035	1	0.632
Education	8	-0.064	0.542	0.073	0.345	0.431	-0.032	0.632	1

2015-2020

Variable	#	1	2	3	4	5	6	7	8
SDG_share	1	1	0.059	0.370	0.083	-0.091	-0.235	-0.081	-0.107
KCI	2	0.059	1	0.121	0.249	0.305	-0.180	0.528	0.579
RD_SDG	3	0.370	0.121	1	0.085	-0.072	-0.121	-0.006	0.069
Loc_1	4	0.083	0.249	0.085	1	0.290	-0.183	0.234	0.288
Loc_2	5	-0.091	0.305	-0.072	0.290	1	-0.124	0.404	0.349
Population	6	-0.235	-0.180	-0.121	-0.183	-0.124	1	0.048	-0.028
GDP	7	-0.081	0.528	-0.006	0.234	0.404	0.048	1	0.692
Education	8	-0.107	0.579	0.069	0.288	0.349	-0.028	0.692	1

SDG16

2010-2014

Variable	#	1	2	3	4	5	6
SDG_share	**1**	1	0.641	0.491	-0.029	0.275	0.365
KCI	**2**	0.641	1	0.467	-0.172	0.502	0.542
RD_SDG	**3**	0.491	0.467	1	0.062	0.266	0.325
Population	**4**	-0.029	-0.172	0.062	1	0.035	-0.032
GDP	**5**	0.275	0.502	0.266	0.035	1	0.632
Education	**6**	0.365	0.542	0.325	-0.032	0.632	1

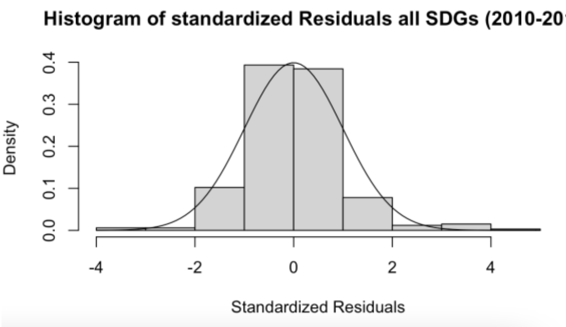
2015-2020

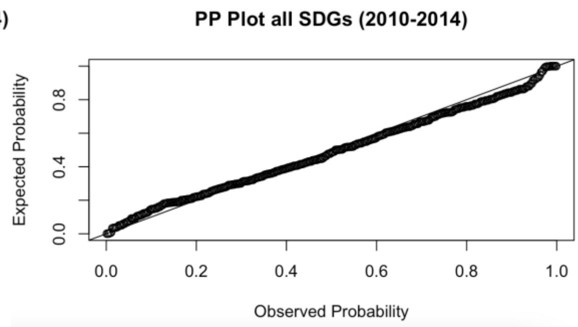
Variable	#	1	2	3	4	5	6
SDG_share	**1**	1	0.619	0.505	-0.082	0.313	0.362
KCI	**2**	0.619	1	0.565	-0.180	0.528	0.579
RD_SDG	**3**	0.505	0.565	1	-0.038	0.293	0.367
Population	**4**	-0.082	-0.180	-0.038	1	0.048	-0.028
GDP	**5**	0.313	0.528	0.293	0.048	1	0.692
Education	**6**	0.362	0.579	0.367	-0.028	0.692	1

Appendix C: Histograms and PP Plots Residuals

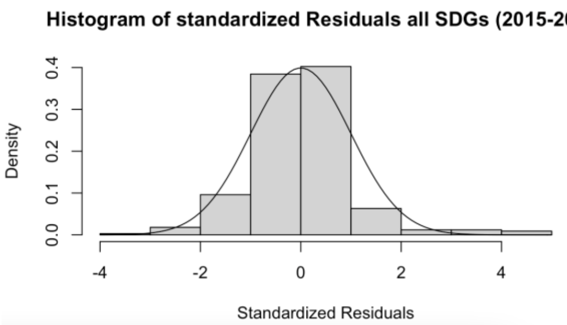
SDGs combined

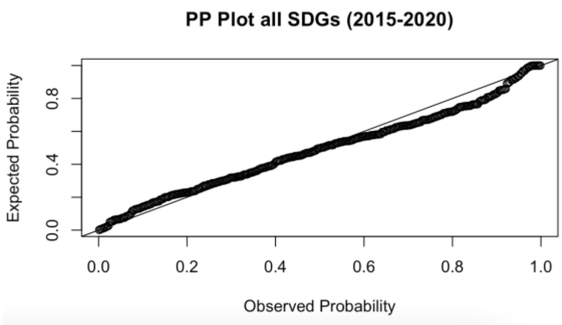
2010-2014





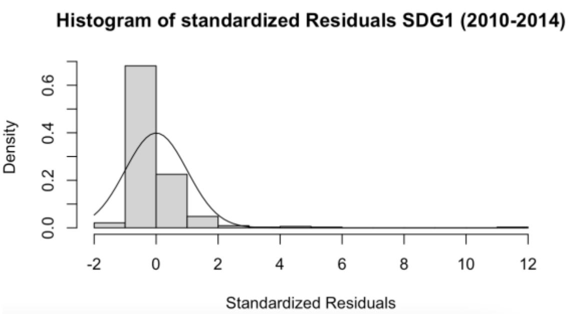
2015-2020

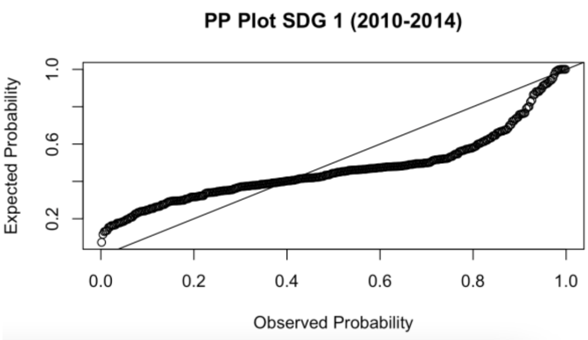




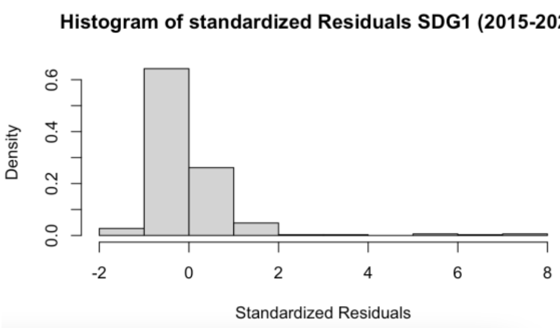
SDG1

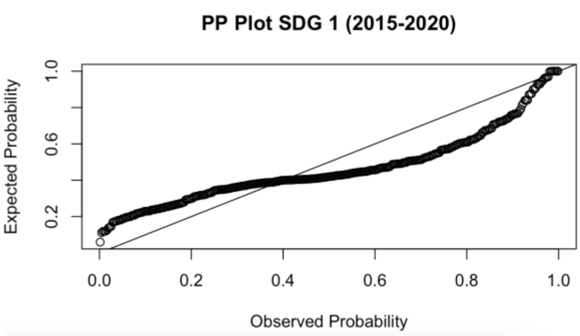
2010-2014





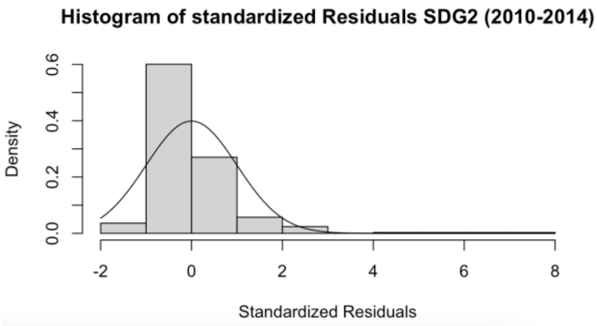
2015-2020

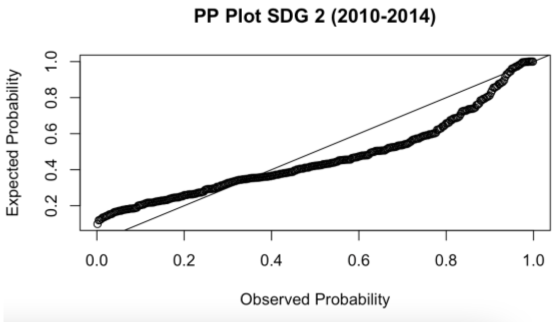




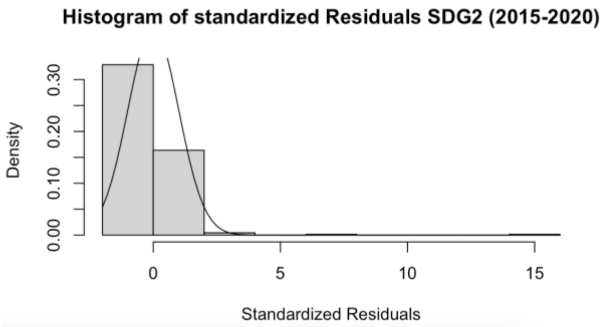
SDG2

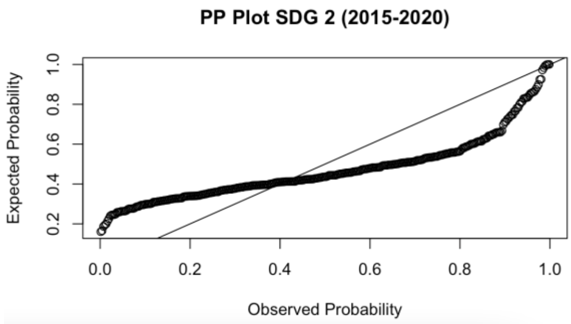
2010-2014





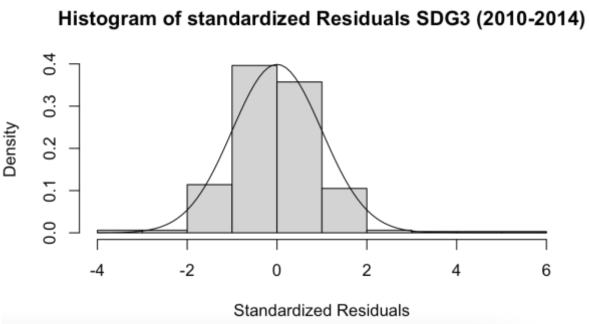
2015-2020

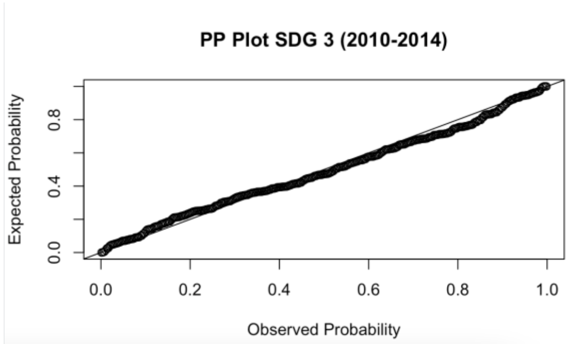




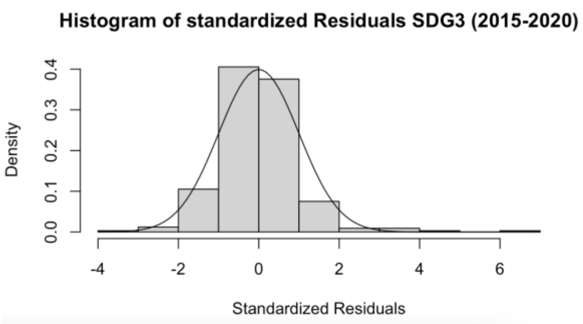
SDG3

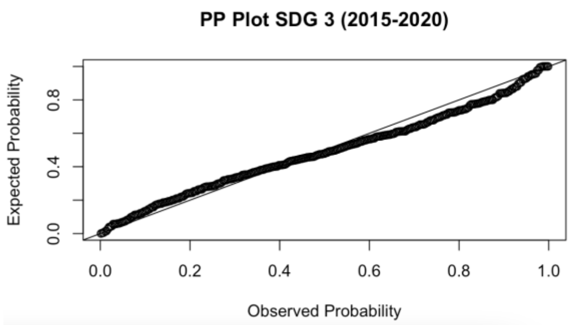
2010-2014



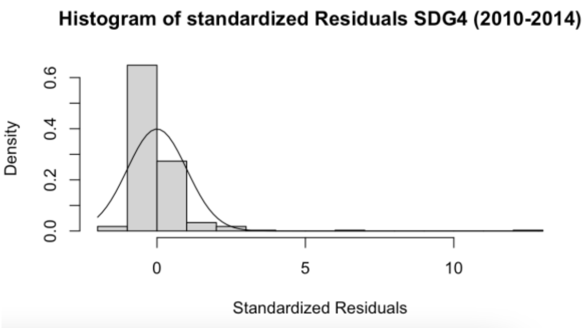


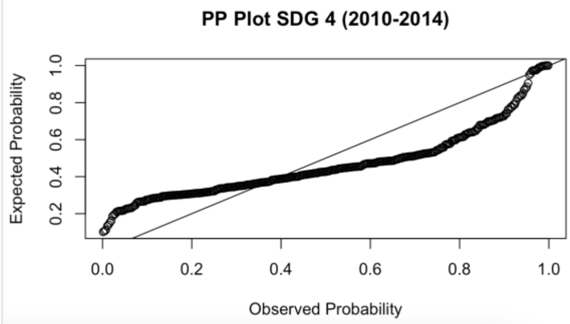
2015-2020



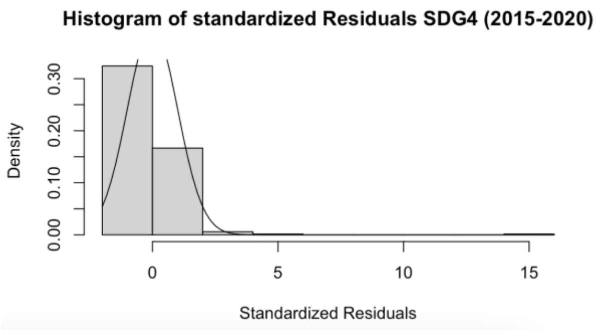


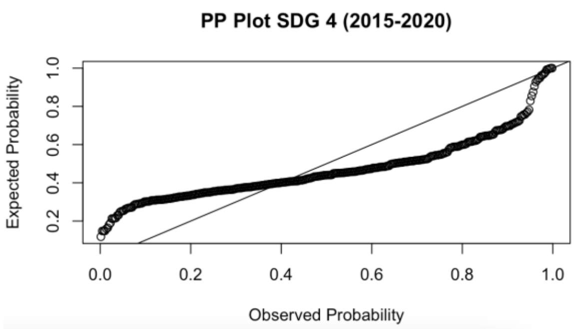
SDG4
2010-2014



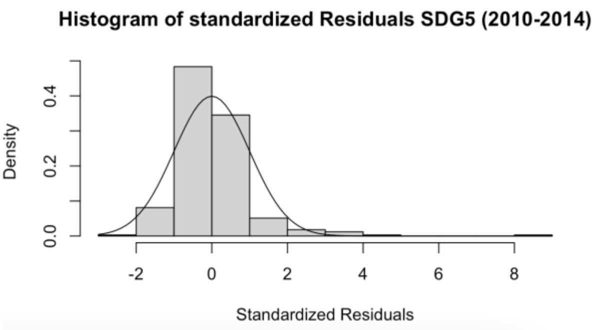


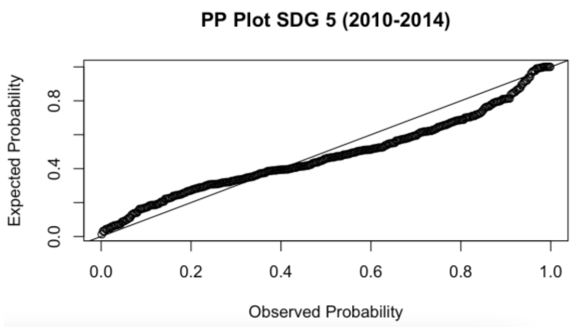
2015-2020



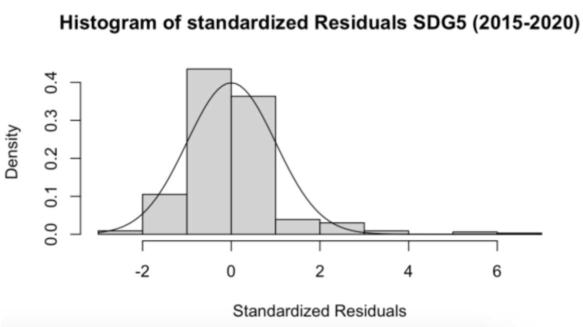


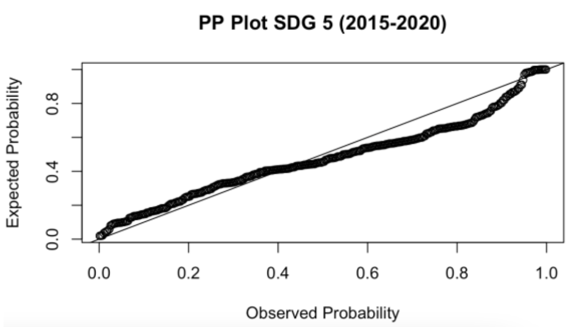
SDG5
2010-2014



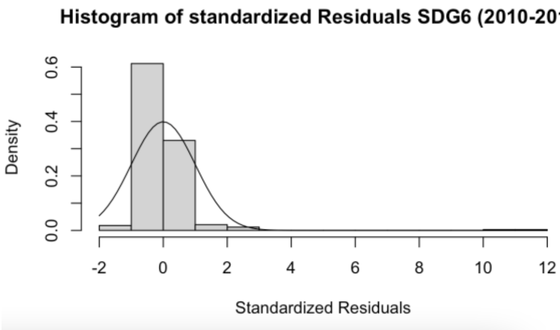


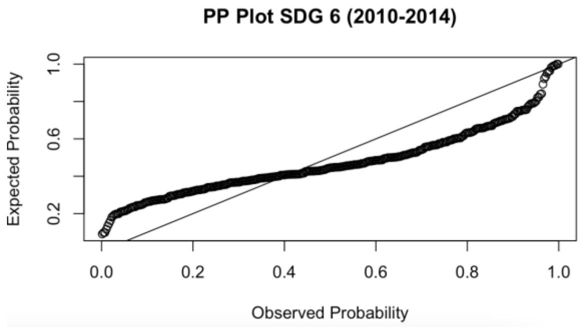
2015-2020



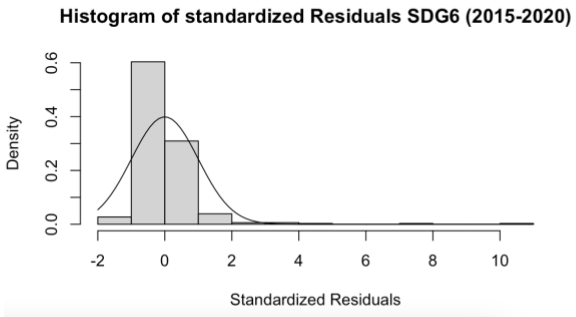


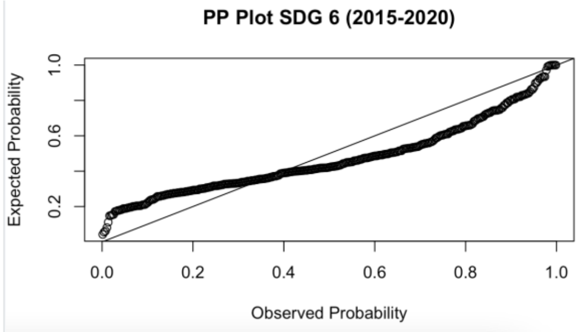
SDG6
2010-2014



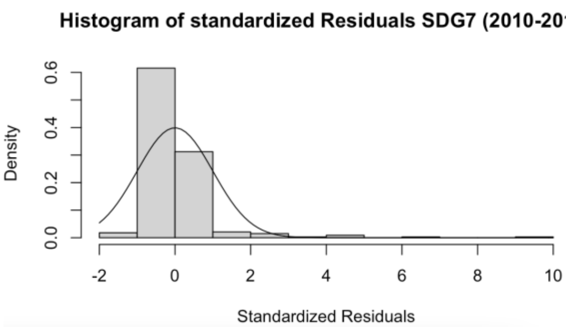


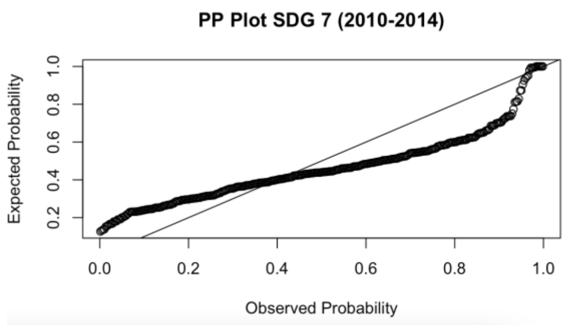
2015-2020



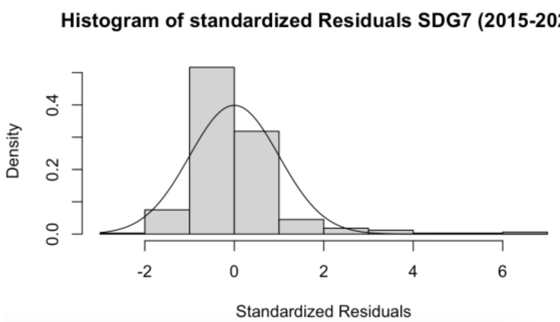


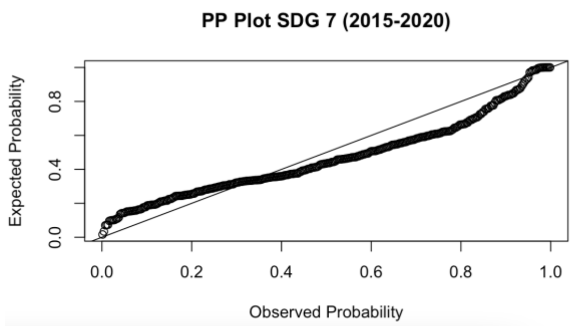
SDG7
2010-2014



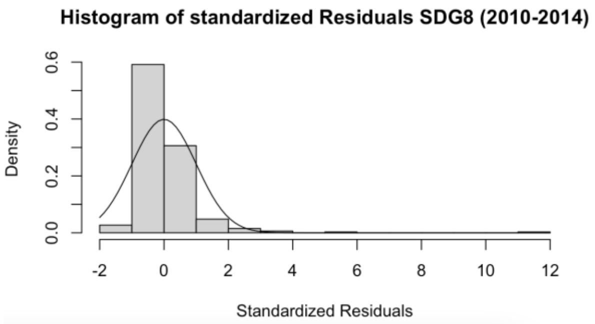


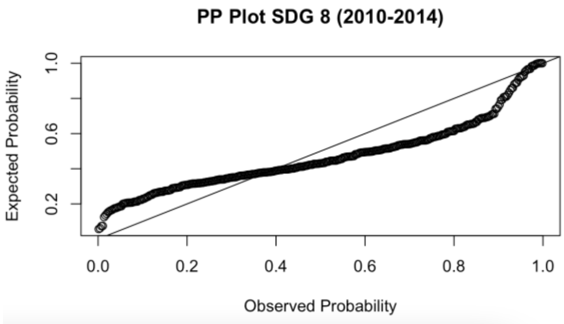
2015-2020



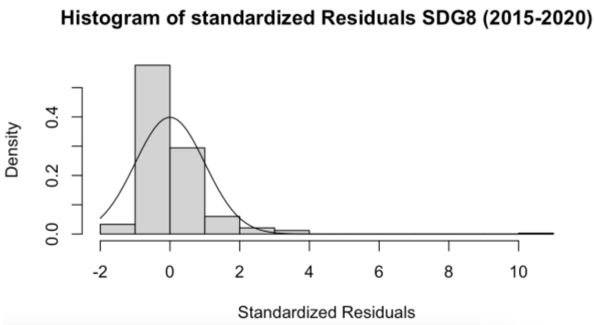


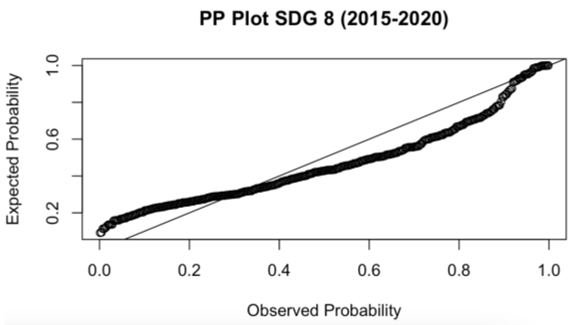
SDG8
2010-2014



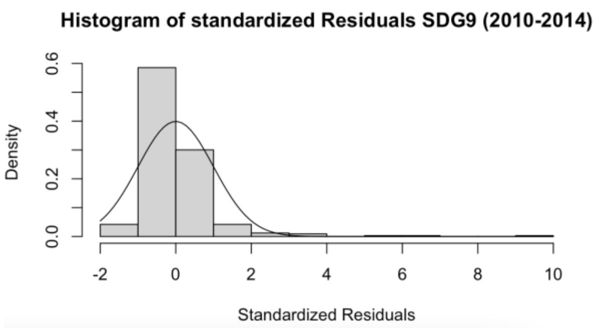


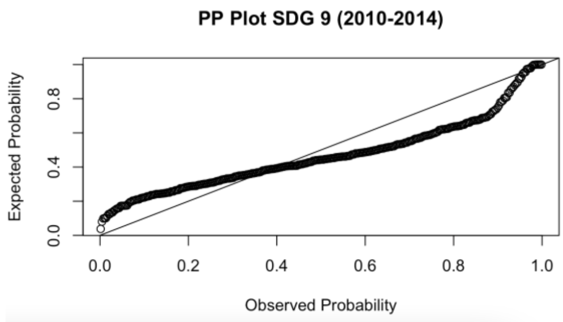
2015-2020



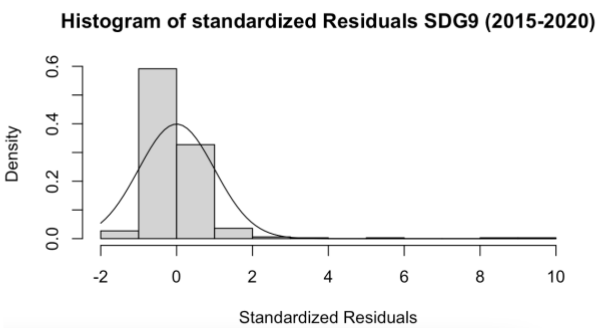


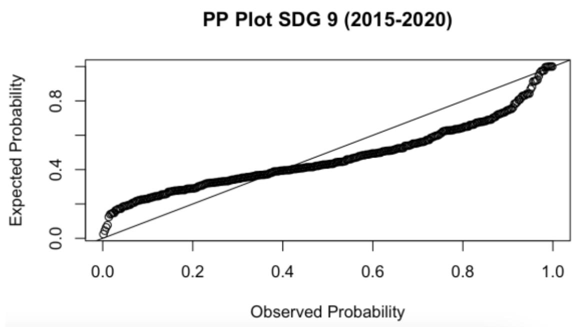
SDG9
2010-2014



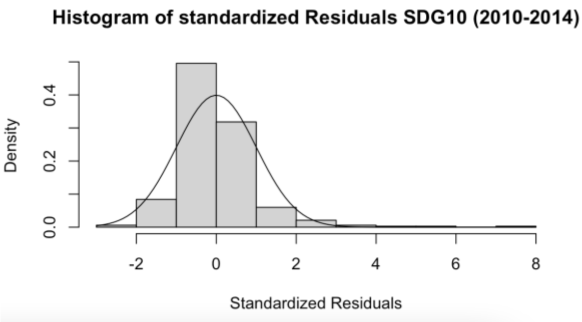


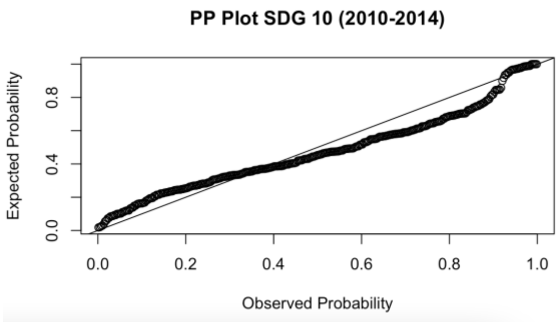
2015-2020



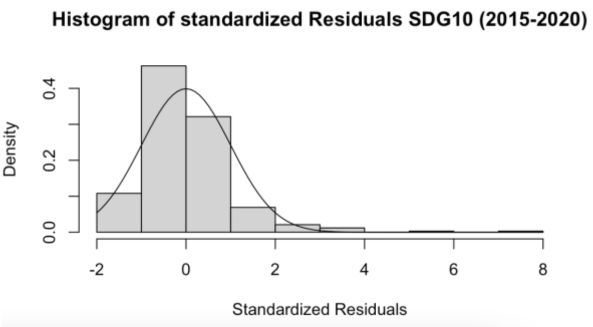


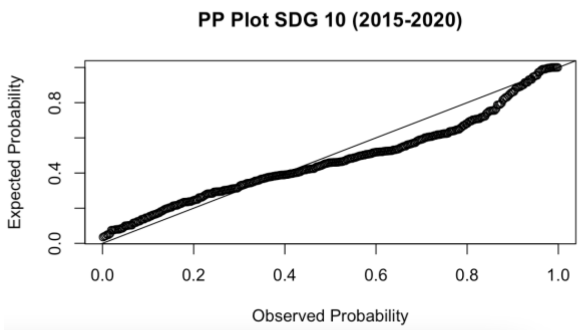
SDG10
2010-2014



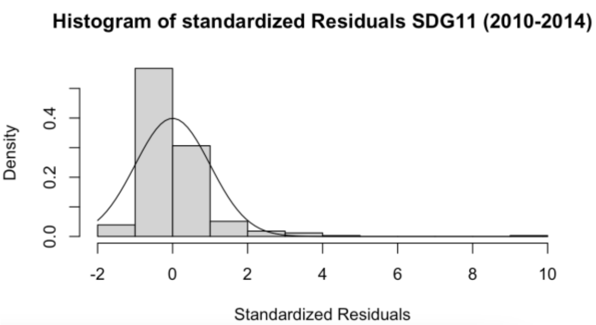


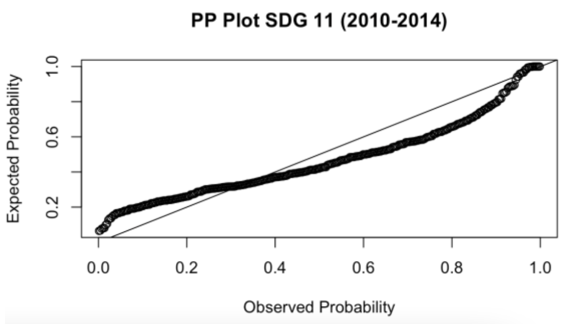
2015-2020



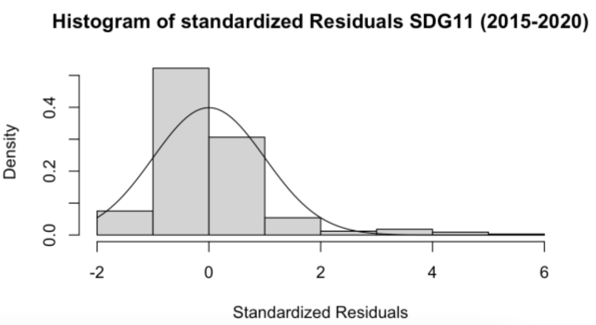


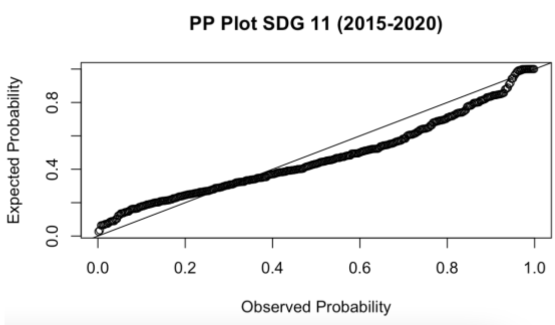
SDG11
2010-2014



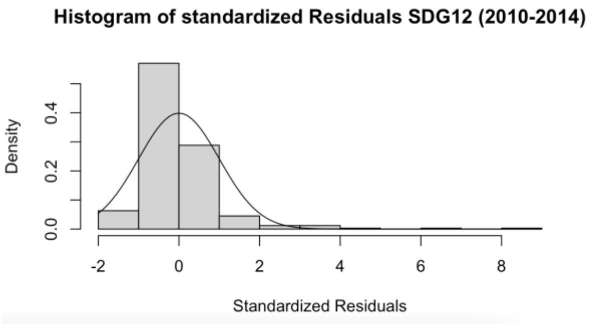


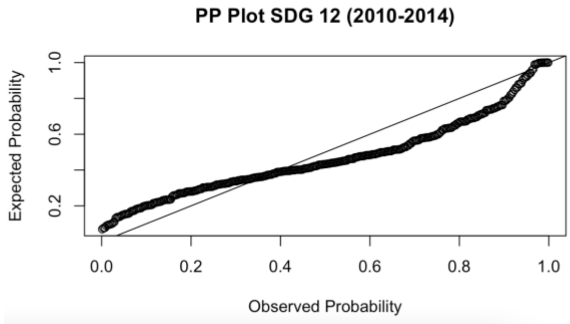
2015-2020



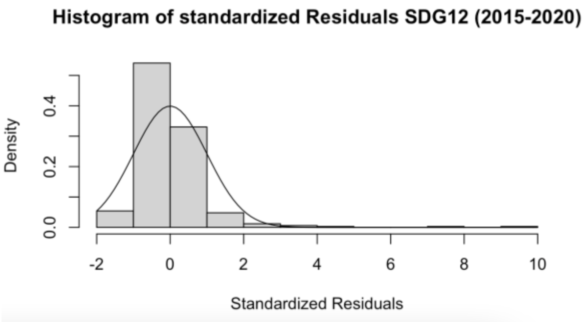


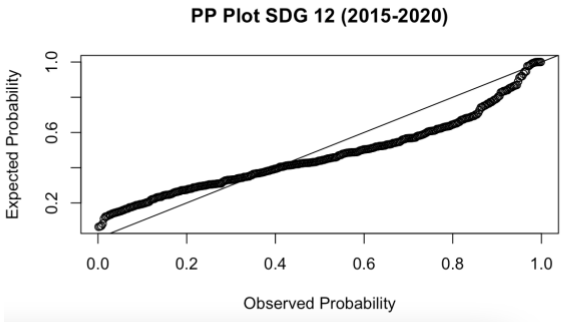
SDG12
2010-2014



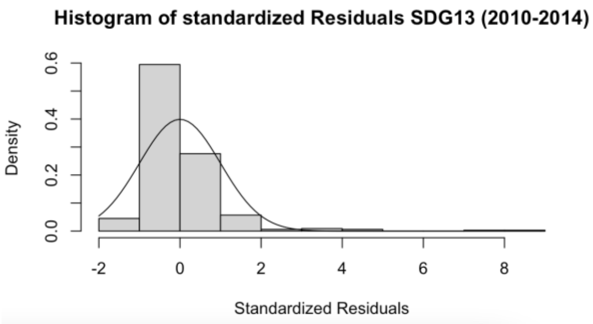


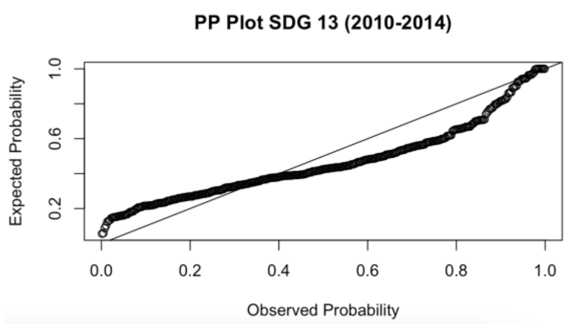
2015-2020



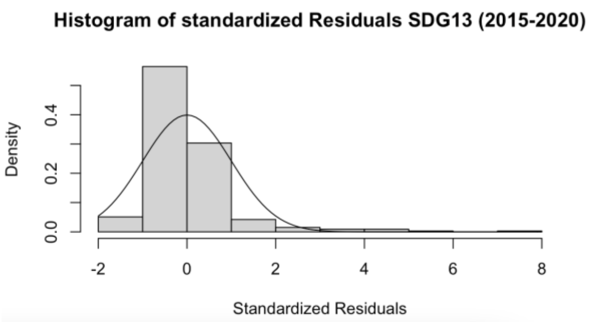


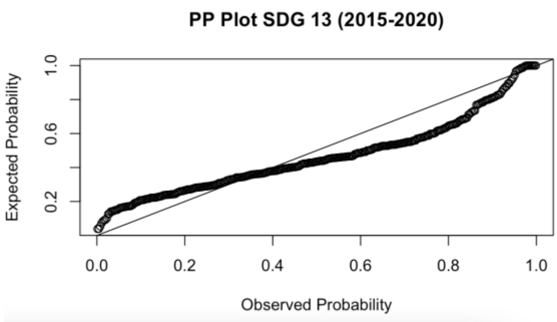
SDG13
2010-2014



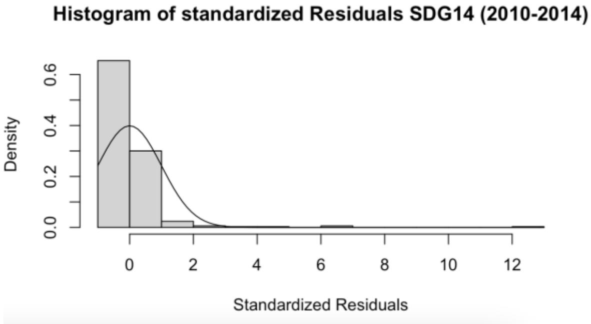


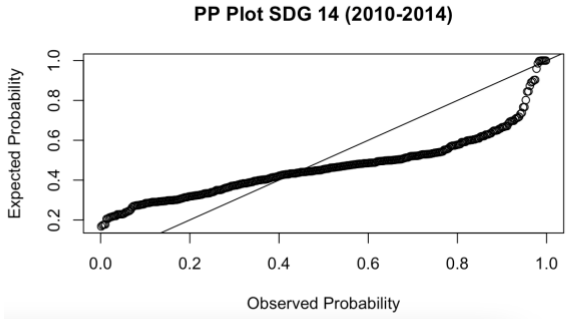
2015-2020



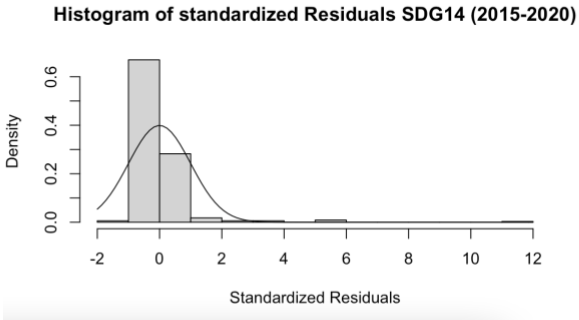


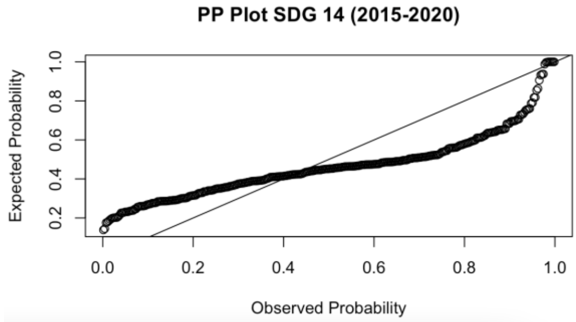
SDG14
2010-2014



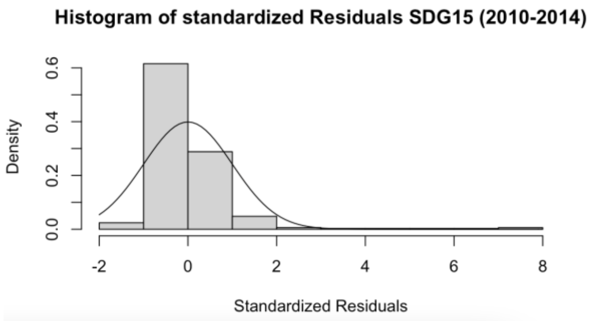


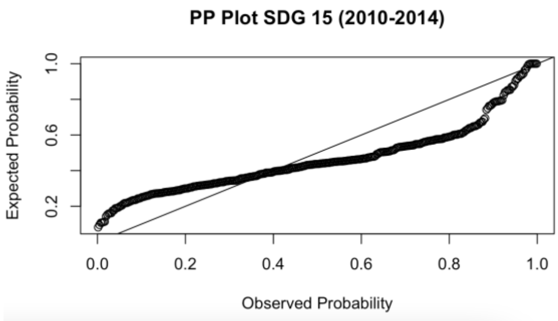
2015-2020



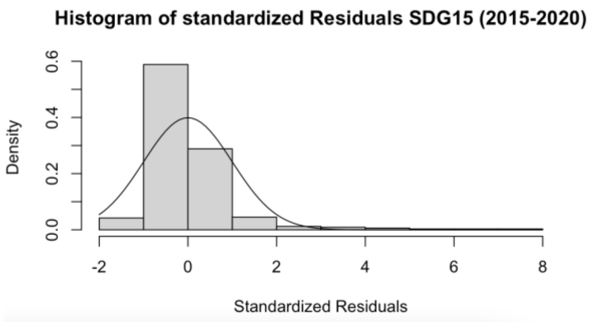


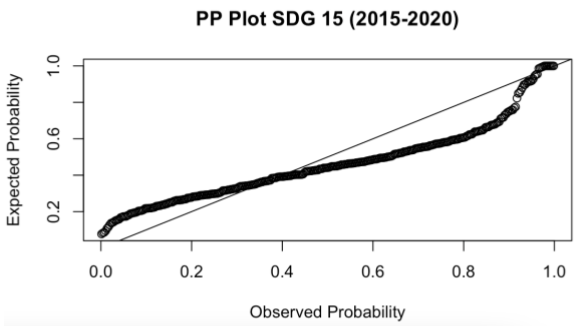
SDG15
2010-2014



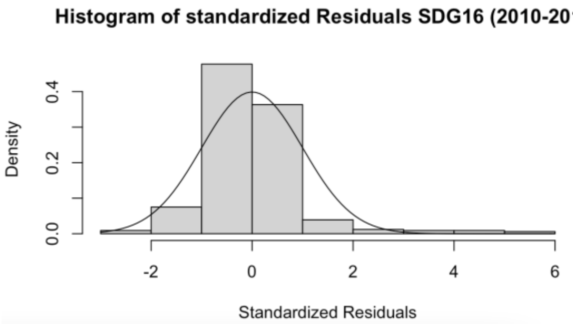


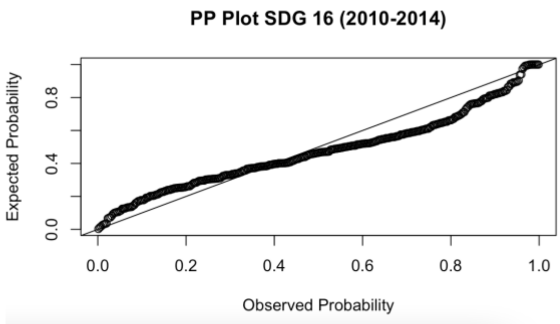
2015-2020



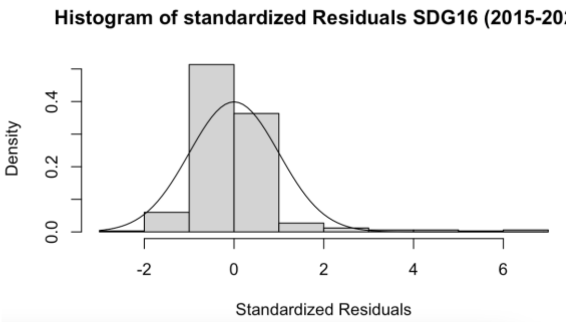


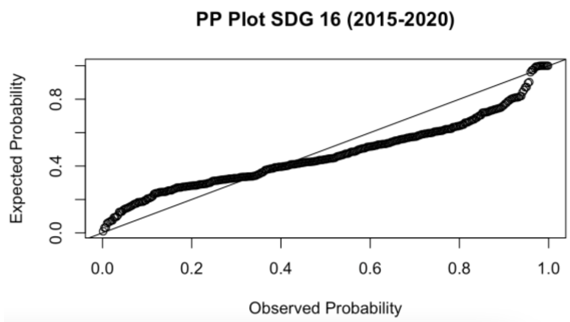
SDG16
2010-2014





2015-2020





Appendix D: Skewness coefficient before data transformation

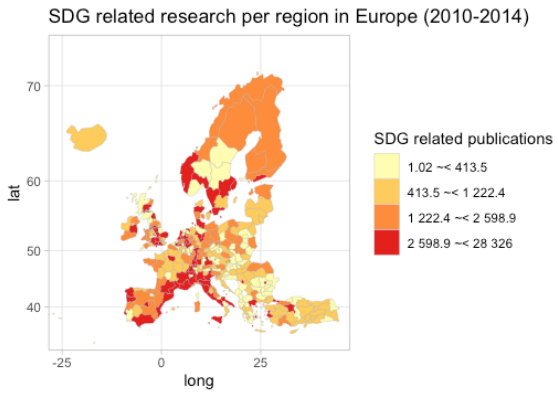
2010-2014

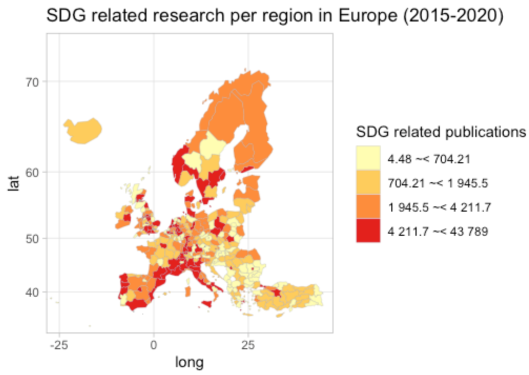
Model	SDG_share	KCI	RD_SDG	RC_1	RC_2	RC_3	Population	GDP	Education
1	1.28	0.45	0.13				3.18	0.76	0.47
SDG1	6.09	0.45	0.88	3.65	-3.77	-2.66	3.18	0.76	0.47
SDG2	3.38	0.45	-0.13	4.74			3.18	0.76	0.47
SDG3	1.02	0.45	0.57	-0.61	0.46		3.18	0.76	0.47
SDG4	6.99	0.45					3.18	0.76	0.47
SDG5	2.24	0.45	0.81				3.18	0.76	0.47
SDG6	6.65	0.45	0.11				3.18	0.76	0.47
SDG7	4.90	0.45	0.02		1.99		3.18	0.76	0.47
SDG8	5.43	0.45	0.29				3.18	0.76	0.47
SDG9	3.78	0.45	0.27				3.18	0.76	0.47
SDG10	2.27	0.45	0.92	0.37	2.81		3.18	0.76	0.47
SDG11	3.93	0.45	0.20				3.18	0.76	0.47
SDG12	3.53	0.45	0.04				3.18	0.76	0.47
SDG13	3.63	0.45	-0.03	3.63	0.72		3.18	0.76	0.47
SDG14	7.54	0.45	0.11		-1.17		3.18	0.76	0.47
SDG15	4.70	0.45	0.41	-1.58	-0.26		3.18	0.76	0.47
SDG16	2.22	0.45	0.59				3.18	0.76	0.47

2015-2020

Model	SDG_share	KCI	RD_SDG	RC_1	RC_2	RC_3	Population	GDP	Education
2	1.54	0.39	0.19				3.36	0.78	0.44
SDG1	4.36	0.39	0.80	2.88	-2.68	-3.63	3.36	0.78	0.44
SDG2	9.97	0.39	0.19	6.19	3.81		3.36	0.78	0.44
SDG3	1.45	0.39	0.40	-1.17	0.18		3.36	0.78	0.44
SDG4	9.15	0.39	0.79				3.36	0.78	0.44
SDG5	1.86	0.39	0.95				3.36	0.78	0.44
SDG6	4.31	0.39	0.12				3.36	0.78	0.44
SDG7	2.26	0.39	0.04		2.17		3.36	0.78	0.44
SDG8	4.24	0.39	0.24				3.36	0.78	0.44
SDG9	4.51	0.39	0.10				3.36	0.78	0.44
SDG10	1.69	0.39	0.96	0.03	5.50		3.36	0.78	0.44
SDG11	2.06	0.39	0.20				3.36	0.78	0.44
SDG12	3.79	0.39	0.10				3.36	0.78	0.44
SDG13	3.01	0.39	-0.05	5.43	0.46		3.36	0.78	0.44
SDG14	7.03	0.39	0.01	4.01	-1.37		3.36	0.78	0.44
SDG15	3.81	0.39	0.38	-1.84	-0.25		3.36	0.78	0.44
SDG16	2.66	0.39	0.67				3.36	0.78	0.44

Appendix E: SDG related publication count





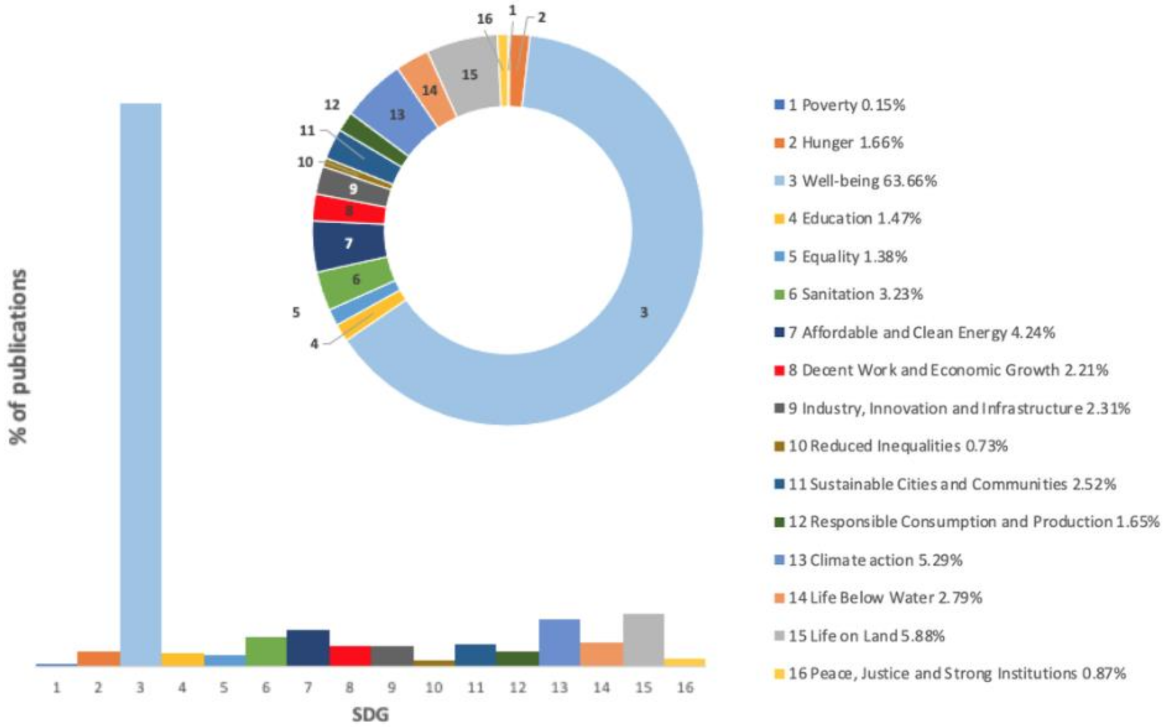
Top 10 regions of SDG related publications

2010-2014			**2015-2020**		
NUTS	**Region**	**SDG pub**	**NUTS**	**Region**	**SDG pub**
UKI3	Inner Londen-West	28 327	**UKI3**	Inner Londen-West	43 789
FR10	Île-de-France	18 922	**FR10**	Île-de-France	27 297
ES51	Cataluña	13 316	**ITC4**	Lombardia	22 706
ITC4	Lombardia	12 391	**ES51**	Cataluña	20 359
ES30	Comunidad de Madrid	11 376	**ES30**	Comunidad de Madrid	20 082
ITI4	Lazio	11377	**ITI4**	Lazio	19 099
NL33	Zuid-Holland	10 144	**DK01**	Hovedstaden	16 032
NL32	Noord-Holland	10 033	**NL33**	Zuid-Holland	15 528
DK01	Hovedstaden	9 483	**NL32**	Noord-Holland	14 805
UKJ1	Berkshire, Buckinghamshire and Oxfordshire	8 865	**ES61**	Andalucía	14 796

Bottom 10 regions of SDG related publications

2010-2014			**2015-2020**		
NUTS	**Region**	**SDG pub**	**NUTS**	**Region**	**SDG pub**
AL03	Jug	1.02	**FI20**	Åland	4.47
FI20	Åland	1.65	**AL03**	Jug	5.65
AL01	Veri	1.79	**AL01**	Veri	6.87
FRY5	Mayotte	7.37	**FRY5**	Mayotte	9.75
ES63	Ciudad de Ceuta	8.61	**EL42**	Notio Aigaio	23.24
BG32	Severen Tsentralen	9.72	**UKI4**	Inner Londen-East	24.94
ES64	Corse	11.37	**ES63**	Ciudad de Ceuta	25.12
AT11	Highlands and Islands	16.35	**AT11**	Azores	28.76
UKI4	Inner Londen-East	21.62	**BG31**	Severozapaden	28.95
FRY1	Guadeloupe	23.00	**BE34**	Province Luxembourg (BE)	36.32

Appendix F: Distribution of SDG related research (2010-2014)

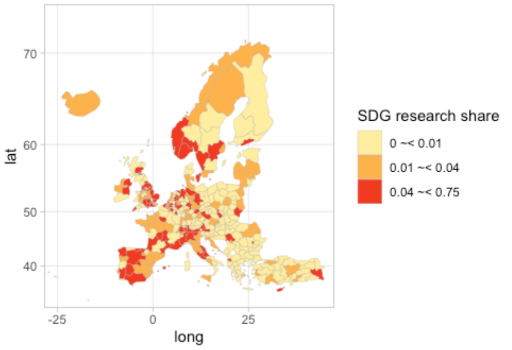


Appendix G: Most represented subfields

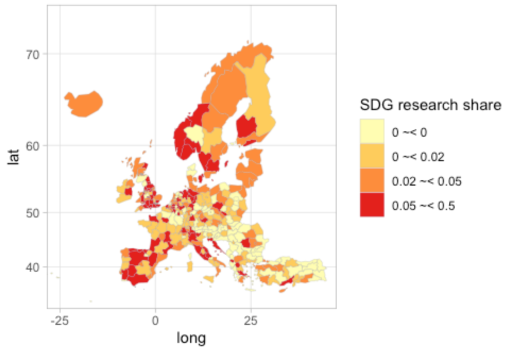
2010-2014		2015-2020	
Subfield	**#**	**Subfield**	**#**
Engineering, Electrical & Electronic	73 932	Multidisciplinary Sciences	136 024
Multidisciplinary Sciences	66 390	Engineering, Electrical & Electronic	107 400
Biochemistry & Molecular Biology	51 100	Environmental Sciences	84 808
Materials Science, Multidisciplinary	50 657	Chemistry, Multidisciplinary	82 344
Physics, Applied	49 874	Materials Science, Multidisciplinary	76 413
Optics	48 400	Medicine, General & Internal	69 896
Chemistry, Multidisciplinary	48 323	Oncology	64 839
Medicine, General & Internal	44 550	Biochemistry & Molecular Biology	63 611
Oncology	44 482	Neurosciences	57 914
Chemistry, Physical	41 435	Physics, Applied	56 473

Appendix H: Research share maps individual SDGs

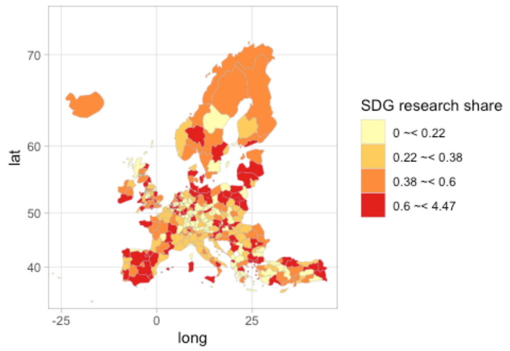
SDG1 related research per region in Europe (2010-2014)



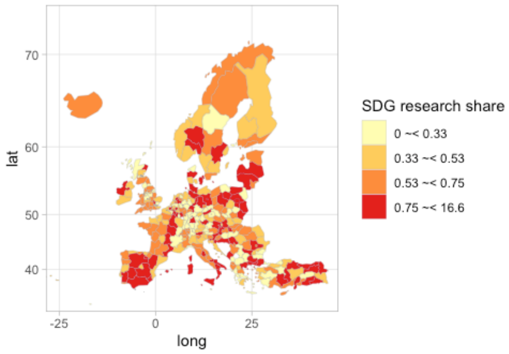
SDG1 related research per region in Europe (2015-2020)



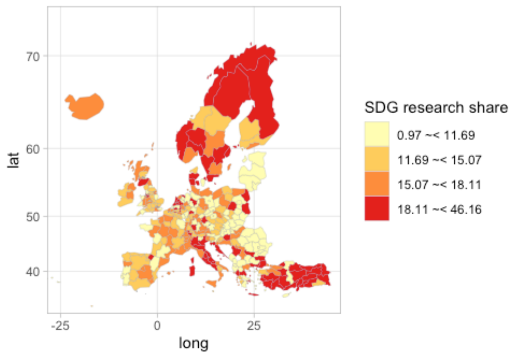
SDG2 related research per region in Europe (2010-2014)



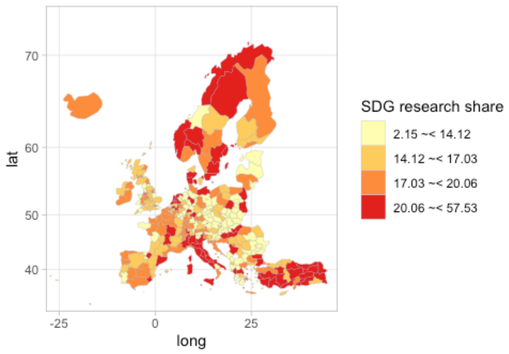
SDG2 related research per region in Europe (2015-2020)



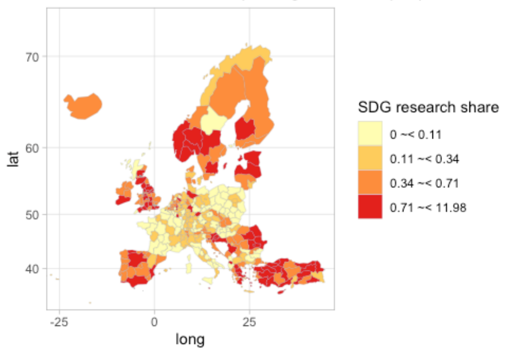
SDG3 related research per region in Europe (2010-2014)



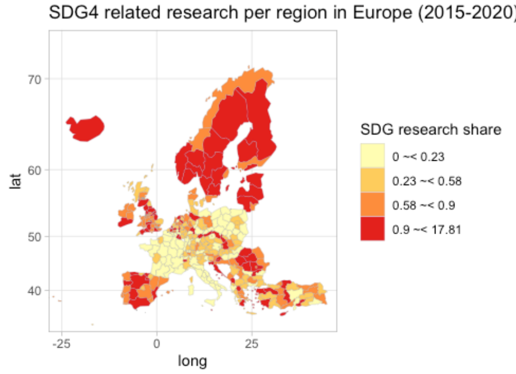
SDG3 related research per region in Europe (2015-2020)



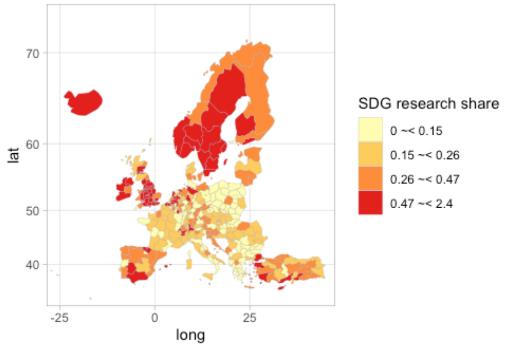
SDG4 related research per region in Europe (2010-2014)



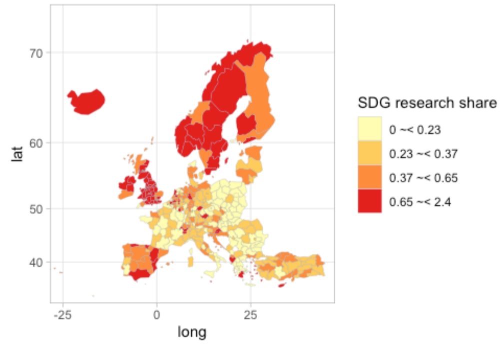
SDG4 related research per region in Europe (2015-2020)



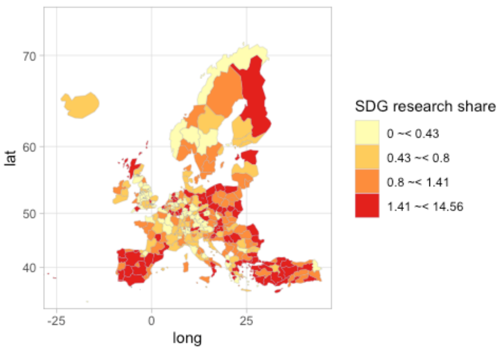
SDG5 related research per region in Europe (2010-2014)



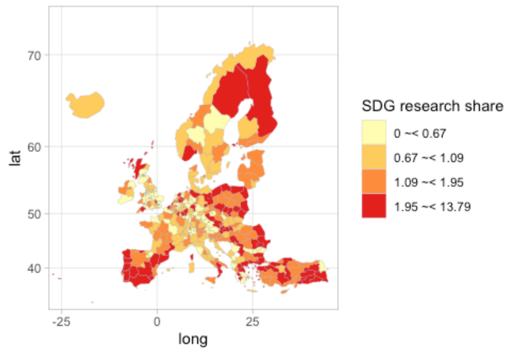
SDG5 related research per region in Europe (2015-2020)



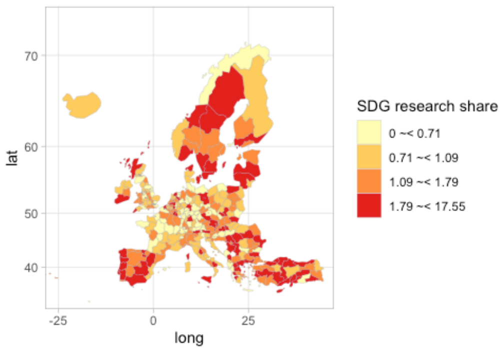
SDG6 related research per region in Europe (2010-2014)



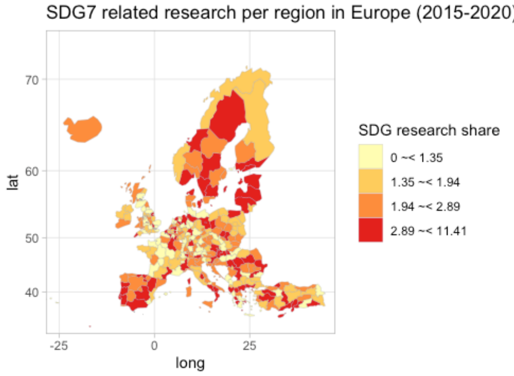
SDG6 related research per region in Europe (2015-2020)



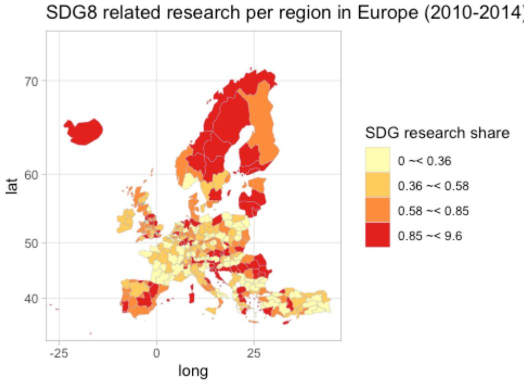
SDG7 related research per region in Europe (2010-2014)



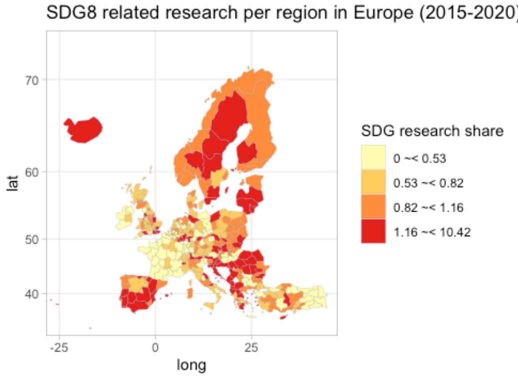
SDG7 related research per region in Europe (2015-2020)



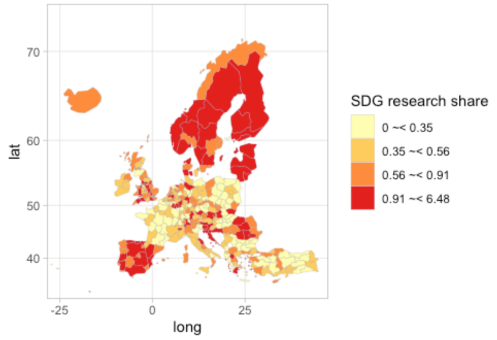
SDG8 related research per region in Europe (2010-2014)



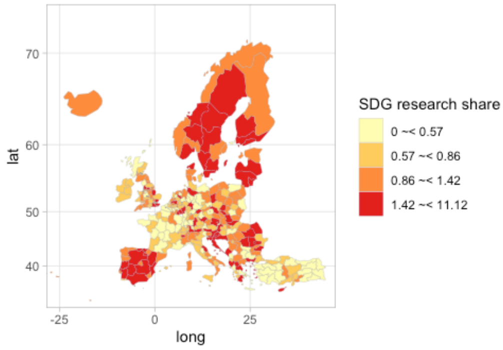
SDG8 related research per region in Europe (2015-2020)



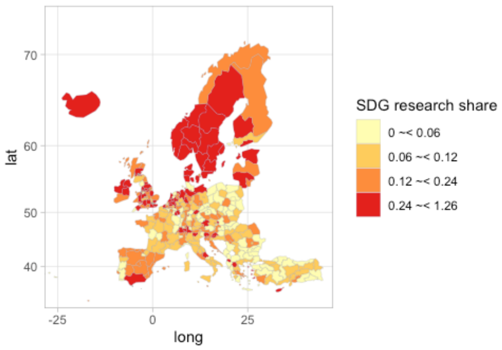
SDG9 related research per region in Europe (2010-2014)



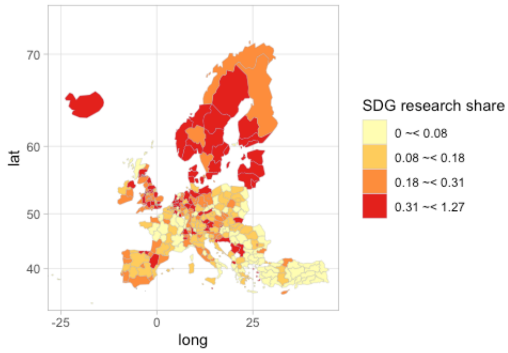
SDG9 related research per region in Europe (2015-2020)



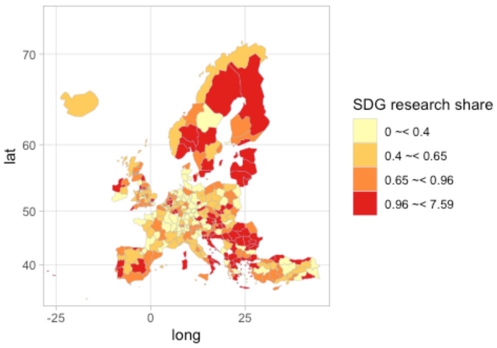
SDG10 related research per region in Europe (2010-2014)



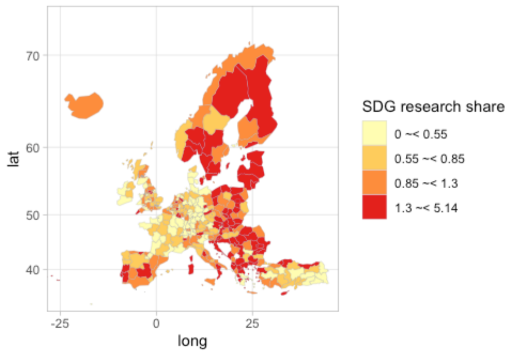
SDG10 related research per region in Europe (2015-2020)



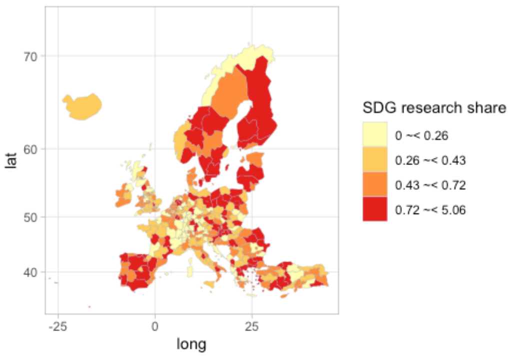
SDG11 related research per region in Europe (2010-2014)



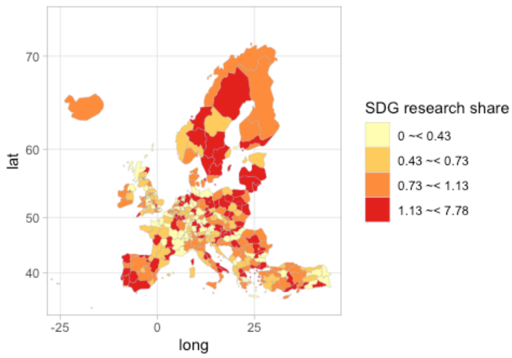
SDG11 related research per region in Europe (2015-2020)



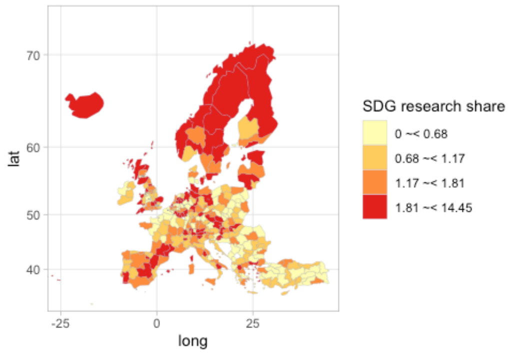
SDG12 related research per region in Europe (2010-2014)



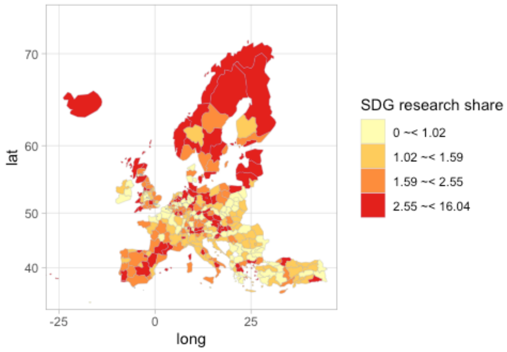
SDG12 related research per region in Europe (2015-2020)



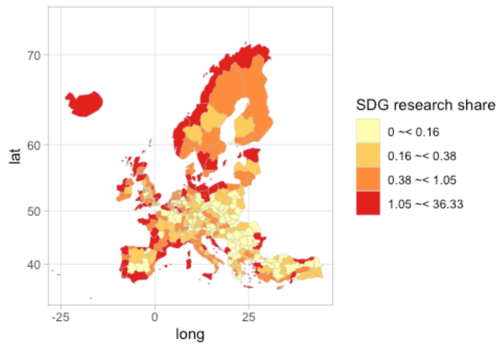
SDG13 related research per region in Europe (2010-2014)



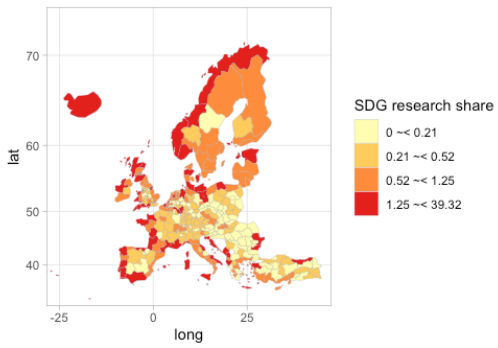
SDG13 related research per region in Europe (2015-2020)



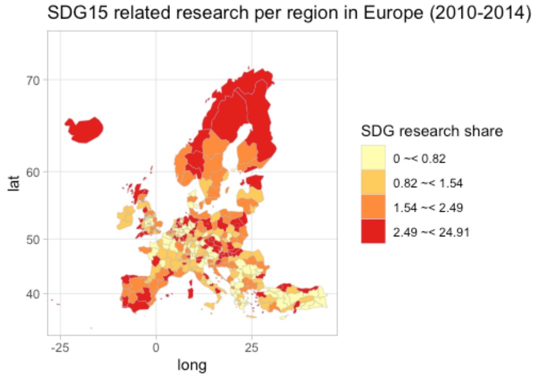
SDG14 related research per region in Europe (2010-2014)



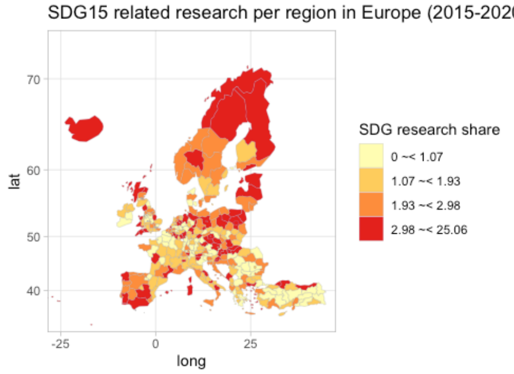
SDG14 related research per region in Europe (2015-2020)



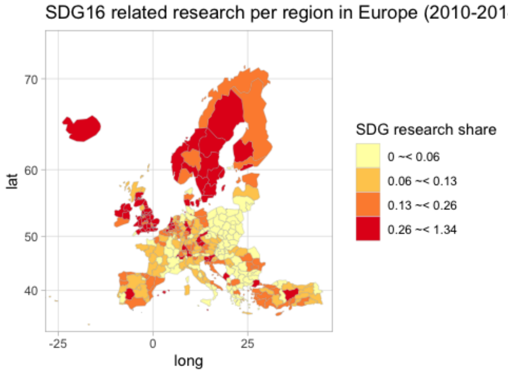
SDG15 related research per region in Europe (2010-2014)



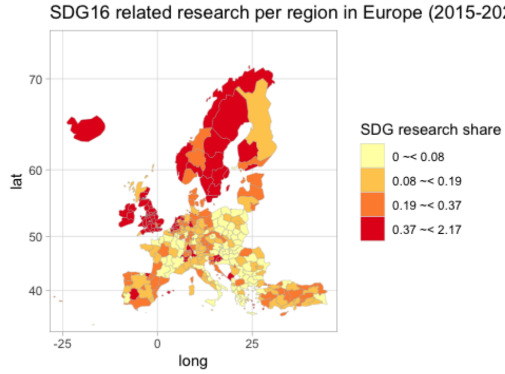
SDG15 related research per region in Europe (2015-2020)



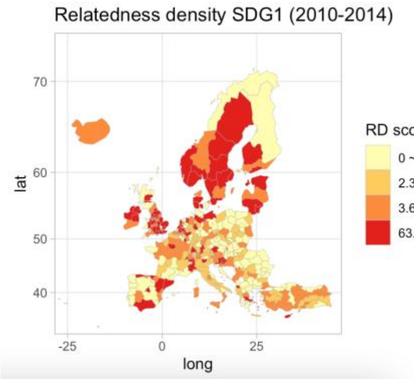
SDG16 related research per region in Europe (2010-2014)

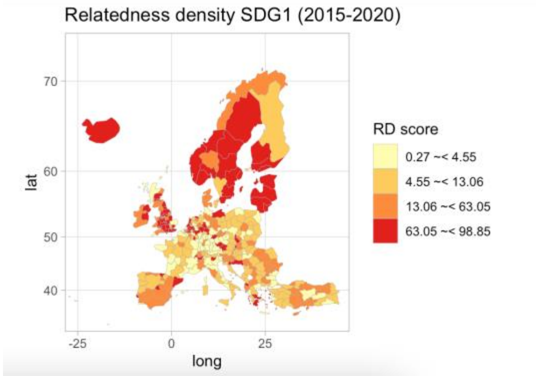


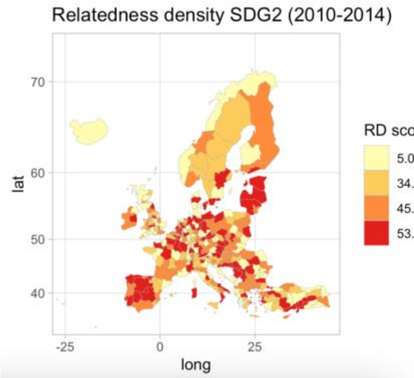
SDG16 related research per region in Europe (2015-2020)

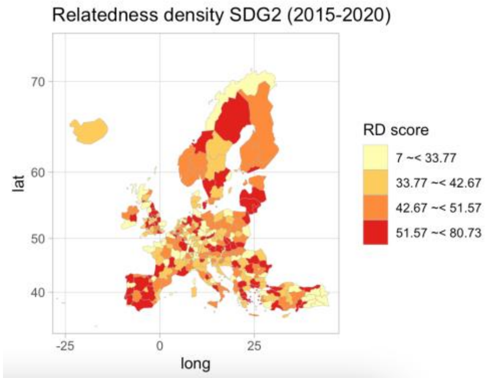


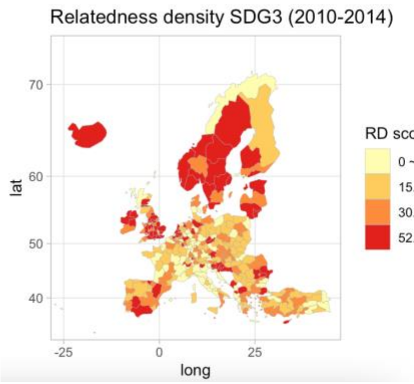
Appendix I: Relatedness density scores maps individual SDGs

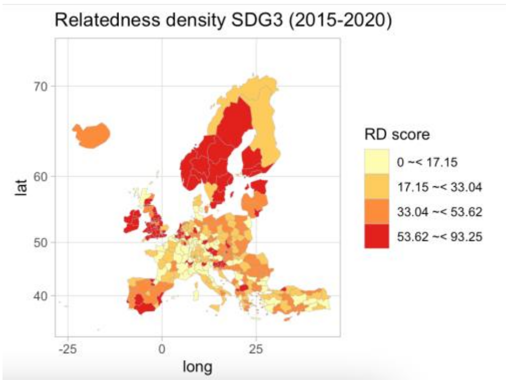


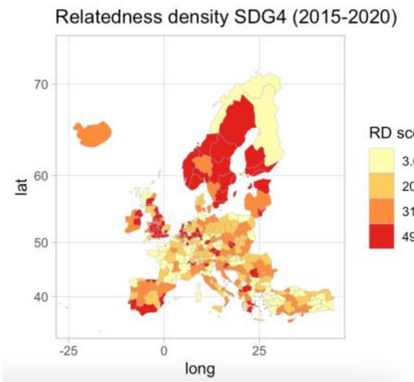


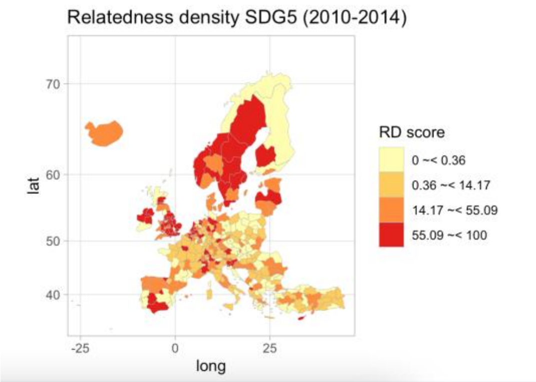


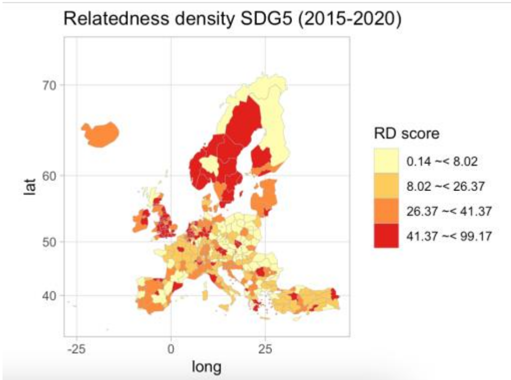


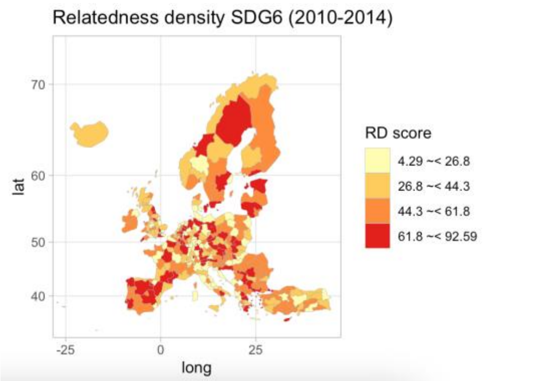


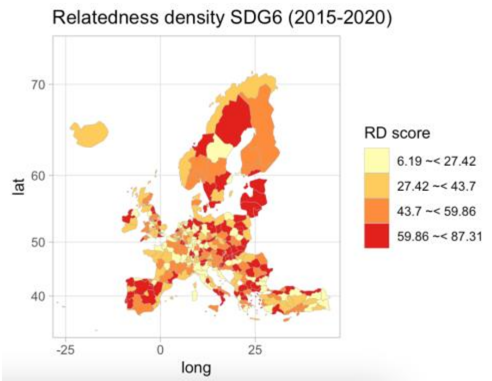


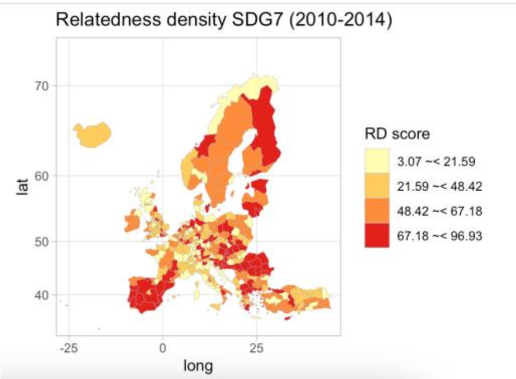


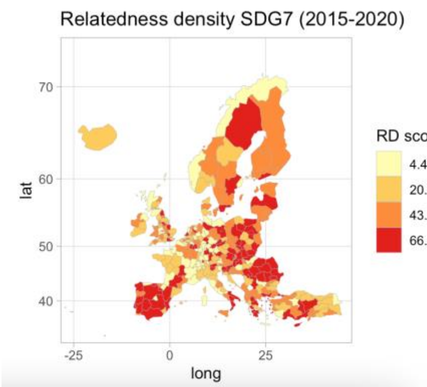

Relatedness density SDG5 (2010-2014)
RD score
0 ~< 0.36
0.36 ~< 14.17
14.17 ~< 55.09
55.09 ~< 100
lat
long

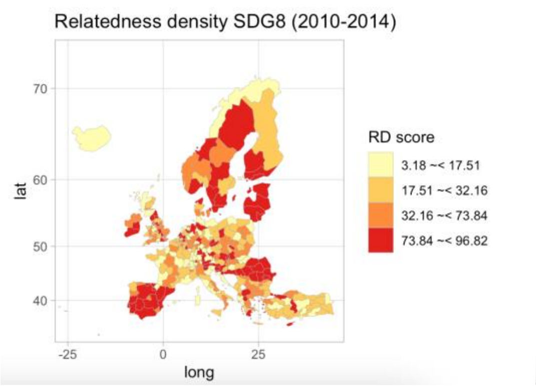

Relatedness density SDG5 (2015-2020)
RD score
0.14 ~< 8.02
8.02 ~< 26.37
26.37 ~< 41.37
41.37 ~< 99.17
lat
long

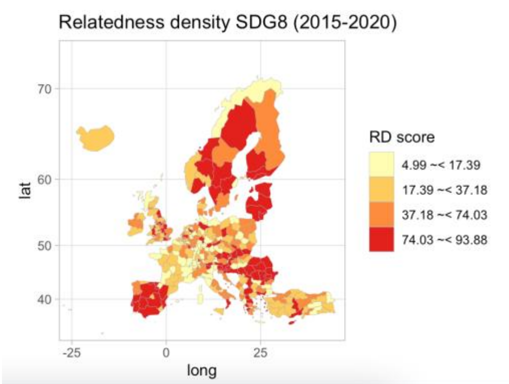

Relatedness density SDG6 (2010-2014)
RD score
4.29 ~< 26.8
26.8 ~< 44.3
44.3 ~< 61.8
61.8 ~< 92.59
lat
long

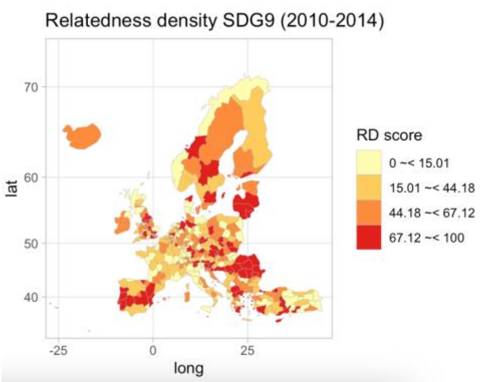

Relatedness density SDG6 (2015-2020)
RD score
6.19 ~< 27.42
27.42 ~< 43.7
43.7 ~< 59.86
59.86 ~< 87.31
lat
long

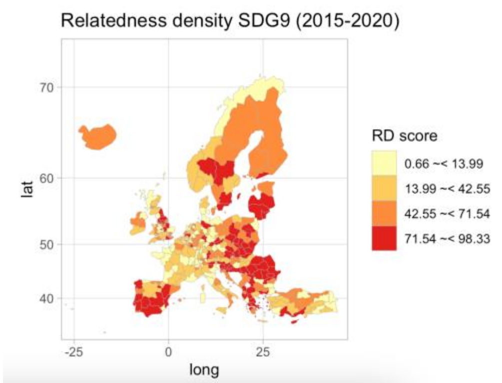

Relatedness density SDG7 (2010-2014)
RD score
3.07 ~< 21.59
21.59 ~< 48.42
48.42 ~< 67.18
67.18 ~< 96.93
lat
long

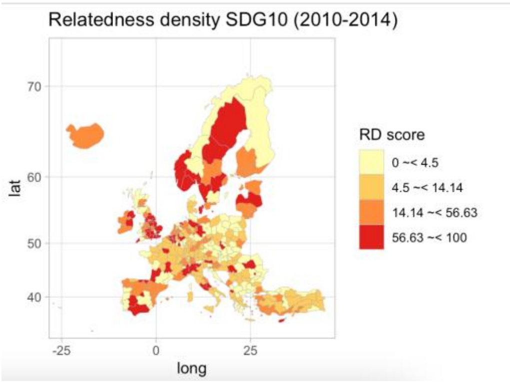

Relatedness density SDG7 (2015-2020)
RD score
4.44 ~< 20.58
20.58 ~< 43.38
43.38 ~< 66.11
66.11 ~< 94.91
lat
long

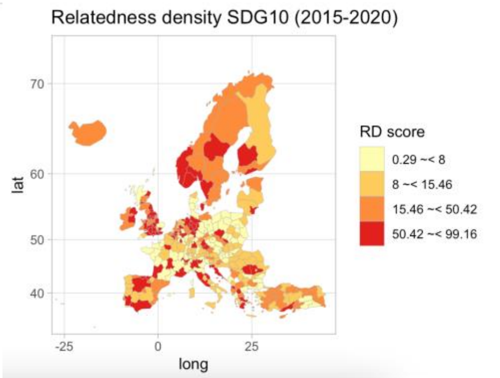

Relatedness density SDG8 (2010-2014)
RD score
3.18 ~< 17.51
17.51 ~< 32.16
32.16 ~< 73.84
73.84 ~< 96.82
lat
long

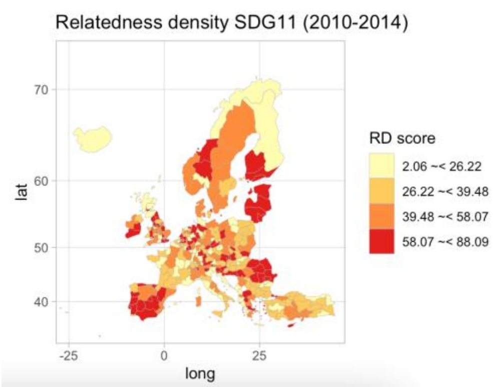

Relatedness density SDG8 (2015-2020)
RD score
4.99 ~< 17.39
17.39 ~< 37.18
37.18 ~< 74.03
74.03 ~< 93.88
lat
long

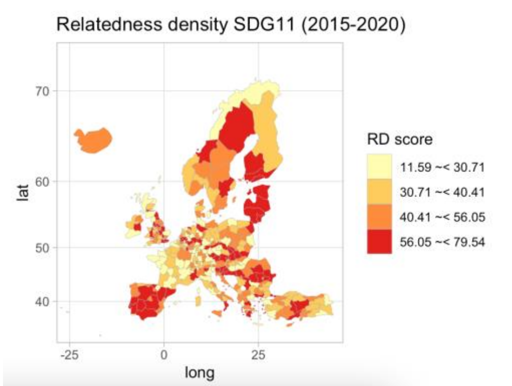

Relatedness density SDG9 (2010-2014)
RD score
0 ~< 15.01
15.01 ~< 44.18
44.18 ~< 67.12
67.12 ~< 100
lat
long
70
60
50
40
-25
0
25

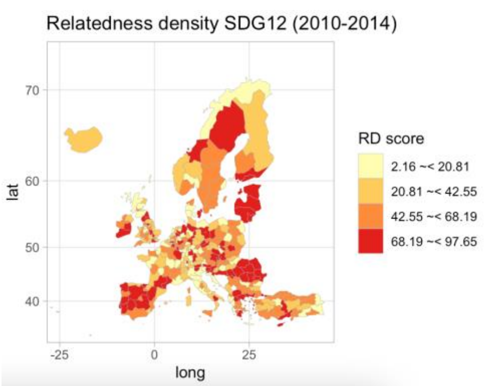

Relatedness density SDG9 (2015-2020)
RD score
0.66 ~< 13.99
13.99 ~< 42.55
42.55 ~< 71.54
71.54 ~< 98.33
lat
long
70
60
50
40
-25
0
25

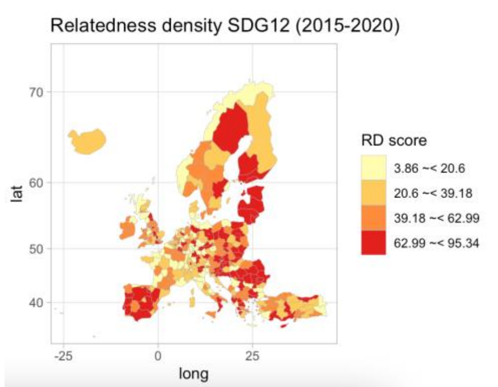

Relatedness density SDG10 (2010-2014)
RD score
0 ~< 4.5
4.5 ~< 14.14
14.14 ~< 56.63
56.63 ~< 100
lat
long
70
60
50
40
-25
0
25

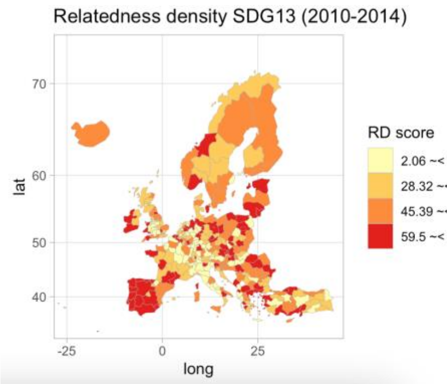

Relatedness density SDG10 (2015-2020)
RD score
0.29 ~< 8
8 ~< 15.46
15.46 ~< 50.42
50.42 ~< 99.16
lat
long
70
60
50
40
-25
0
25

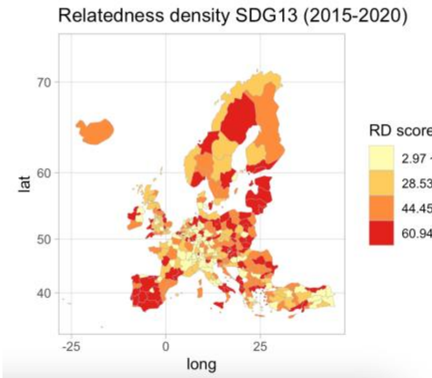

Relatedness density SDG11 (2010-2014)
RD score
2.06 ~< 26.22
26.22 ~< 39.48
39.48 ~< 58.07
58.07 ~< 88.09
lat
long
70
60
50
40
-25
0
25

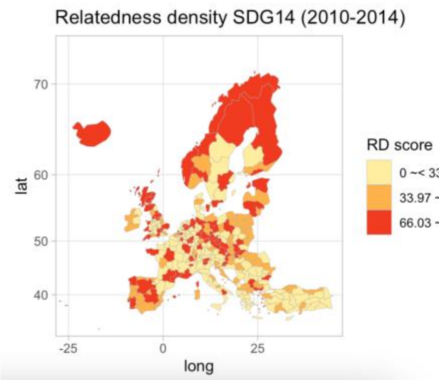

Relatedness density SDG11 (2015-2020)
RD score
11.59 ~< 30.71
30.71 ~< 40.41
40.41 ~< 56.05
56.05 ~< 79.54
lat
long
70
60
50
40
-25
0
25

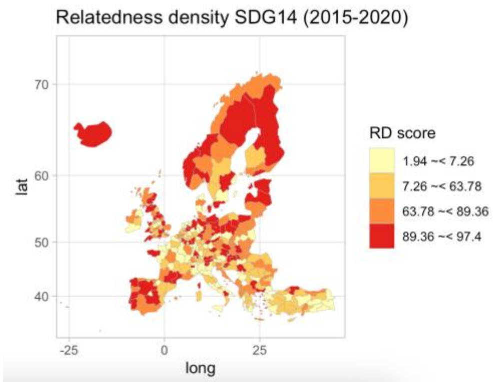

Relatedness density SDG12 (2010-2014)
RD score
2.16 ~< 20.81
20.81 ~< 42.55
42.55 ~< 68.19
68.19 ~< 97.65
lat
long
70
60
50
40
-25
0
25

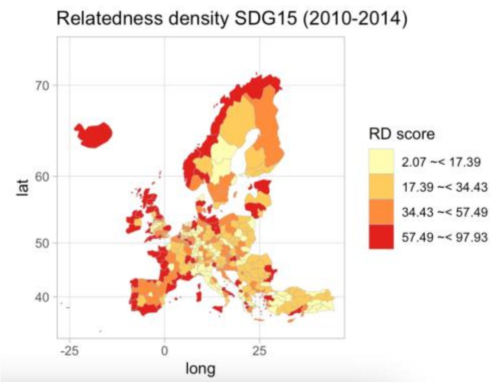

Relatedness density SDG12 (2015-2020)
RD score
3.86 ~< 20.6
20.6 ~< 39.18
39.18 ~< 62.99
62.99 ~< 95.34
lat
long
70
60
50
40
-25
0
25

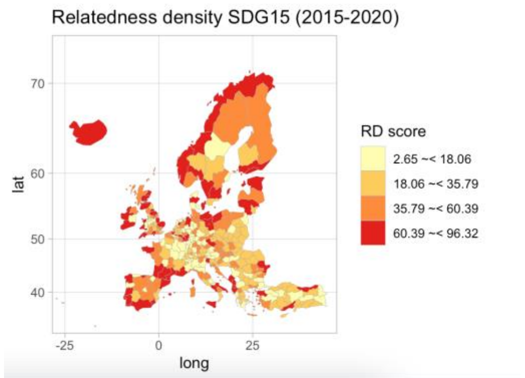

Relatedness density SDG13 (2010-2014)
lat
long
RD score
2.06 ~< 28.32
28.32 ~< 45.39
45.39 ~< 59.5
59.5 ~< 97.94

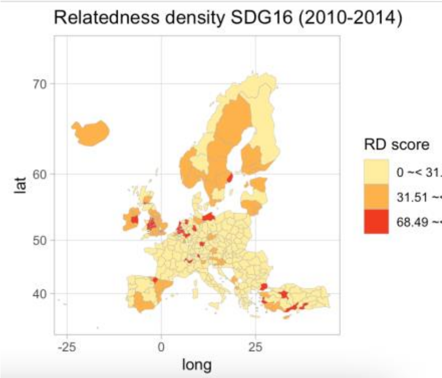

Relatedness density SDG13 (2015-2020)
lat
long
RD score
2.97 ~< 28.53
28.53 ~< 44.45
44.45 ~< 60.94
60.94 ~< 96.43

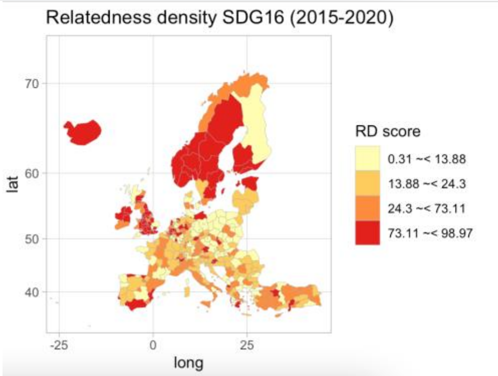

Relatedness density SDG14 (2010-2014)
lat
long
RD score
0 ~< 33.97
33.97 ~< 66.03
66.03 ~<100


Relatedness density SDG14 (2015-2020)
lat
long
RD score
1.94 ~< 7.26
7.26 ~< 63.78
63.78 ~< 89.36
89.36 ~< 97.4

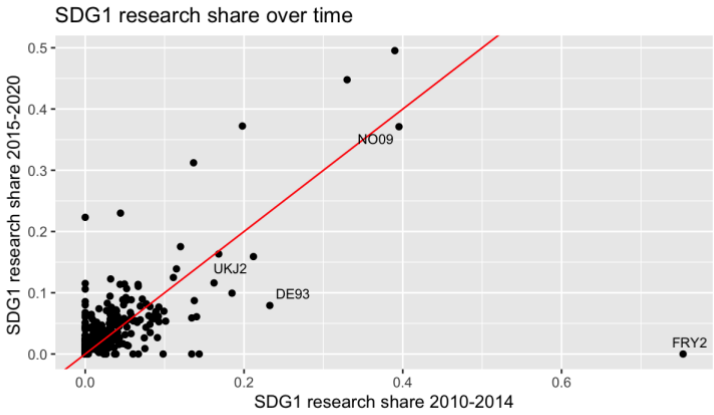

Relatedness density SDG15 (2010-2014)
lat
long
RD score
2.07 ~< 17.39
17.39 ~< 34.43
34.43 ~< 57.49
57.49 ~< 97.93

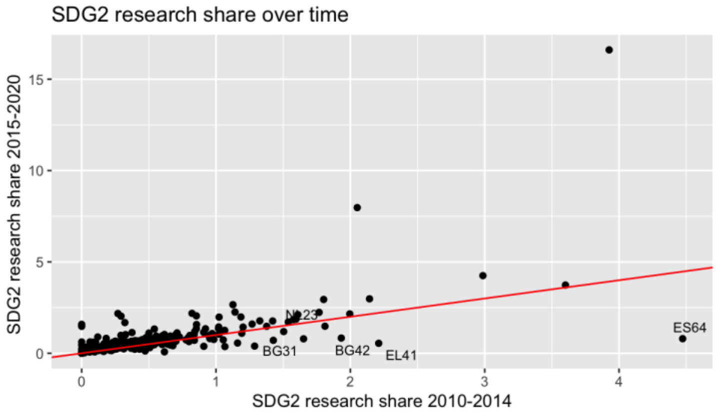

Relatedness density SDG15 (2015-2020)
lat
long
RD score
2.65 ~< 18.06
18.06 ~< 35.79
35.79 ~< 60.39
60.39 ~< 96.32

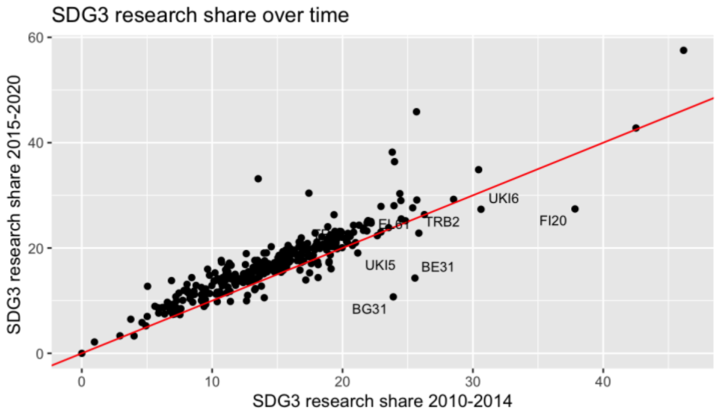

Relatedness density SDG16 (2010-2014)
lat
long
RD score
0 ~< 31.51
31.51 ~< 68.49
68.49 ~<100

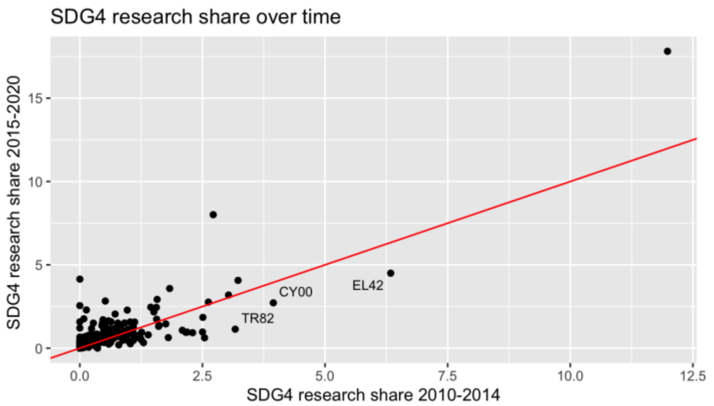

Relatedness density SDG16 (2015-2020)
lat
long
RD score
0.31 ~< 13.88
13.88 ~< 24.3
24.3 ~< 73.11
73.11 ~< 98.97

Appendix J: SDG research share over time

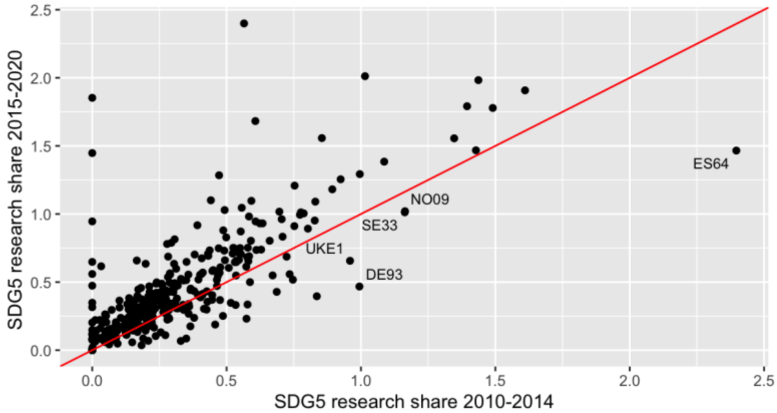






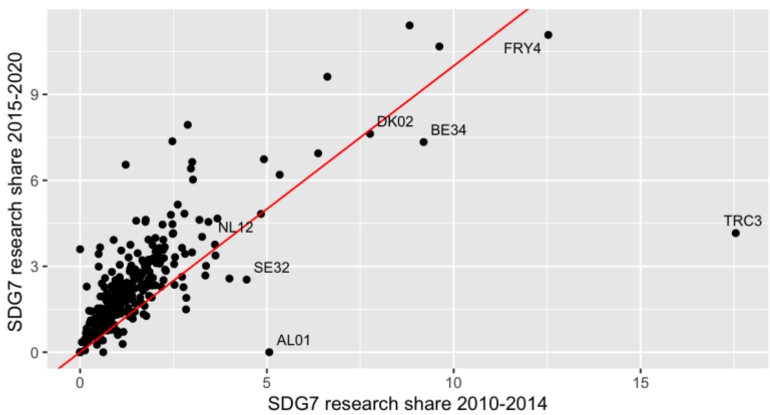


SDG5 research share over time

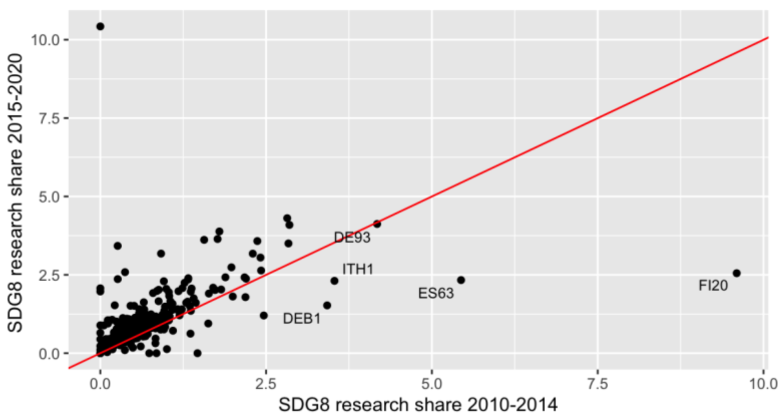


SDG6 research share over time

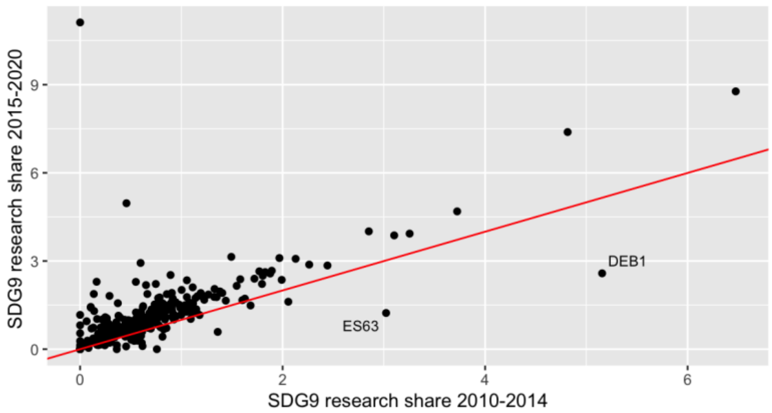
SDG7 research share over time



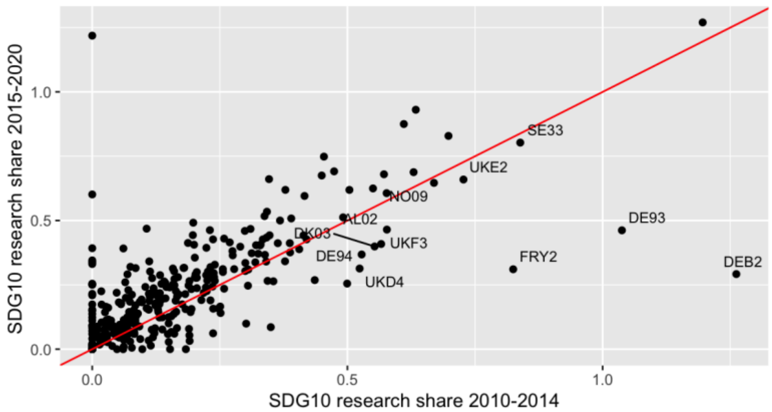
SDG8 research share over time



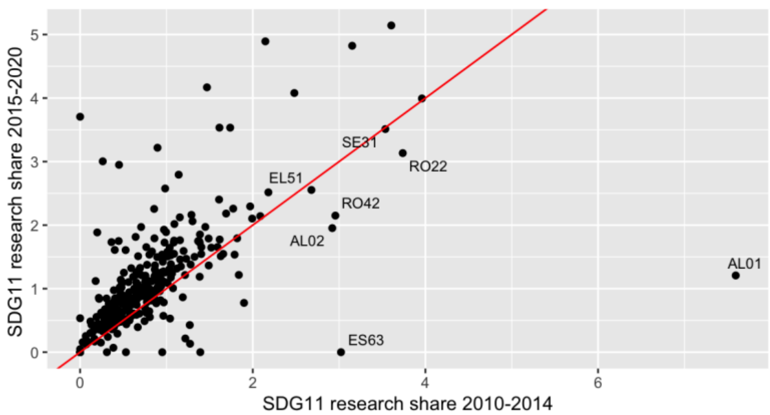
SDG9 research share over time



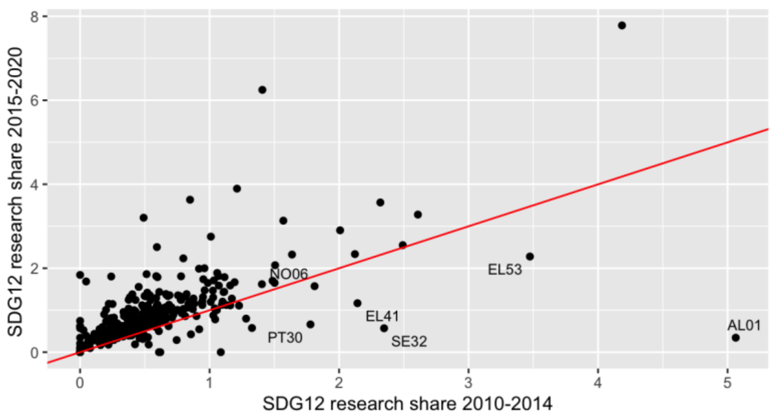
SDG10 research share over time

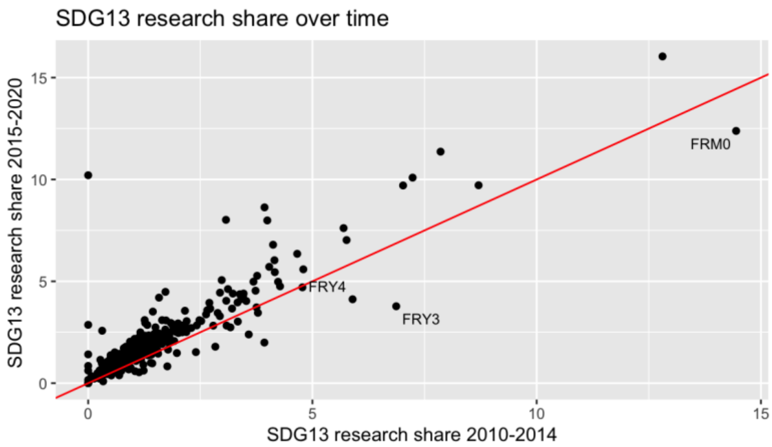


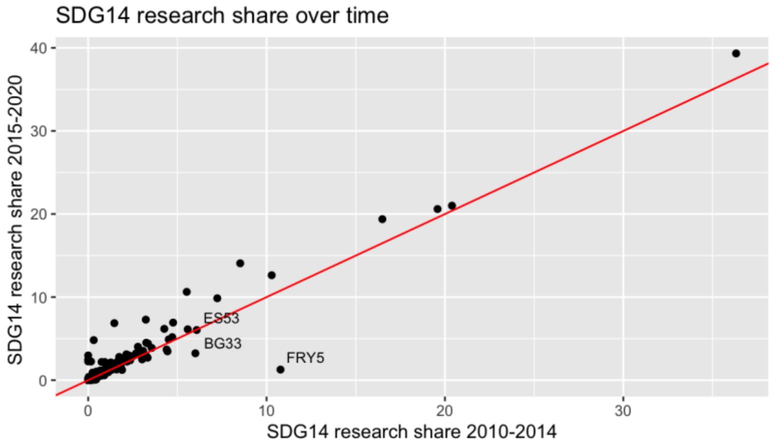
SDG11 research share over time

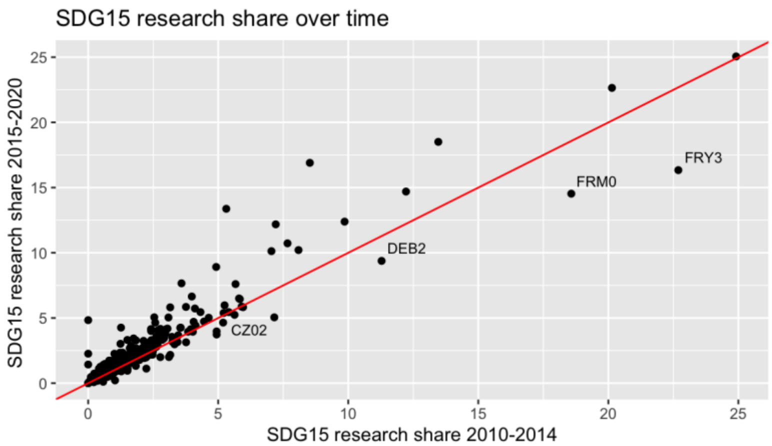


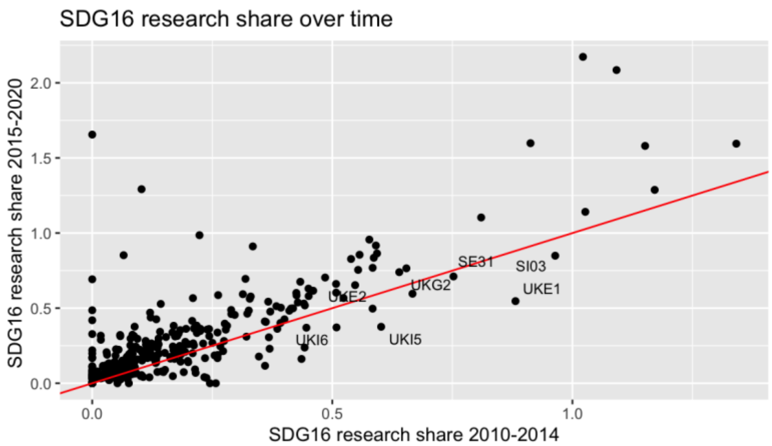
SDG12 research share over time



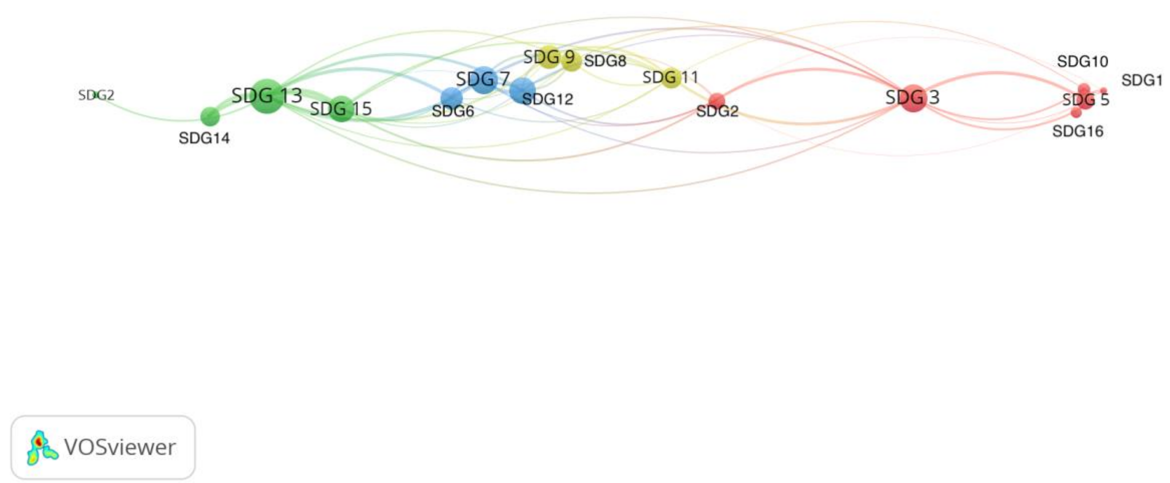

SDG13 research share over time
SDG13 research share 2015-2020
SDG13 research share 2010-2014
0
5
10
15
FRM0
FRY4
FRY3


SDG14 research share over time
SDG14 research share 2015-2020
SDG14 research share 2010-2014
0
10
20
30
40
ES53
BG33
FRY5


SDG15 research share over time
SDG15 research share 2015-2020
SDG15 research share 2010-2014
0
5
10
15
20
25
FRY3
FRM0
DEB2
CZ02

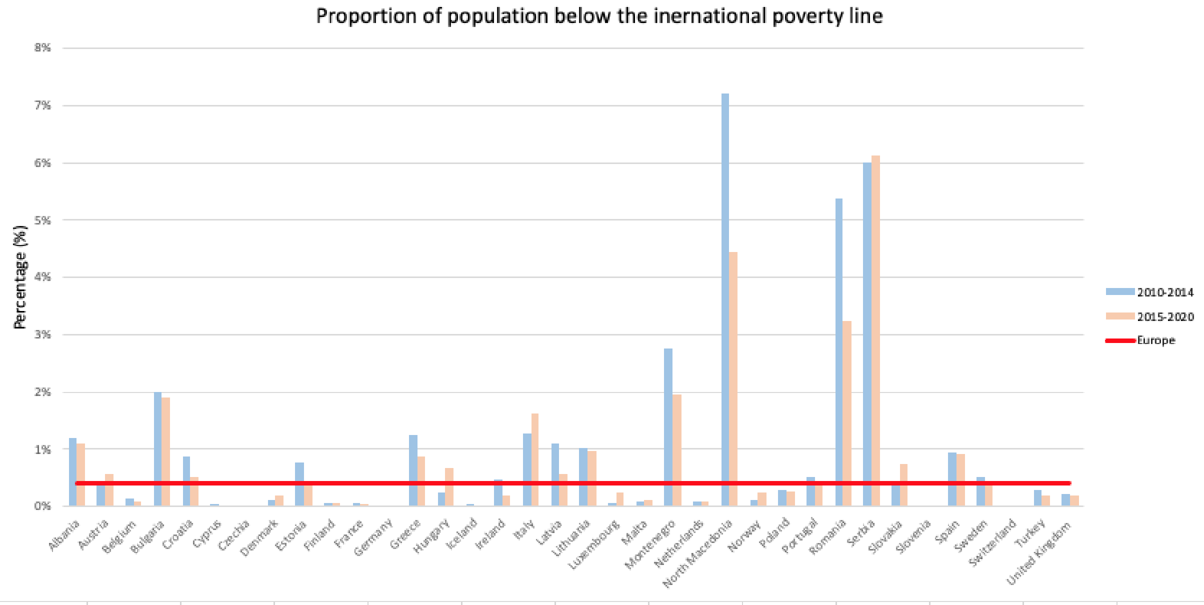

SDG16 research share over time
SDG16 research share 2015-2020
SDG16 research share 2010-2014
0.0
0.5
1.0
1.5
2.0
SE31
SI03
UKE1
UKG2
UKE2
UKI6
UKI5

Appendix K: Scientific relatedness space of SDGS (2010-2014)

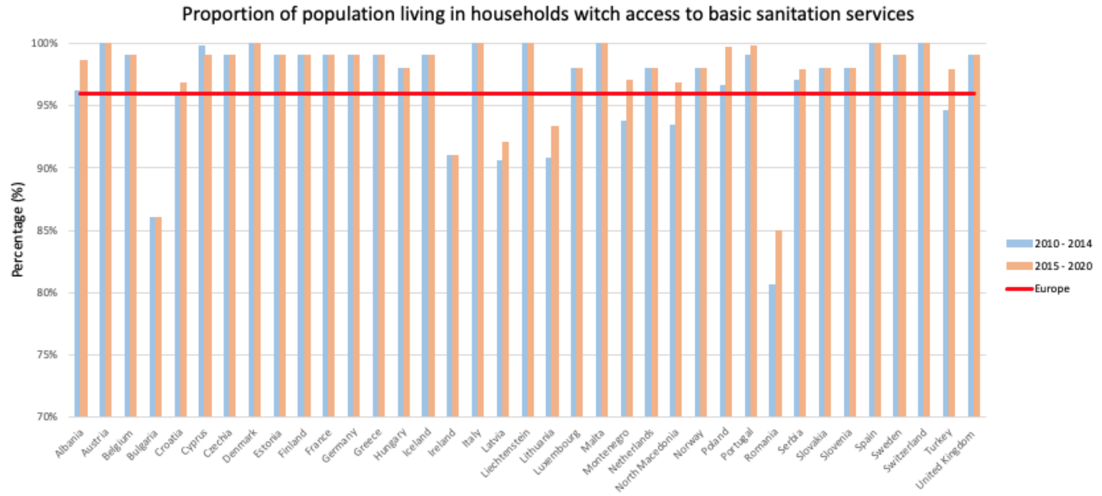


Appendix L: Plots regional characteristics

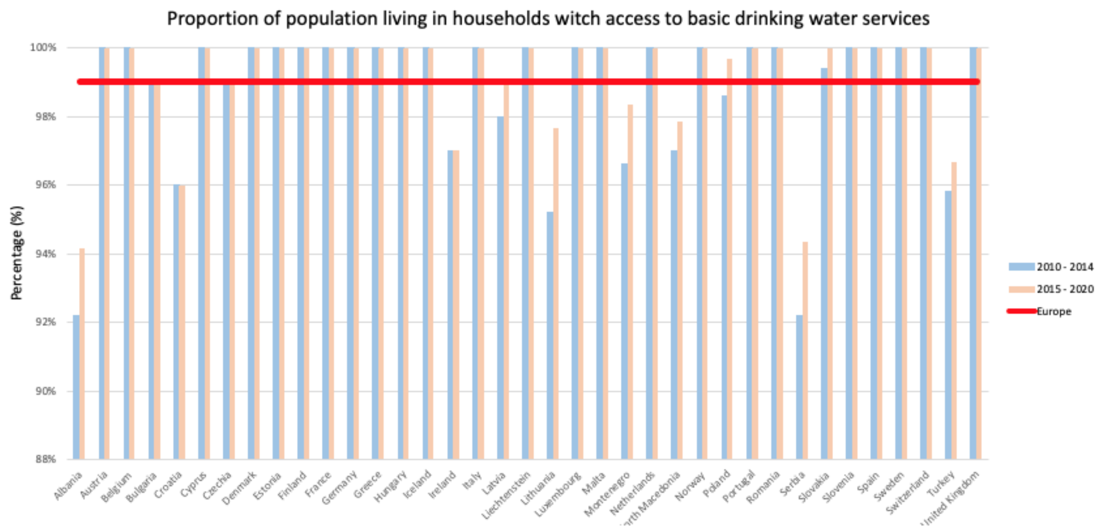
SDG1.1.1



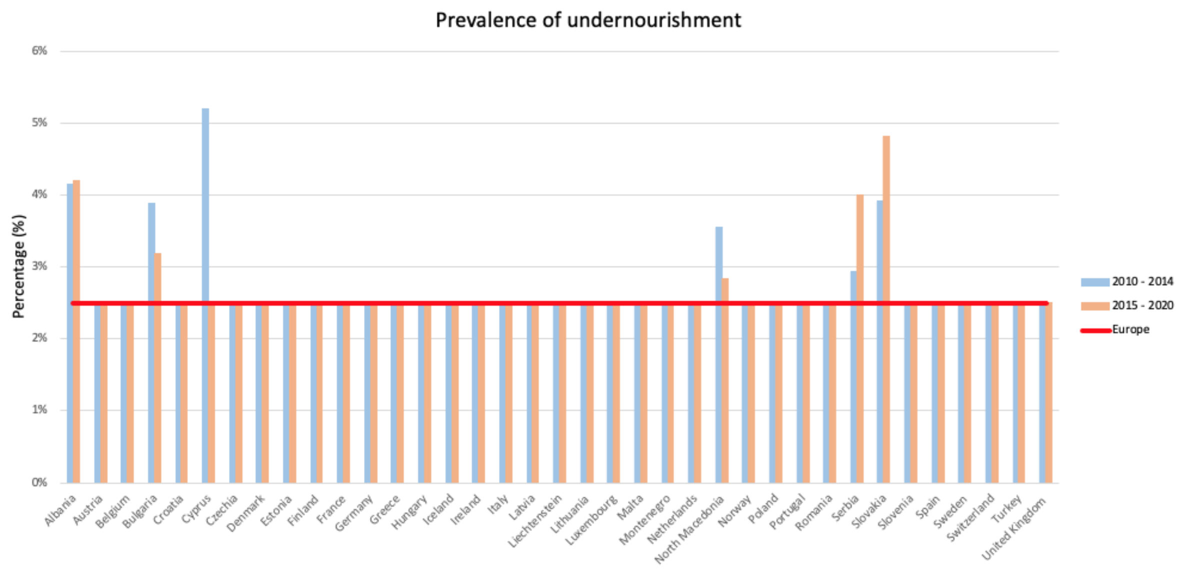
SDG1.4.1 sanitation



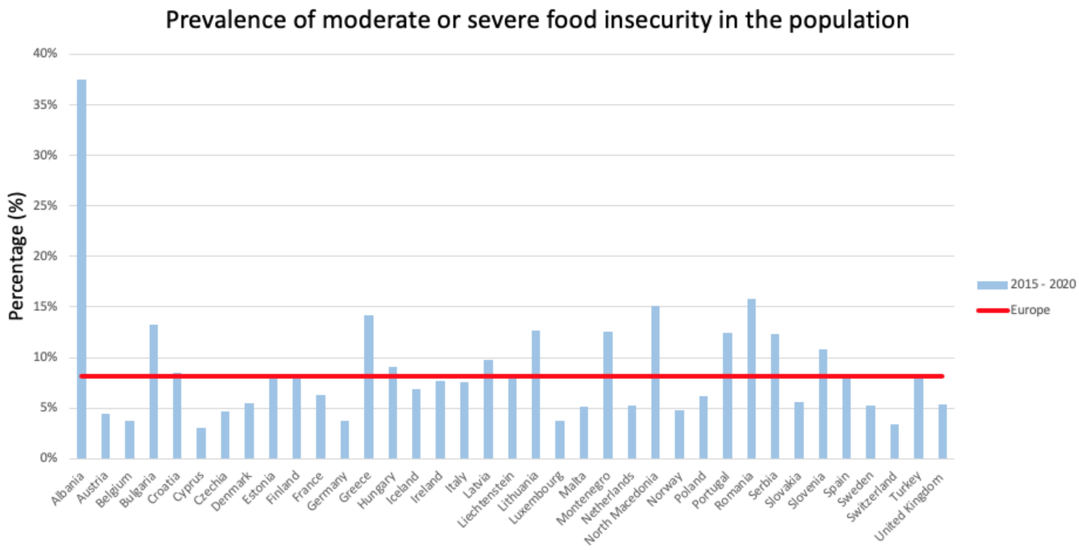
SDG1.4.1 water



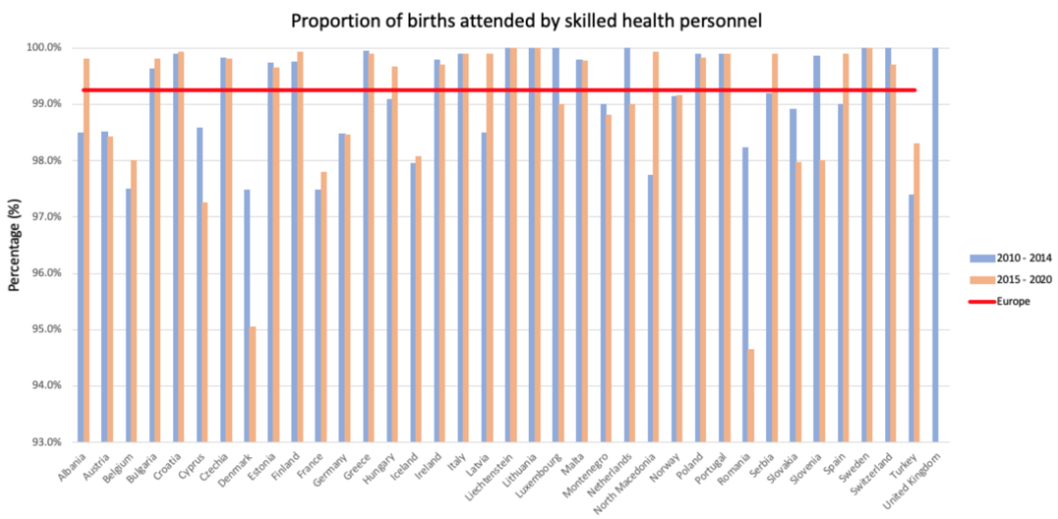
SDG2.1.1



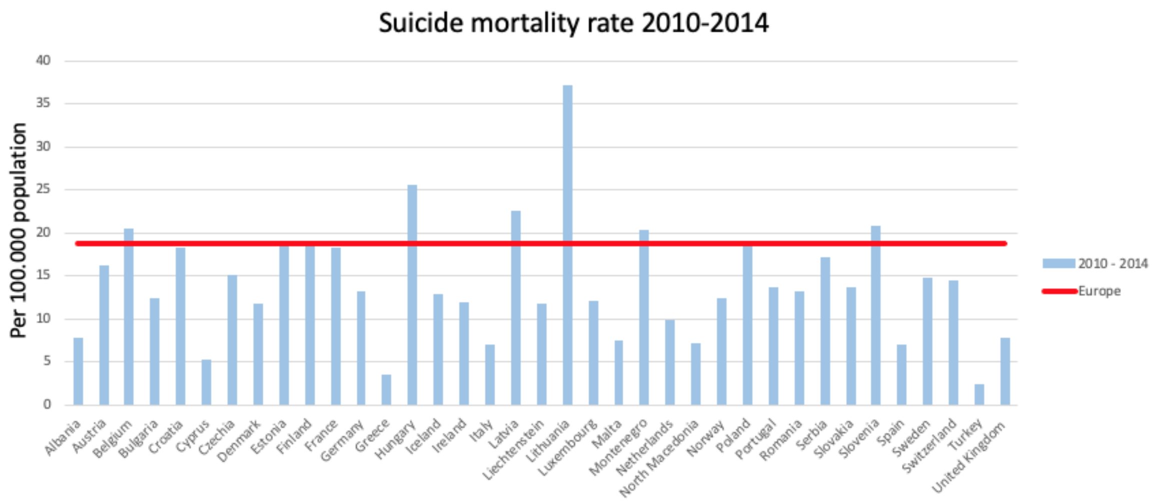
SDG2.1.2

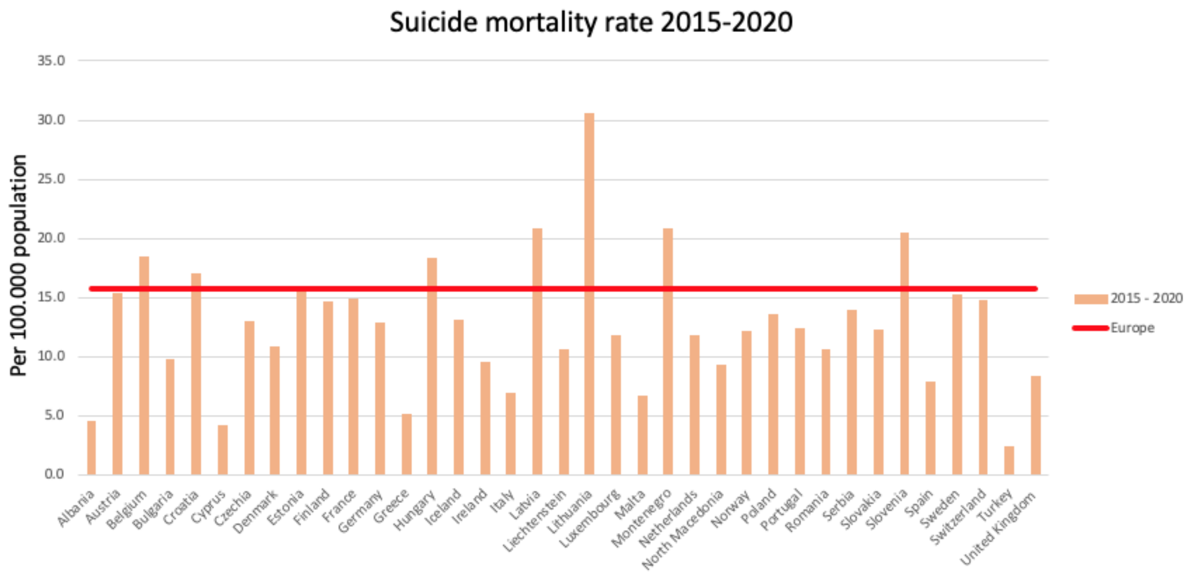


SDG3.1.2



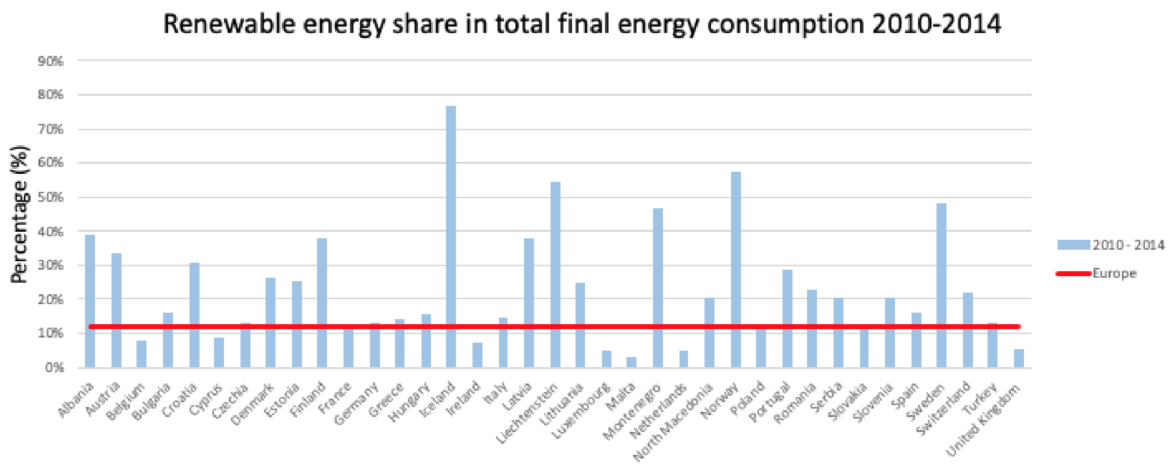
SDG3.4.2



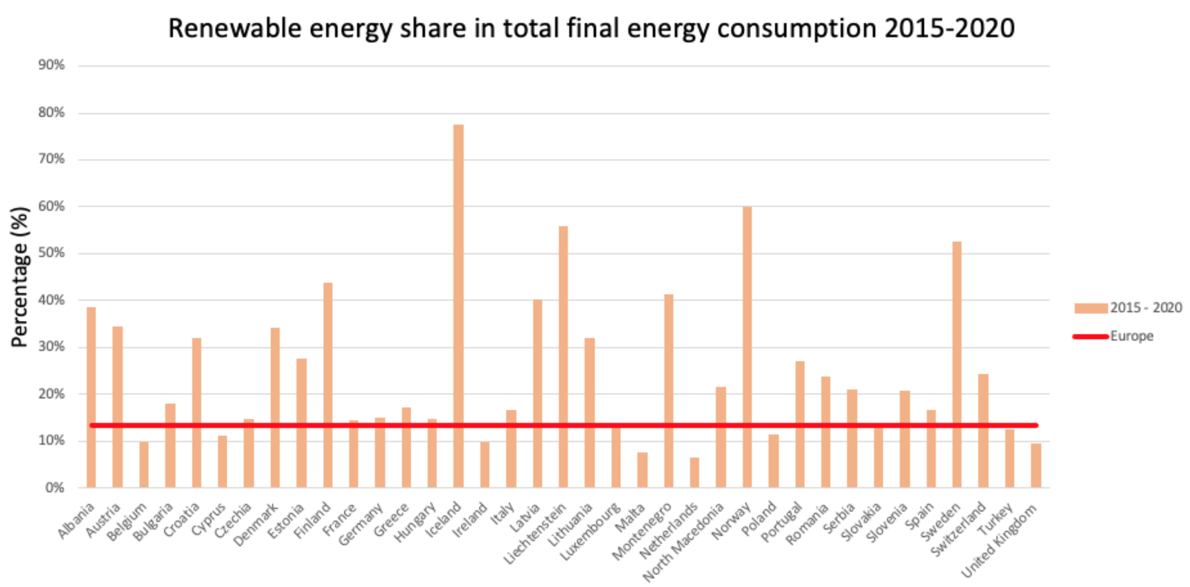


Suicide mortality rate 2015-2020
Per 100.000 population
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Romania
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
2015 - 2020
Europe

SDG7.2.1

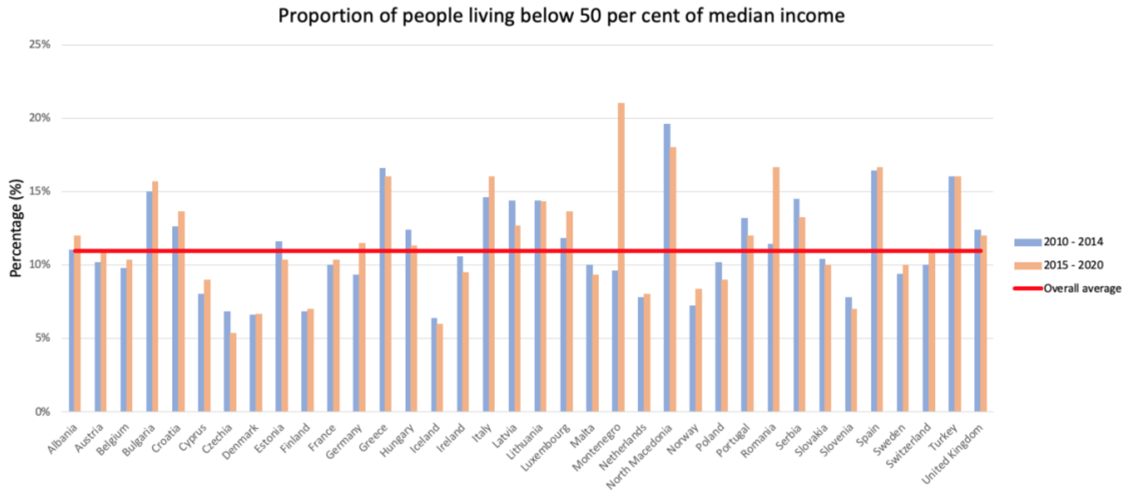


Renewable energy share in total final energy consumption 2010-2014
Percentage (%)
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Romania
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
2010 - 2014
Europe

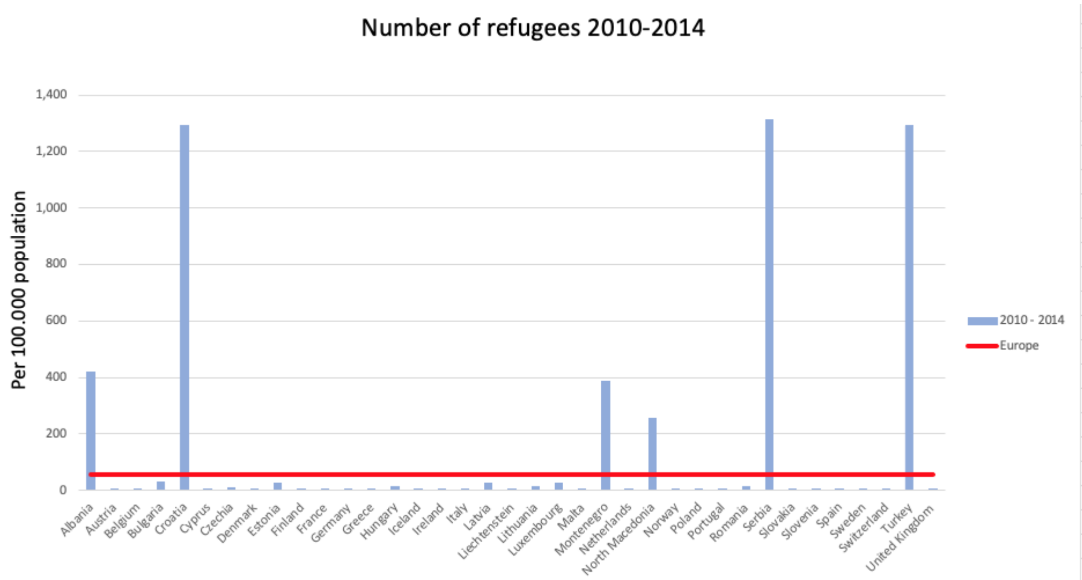


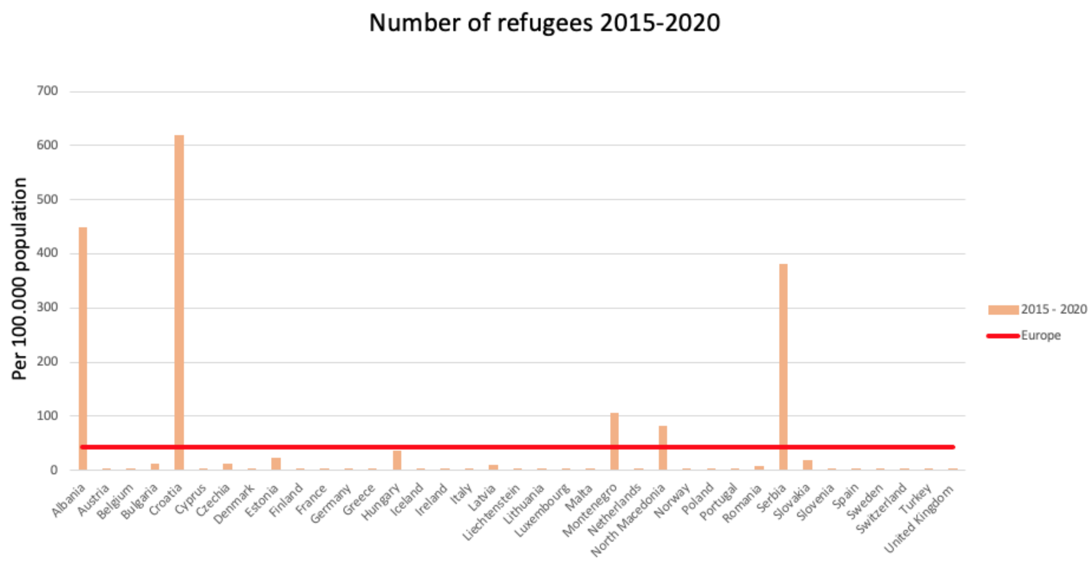
Renewable energy share in total final energy consumption 2015-2020
Percentage (%)
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Romania
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
2015 - 2020
Europe

SDG10.2.1

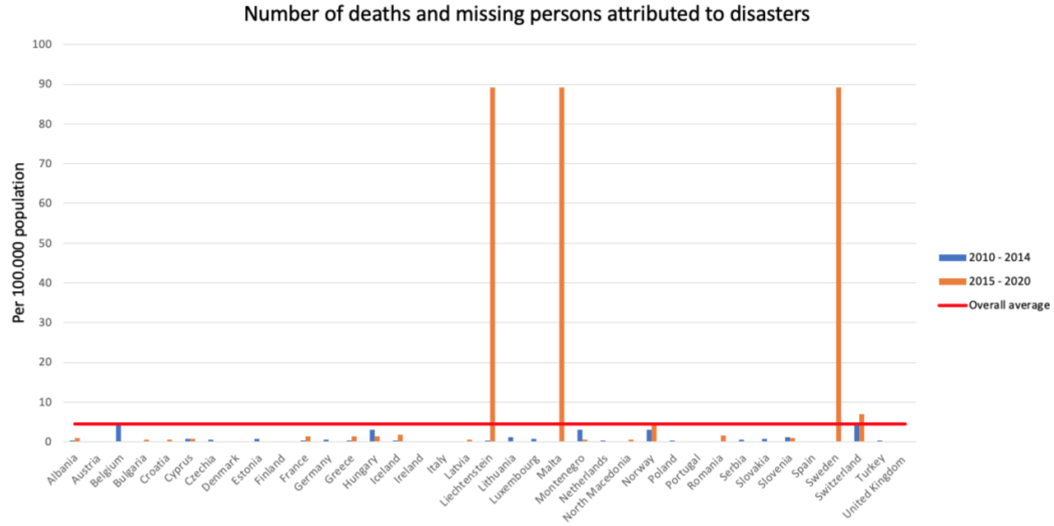


SDG10.7.4

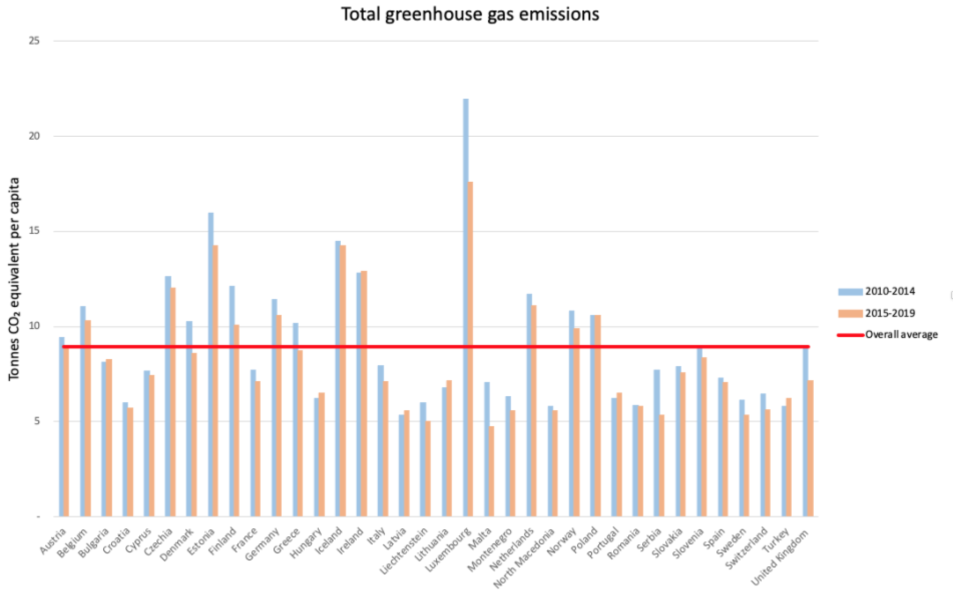




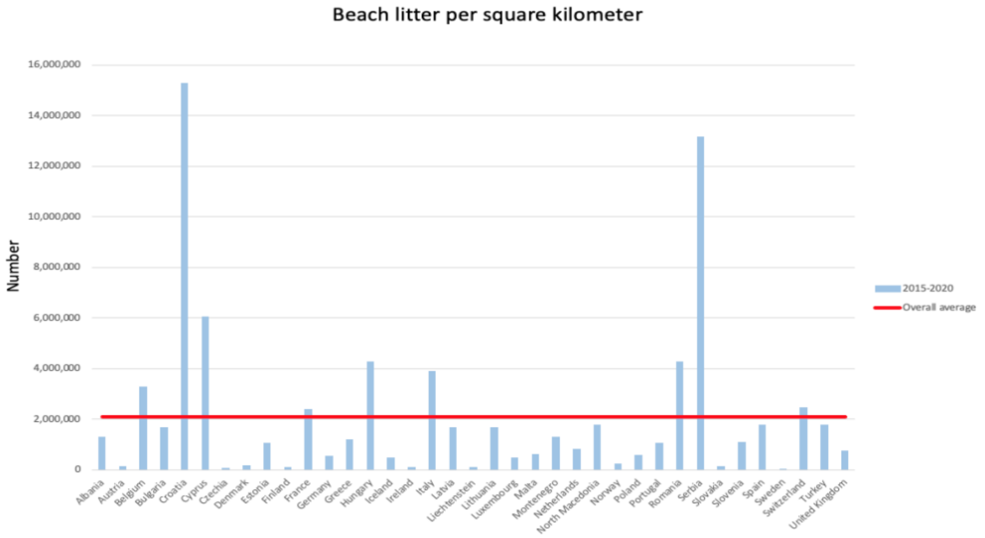
SDG13.1.1



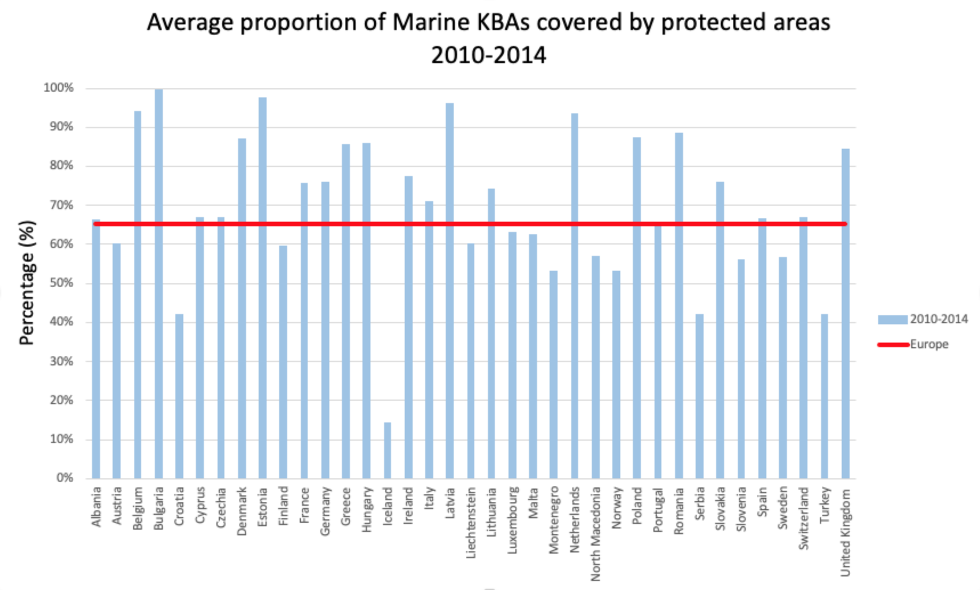
SDG13.2.2

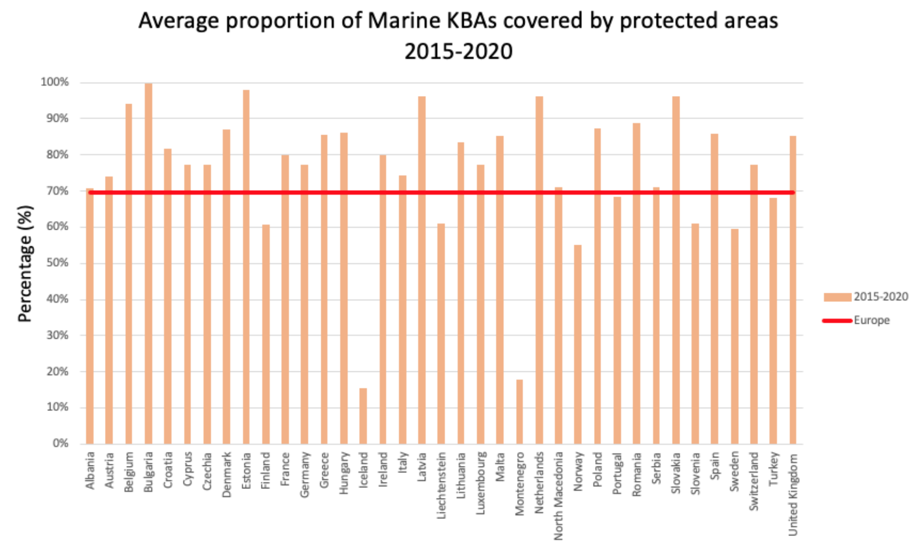


SDG14.1.1

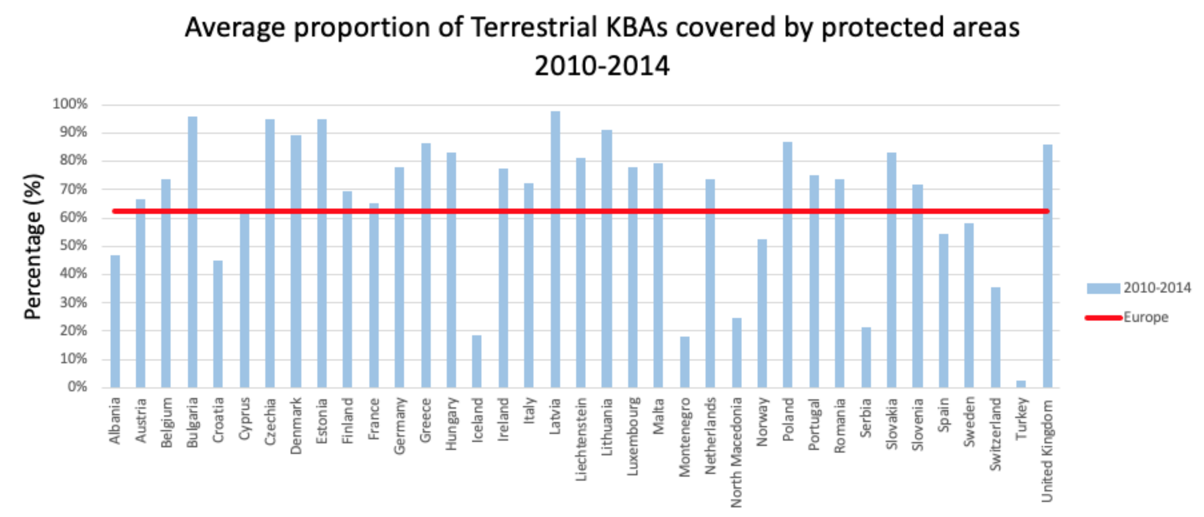


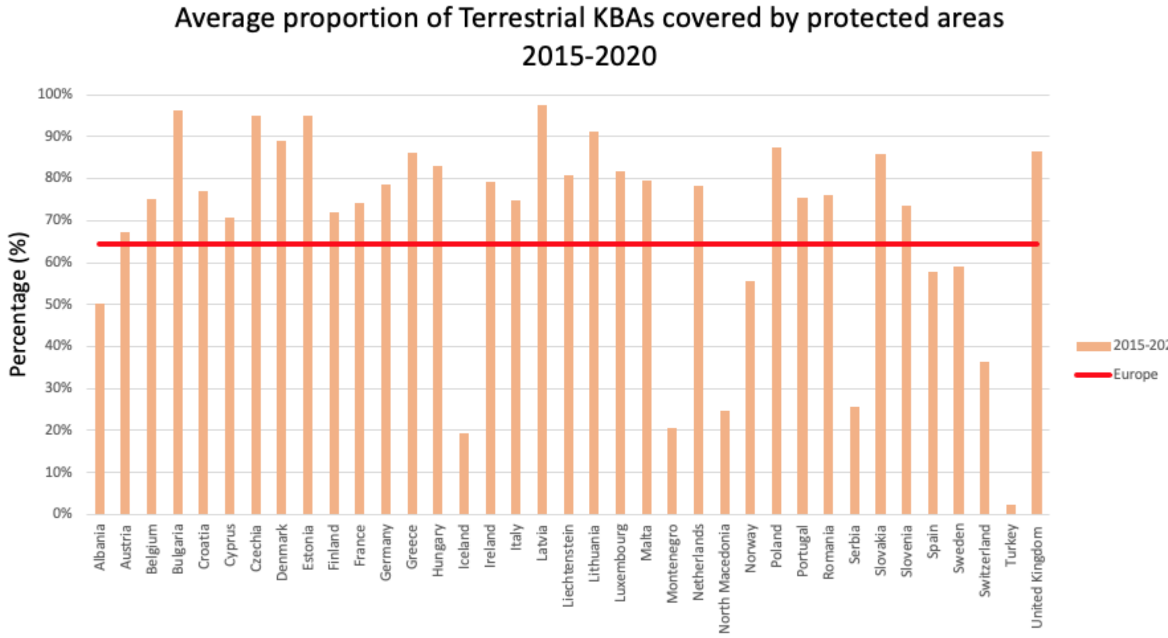
SDG14.5.1



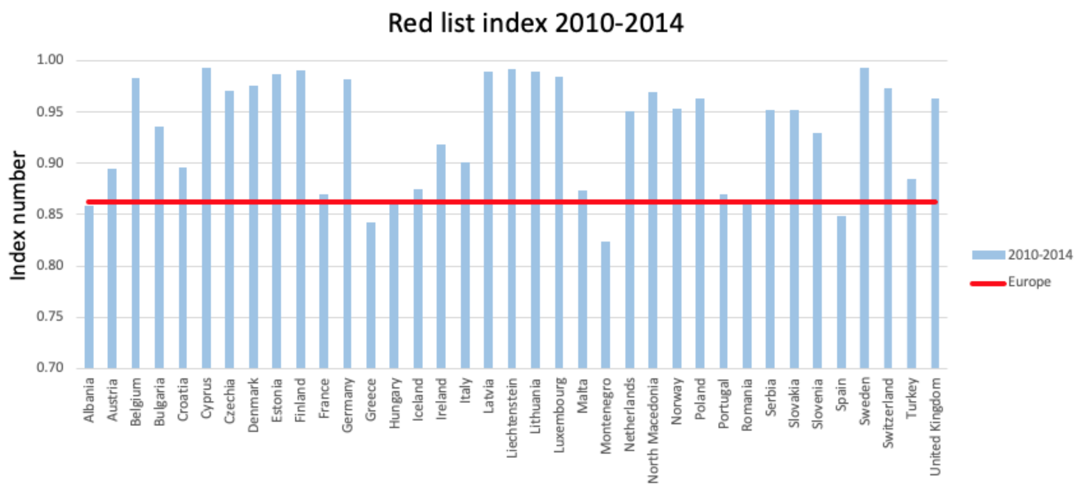


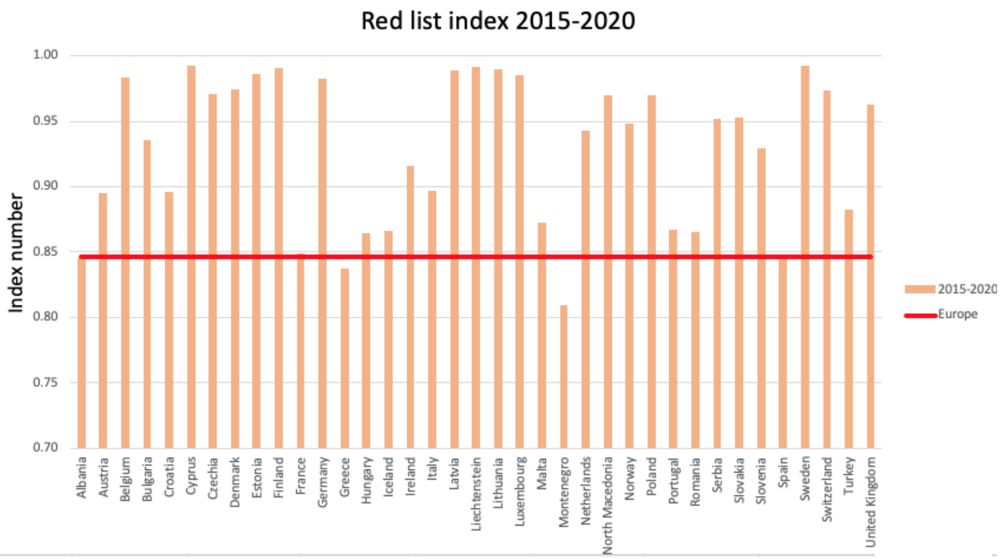
SDG15.1.2




Average proportion of Terrestrial KBAs covered by protected areas 2015-2020
Percentage (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Romania
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
2015-2020
Europe

SDG15.5.1


Red list index 2010-2014
Index number
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Romania
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
2010-2014
Europe


Red list index 2015-2020
Index number
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Albania
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Romania
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
2015-2020
Europe

Appendix M: Descriptive statistics

Dependent variables

2010-2014

	N	Mean	Median	Standard Deviation	Min	Max
SDG_share	333	23.22	22.32	6.87	5.04	56.95
SDG1_share	333	0.03	0.01	0.06	0	0.75
SDG2_share	333	0.52	0.38	0.54	0	4.47
SDG3_share	333	14.99	15.01	5.66	0	46.16
SDG4_Share	333	0.56	0.33	0.92	0	11.98
SDG5_share	333	0.33	0.26	0.3	0	2.4
SDG6_share	333	1.09	0.82	1.28	0	14.56
SDG7_share	333	1.51	1.1	1.66	0	17.55
SDG8_share	333	0.73	0.57	0.8	0	9.6
SDG9_share	333	0.71	0.55	0.71	0	6.48
SDG10_share	333	0.17	0.12	0.19	0	1.26
SDG11_share	333	0.8	0.65	0.71	0	7.59
SDG12_share	333	0.58	0.46	0.56	0	5.06
SDG13_share	333	1.55	1.11	1.62	0	14.45
SDG14_share	333	1.11	0.33	2.94	0	36.33
SDG15_share	333	2.16	1.48	2.83	0	24.91
SDG16_share	333	0.19	0.11	0.22	0	1.34

2015-2020

Indicator	N	Mean	Median	Standard Deviation	Min	Max
SDG_share	333	28.14	27.1	7.62	8.92	64.01
SDG1_share	333	0.04	0.02	0.06	0	0.5
SDG2_share	333	0.73	0.54	1.11	0	16.6
SDG3_share	333	17.21	16.93	6.21	0	57.53
SDG4_share	333	0.75	0.51	1.22	0	17.81
SDG5_share	333	0.46	0.35	0.39	0	2.4
SDG6_share	333	1.41	1.11	1.28	0	13.79
SDG7_share	333	2.3	1.95	1.65	0	11.41
SDG8_share	333	1.03	0.83	0.91	0	10.42
SDG9_share	333	1.12	0.87	1.08	0	11.12
SDG10_share	333	0.22	0.17	0.2	0	1.27
SDG11_share	333	1.07	0.86	0.8	0	5.14
SDG12_share	333	0.91	0.74	0.78	0	7.78
SDG13_share	333	2.12	1.49	1.98	0	16.04
SDG14_share	333	1.34	0.49	3.26	0	39.32
SDG15_share	333	2.62	1.84	3.09	0	25.06
SDG16_share	333	0.28	0.18	0.31	0	2.17

Independent variables

KCI and Relatedness density

2010-2014

Variable	N	Mean	Median	Standard Deviation	Min	Max
KCI	333	42.14	39.28	21.39	3.61	94.15
RD_SDG	333	39.85	40.23	16.27	9.43	82.71
RD_SDG1	333	30.31	3.64	37.79	0	100
RD_SDG2	333	45.26	46.27	13.13	5.09	81.85
RD_SDG3	333	34.45	30.44	24.41	0	95.69
RD_SDG5	333	29.18	14.17	27.92	0	100
RD_SDG6	333	45.47	45.76	22.34	4.29	92.59
RD_SDG7	333	46.02	49.05	27.69	3.07	96.93
RD_SDG8	333	43.89	34.56	30.42	3.18	96.82
RD_SDG9	333	43.73	46.63	31.57	0	100
RD_SDG10	333	29.78	14.14	30.8	0	100
RD_SDG11	333	42.98	41.75	18.57	2.06	88.09
RD_SDG12	333	45.19	45.39	26.63	2.16	97.65
RD_SDG13	333	44.48	45.68	24.09	2.06	97.94
RD_SDG14	333	45.36	66.03	42.15	0	100
RD_SDG15	333	36.68	34.43	25.57	2.07	97.93
RD_SDG16	333	34.98	31.51	33.92	0	100

2015-2020

Variable	N	Mean	Median	Standard Deviation	Min	Max
KCI_2	333	43.25	39.83	23.82	1.22	99.86
RD_SDG	333	40.15	39.52	16.27	8.4	83.2
RD_SDG1	333	33.02	12.92	34.97	0.27	98.85
RD_SDG2	333	43.4	43.58	12.57	7	80.73
RD_SDG3	333	36.69	33.46	23.15	0	93.25
RD_SDG4	333	35.91	31.08	19.37	3.6	89.72
RD_SDG5	333	30.77	26.1	27.98	0.14	99.17
RD_SDG6	333	44.44	44.02	20.61	6.19	87.31
RD_SDG7	333	45.24	45.14	25.65	4.44	94.91
RD_SDG8	333	43.81	37.18	28.34	4.99	93.88
RD_SDG9	333	45.08	43.84	30.18	0.66	98.33
RD_SDG10	333	31.08	14.67	29.76	0.29	99.16
RD_SDG11	333	43.17	41.18	16.07	11.59	79.54
RD_SDG12	333	44.28	43.99	24.42	3.86	95.34
RD_SDG13	333	44.41	44.98	21.65	2.97	96.43
RD_SDG14	333	46.93	63.01	37.3	1.94	97.4
RD_SDG15	333	38.13	35.52	24.04	2.65	96.32
RD_SDG16	333	36.09	24.22	31.25	0.31	98.97

Regional characteristics

2010-2014

Indicator	N	Mean	Median	Standard Deviation	Min	Max
SDG_1.1.1	333	0.01	0	0.01	0	0.07
SDG_1.4.1w	333	0.99	1	0.02	0.92	1
SDG_1.4.1s	333	0.98	0.99	0.03	0.81	1
SDG_2.1.1	333	0.03	0.03	0	0.03	0.05
SDG_3.1.2	333	0.99	0.99	0.01	0.97	1
SDG_3.4.2	333	12.12	13.2	5.98	2.4	37.2
SDG_7.2.1	333	0.16	0.13	0.12	0.03	0.76
SDG_10.2.1	333	0.12	0.11	0.03	0.06	0.2
SDG_10.7.4	333	50.96	0.23	205.58	0.1	1314.07
SDG_13.1.1	333	0.50	0.37	1.01	0.01	4.59
SDG13.2.2	333	8.91	8.83	2.28	5.33	21.94
SDG_14.5.1	333	0.7	0.76	0.23	0.04	1
SDG_15.1.2	333	0.68	0.74	0.24	0.02	0.97
SDG_15.5.1	333	0.92	0.95	0.05	0.82	0.99

2015-2020

Indicator	N	Mean	Median	Standard Deviation	Min	Max
SDG_1.1.1	333	0.01	0	0.01	0	0.06
SDG_1.4.1w	333	0.99	1	0.01	0.94	1
SDG_1.4.1s	333	0.98	0.99	0.03	0.85	1
SDG_2.1.1	333	0.03	0.03	0	0.03	0.05
SDG_2.1.2	333	0.07	0.06	0.04	0.03	0.38
SDG_3.1.2	333	0.99	0.99	0.01	0.95	1
SDG_3.4.2	333	11.12	12.2	4.58	2.4	30.5
SDG_7.2.1	333	0.18	0.15	0.12	0.06	0.77
SDG_10.2.1	333	0.12	0.12	0.03	0.05	0.21
SDG_10.7.4	333	25.43	0.14	90.55	0.08	618.67
SDG_13.1.1	333	3.33	0.1	15.19	0.03	89.26
SDG13.2.2	333	8.46	8.29	2.05	4.75	17.58
SDG_14.1.1	333	1803621.73	1205303	2310076.47	13433	15288985
SDG_14.5.1	333	0.75	0.8	0.23	0.04	1
SDG_15.1.2	333	0.7	0.76	0.24	0.02	0.97
SDG_15.5.1	333	0.92	0.94	0.05	0.81	0.99

Control variables

2010-2014

Variable	N	Mean	Median	Standard Deviation	Min	Max
Population	333	1818274.53	1476895	1572776.63	28252	13562058
GDP	333	24721.98	23420	10693.34	6120	71520
Education	333	26.17	25.5	10.12	7.26	67.2

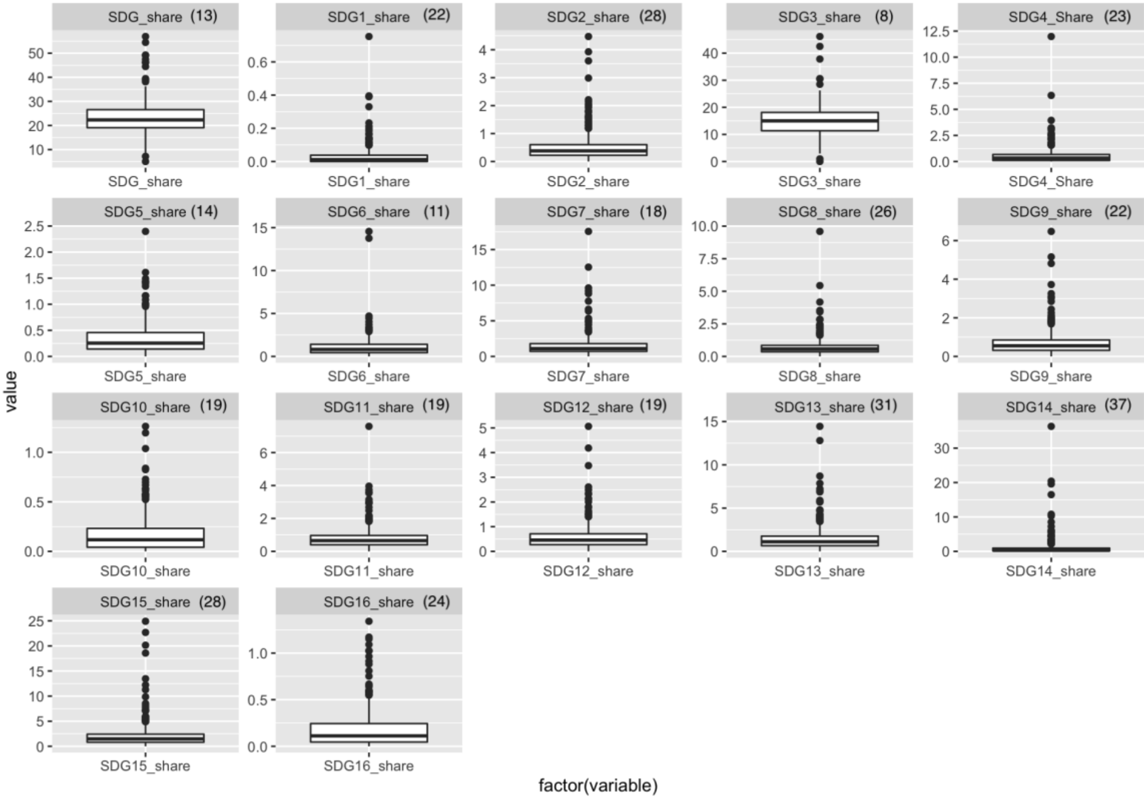
2015-2020

Variable	N	Mean	Median	Standard Deviation	Min	Max
Population	333	1857339.86	1490810	1635472.21	29379	14909132
GDP	333	28421.78	26383	12440.82	7200	78733
Education	333	30.63	30.17	10.77	10.32	72.02

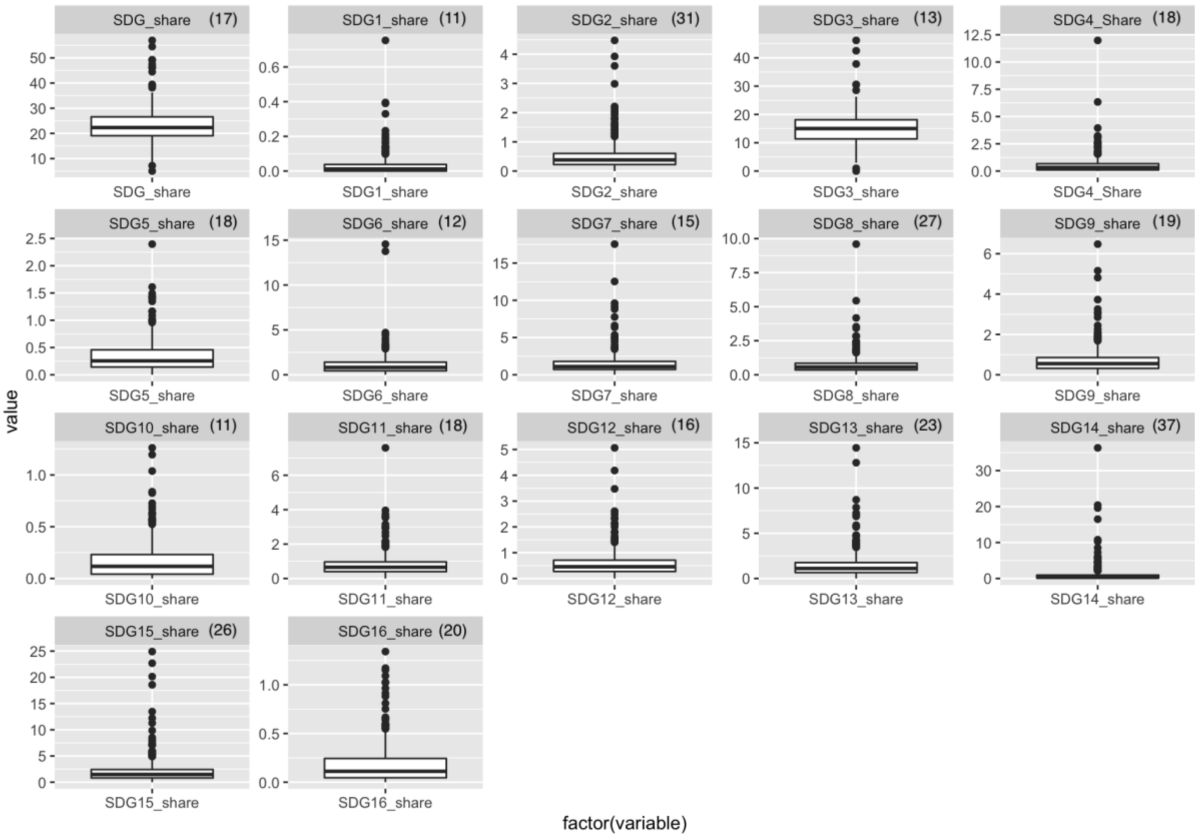
Appendix N: Outlier analysis

Dependent variables

2010-2014



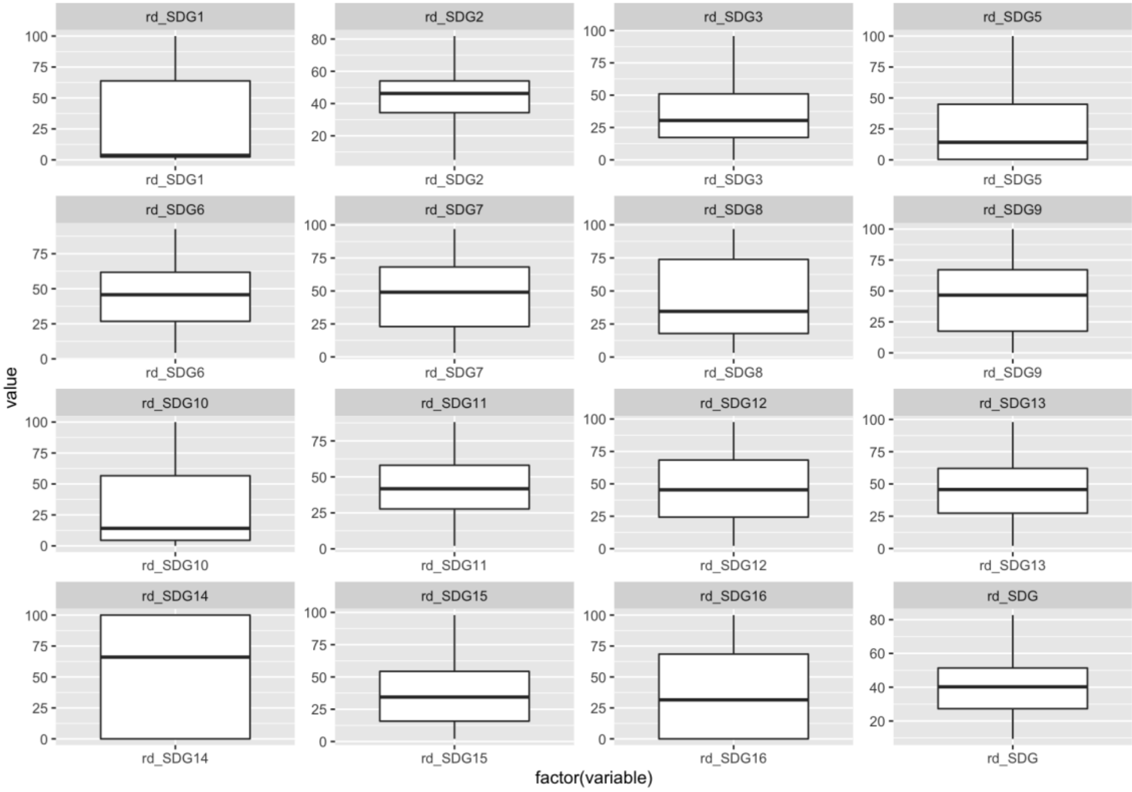
2015-2020



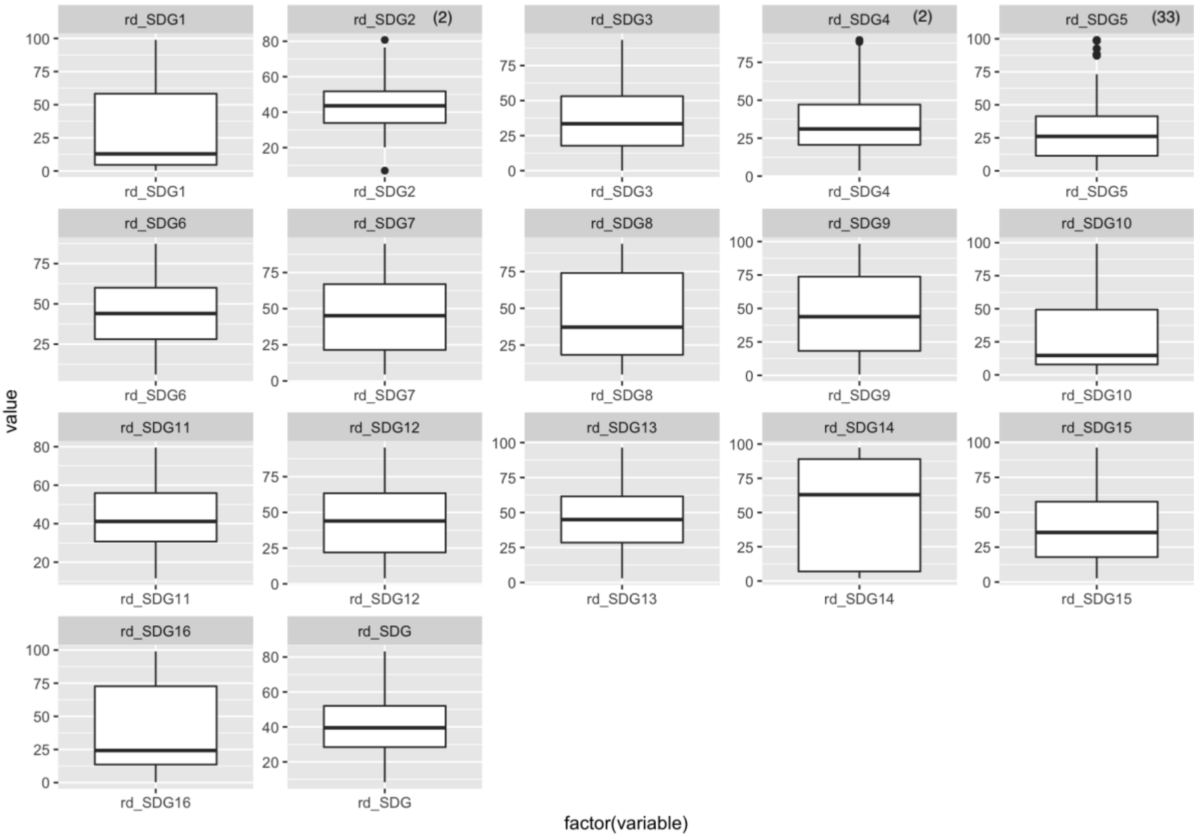
Independent variables

Relatedness density score

2010-2014

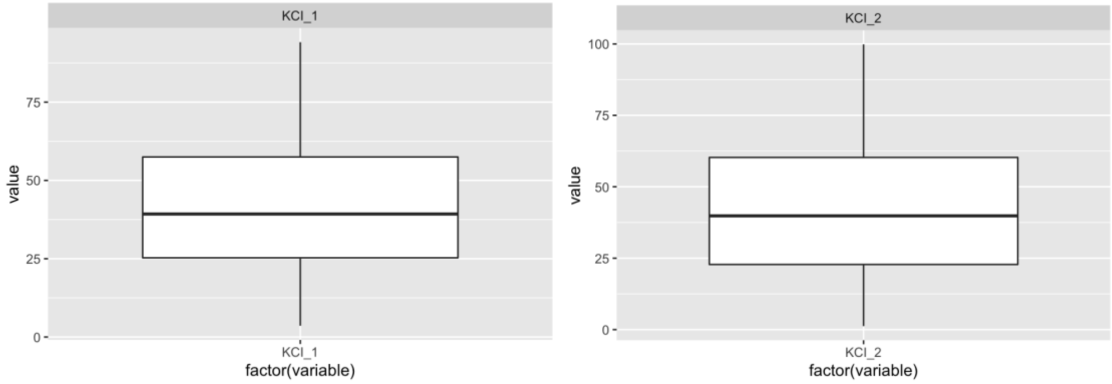


2015-2020



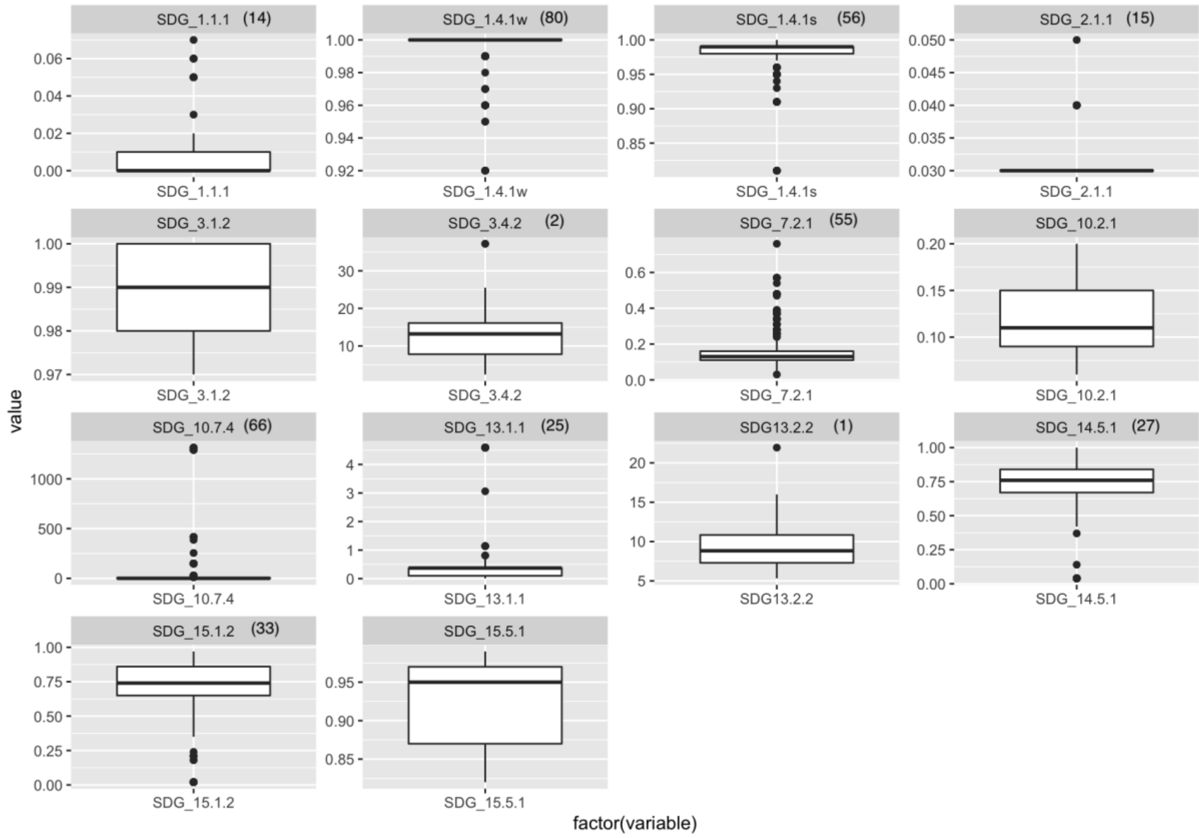
Knowledge complexity

2010-2014 **2015-2020**

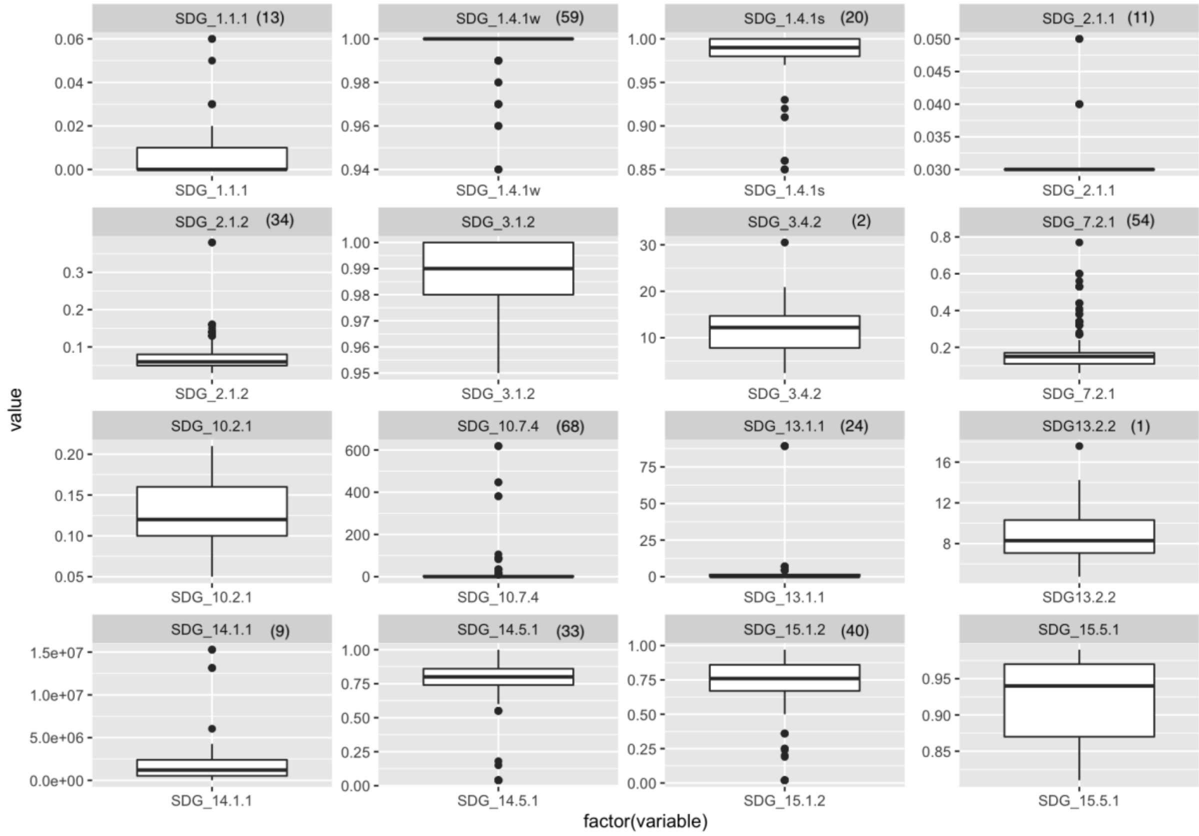


Regional characteristics

2010-2014

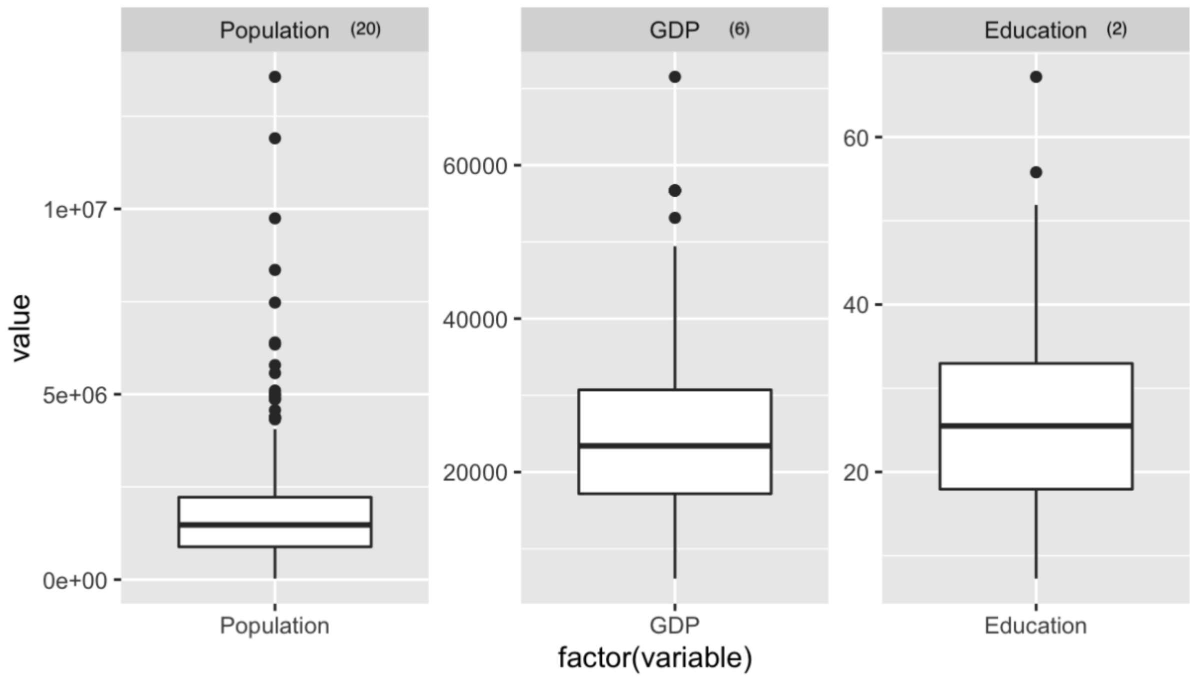


2015-2020

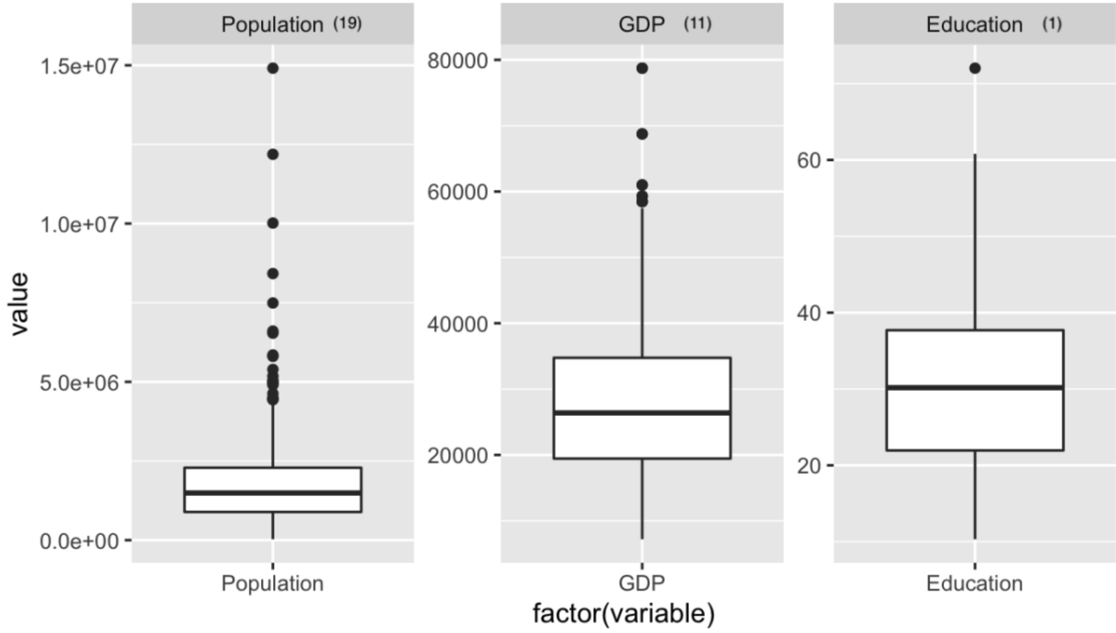


Control variables

2010-2014



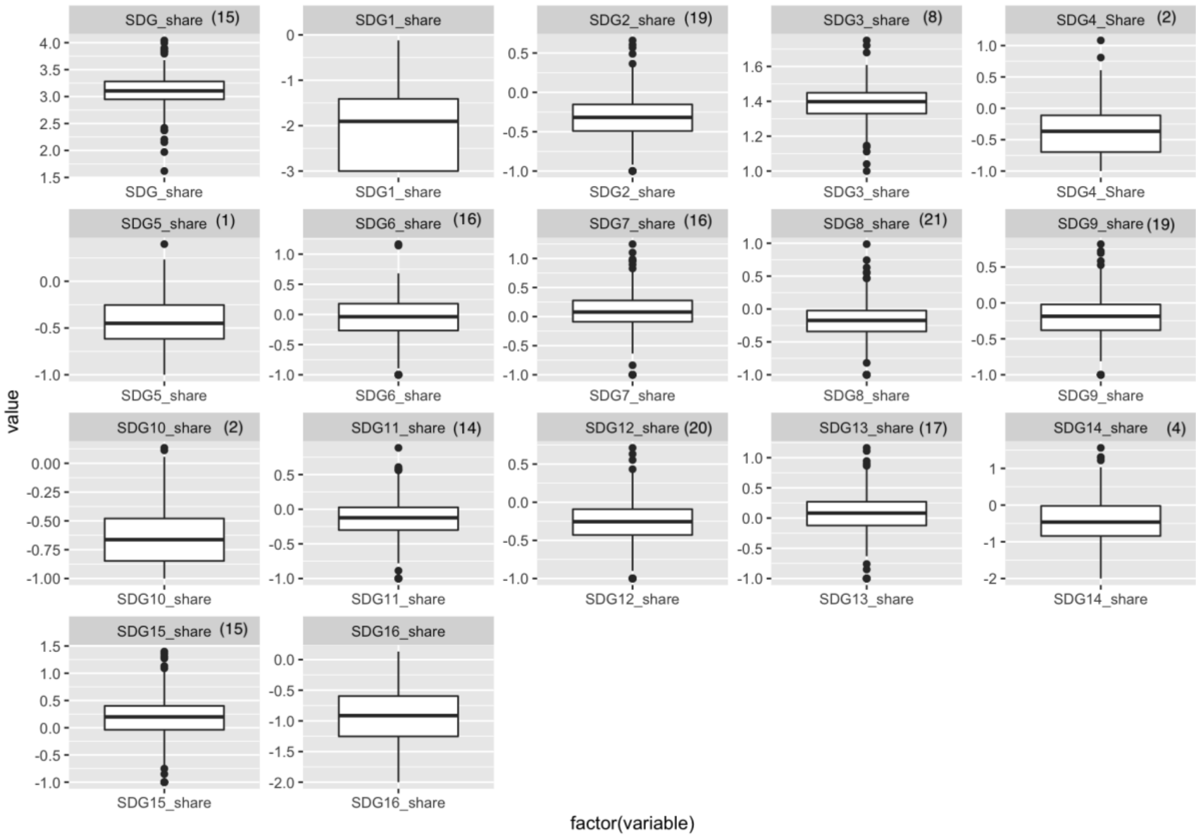
2015-2020



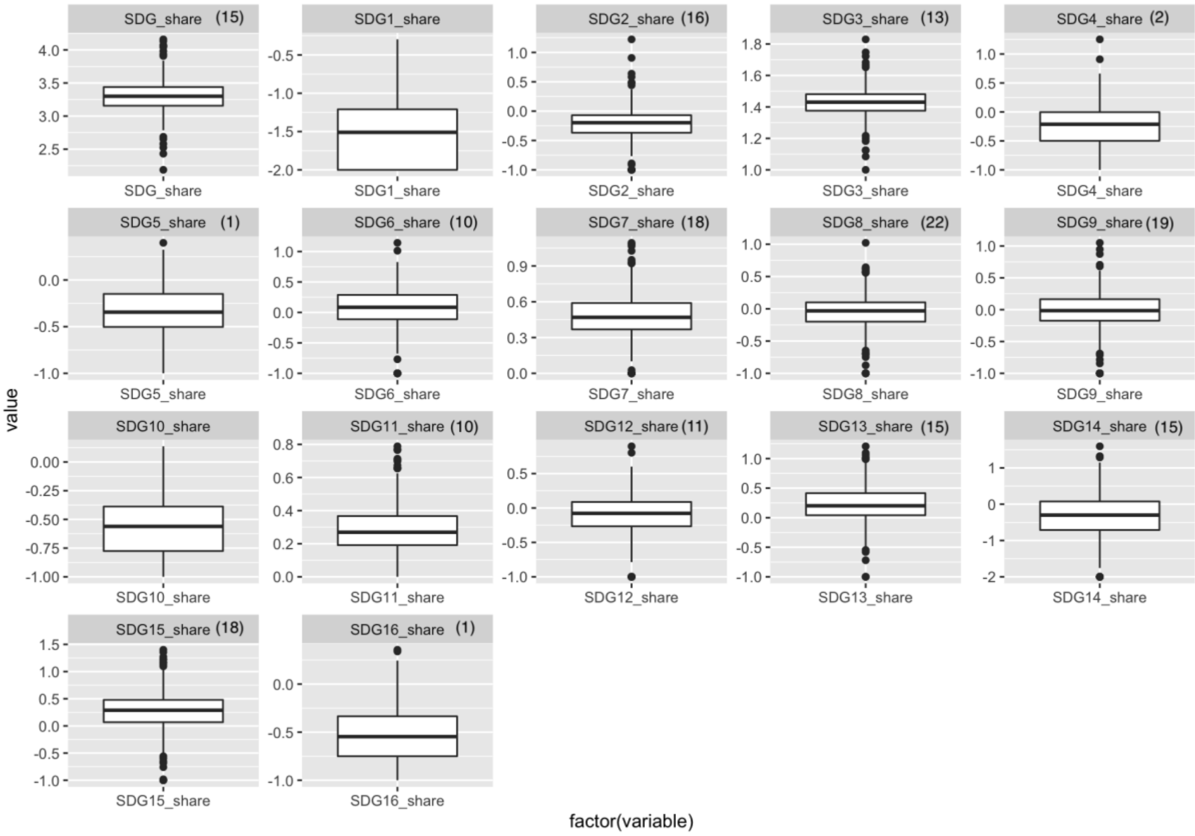
After data transformation

Dependent variables

2010-2014

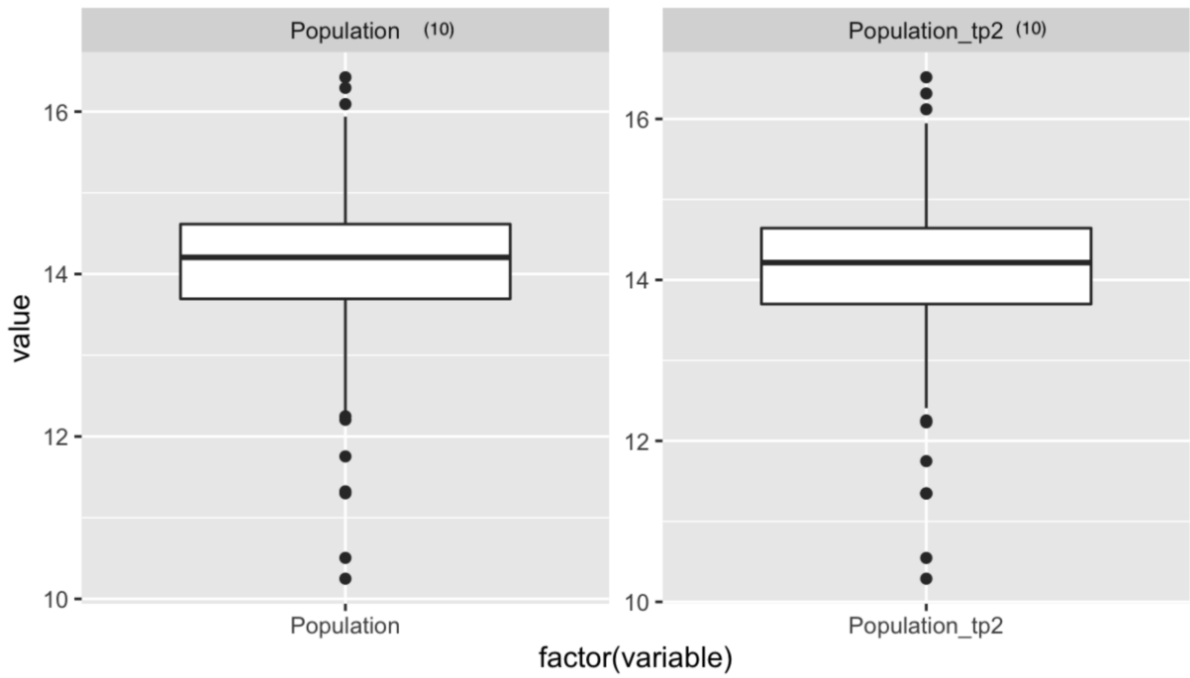


2015-2020



Control Variables

Population



Appendix O: Control models regression analysis

Skewness coefficients 2010-2014

Model	N	SDG_share	KCI	RD_SDG	RC_1	RC_2	Population	GDP	Education
1	277	1.06	0.47	0.23			2.99	0.87	0.40
SDG1	277	7.14	0.47	1.18	3.74	-3.39	2.99	0.87	0.40
SDG2	277	3.31	0.47	-0.04			2.99	0.87	0.40
SDG3	219	0.35	0.63	0.89	-0.17	0.22	3.12	0.63	0.67
SDG4	277	6.85	0.47				2.99	0.87	0.40
SDG5	277	3.13	0.47	0.98			2.99	0.87	0.40
SDG6	277	6.55	0.47	0.087			2.99	0.87	0.40
SDG7	277	4.72	0.47	-0.077		2.05	2.99	0.87	0.40
SDG8	277	5.28	0.47	0.36			2.99	0.87	0.40
SDG9	277	3.02	0.47	0.31			2.99	0.87	0.40
SDG10	249	2.56	0.37	1.08	0.71	13.43	2.59	0.66	0.53
SDG11	277	2.25	0.47	0.28			2.99	0.87	0.40
SDG12	277	2.67	0.47	-0.02			2.99	0.87	0.40
SDG13	114	2.72	0.43	-0.14	3.28	1.89	2.62	0.55	0.54
SDG14	211	7.24	0.30	0.05		-0.13	2.36	0.58	0.50
SDG15	277	4.70	0.47	0.46	-1.60	-0.04	2.99	0.87	0.40
SDG16	277	2.53	0.47	0.86			2.99	0.87	0.40

Skewness coefficients after transformation 2010-2014

Model	N	SDG_share	KCI	RD_SDG	RC_1	RC_2	Population	GDP	Education
1	277	*-0.56*	0.47	0.23			*-0.67*	0.87	0.40
SDG1	277	*0.14 (0.001)*	0.47	*0.36 (1)*	*0.53 (0.001)*	*-0.30 (0.001)*	*-0.67*	0.87	0.40
SDG2	277	*0.21 (0.1)*	0.47	-0.04			*-0.67*	0.87	0.40
SDG3	219	0.35	0.63	0.89	-0.17	0.22	*-0.74*	0.63	0.67
SDG4	277	*0.56 (0.1)*	0.47				*-0.67*	0.87	0.40
SDG5	277	*-0.05 (0.1)*	0.47	0.98			*-0.67*	0.87	0.40
SDG6	277	*-0.50 (0.1)*	0.47	0.087			*-0.67*	0.87	0.40
SDG7	277	*-0.40 (0.1)*	0.47	-0.077		*0.32*	*-0.67*	0.87	0.40
SDG8	277	*0.06 (0.1)*	0.47	0.36			*-0.67*	0.87	0.40
SDG9	277	*-0.08 (0.1)*	0.47	0.31			*-0.67*	0.87	0.40
SDG10	249	*0.48 (0.1)*	0.37	*-0.41 (1)*	0.71	_-0.40*_	*-0.67*	0.66	0.53
SDG11	277	*-0.32 (0.1)*	0.47	0.28			*-0.67*	0.87	0.40
SDG12	277	*-0.30 (0.1)*	0.47	-0.02			*-0.67*	0.87	0.40
SDG13	114	*-0.53 (0.1)*	0.43	-0.14	*0.44*	_-0.75*_	*-0.96*	0.55	0.54
SDG14	211	*-0.38 (0.01)*	0.30	0.05		-0.13	*-0.73*	0.58	0.50
SDG15	277	*-0.15 (0.1)*	0.47	0.46	*-0.42*	-0.04	*-0.67*	0.87	0.40
SDG16	277	*-0.55 (0.01)*	0.47	0.86			*-0.67*	0.87	0.40

*Reciprocal square root transformation

Skewness coefficients 2015-2020

Model	N	SDG_share	KCI	RD_SDG	RC_1	RC_2	Population	GDP	Education
2	277	1.36	0.45	0.24			3.16	0.99	0.43
SDG1	277	4.86	0.45	1.08	1.78	-3.30	3.16	0.99	0.43
SDG2	251	8.98	0.35	0.21		1.18	2.65	0.99	0.46
SDG3	236	0.62	0.52	0.48	-1.52	0.18	3.01	0.96	0.63
SDG4	277	9.20	0.39	0.79			3.16	0.99	0.43
SDG5	277	2.19	0.45	0.12			3.16	0.99	0.43
SDG6	277	4.25	0.45	1.12			3.16	0.99	0.43
SDG7	277	2.24	0.45	-0.01		2.07	3.16	0.99	0.43
SDG8	277	1.83	0.45	0.27			3.16	0.99	0.43
SDG9	277	2.57	0.45	0.14			3.16	0.99	0.43
SDG10	242	1.70	0.42	1.10	0.22	13.952	2.60	0.88	0.53
SDG11	277	1.97	0.45	0.24			3.16	0.99	0.43
SDG12	277	3.61	0.45	0.10			3.16	0.99	0.43
SDG13	135	2.78	0.85	0.03	3.69	0.46	3.23	0.89	0.46
SDG14	211	6.56	0.31	-0.03	3.13	-1.11	2.48	0.93	0.43
SDG15	277	3.79	0.45	0.42	-1.81	-0.05	3.16	0.99	0.43
SDG16	277	2.79	0.45	1.02			3.16	0.99	0.43

Skewness coefficients after transformation 2015-2020

Model	N	SDG_share	KCI	RD_SDG	RC_1	RC_2	Population	GDP	Education
2	277	-0.12	0.45	0.24			-0.64	0.99	0.43
SDG1	277	0.38 (0.01)	0.45	0.19 (1)	0.78 (0.001)	-0.11 (0.001)	-0.64	0.99	0.43
SDG2	251	0.42 (0.1)	0.35	0.21		0.48 (0.001)	-0.64	0.99	0.46
SDG3	236	0.62	0.52	0.48	0.49 (0.01)	0.18	-0.76	0.96	0.63
SDG4	277	0.44 0.1)	0.39	0.79			-0.64	0.99	0.43
SDG5	277	-0.01 (0.1)	0.45	0.12			-0.64	0.99	0.43
SDG6	277	-0.50 (0.1)	0.45	-0.57 (1)			-0.64	0.99	0.43
SDG7	277	0.14 (1)	0.45	-0.01		0.66	-0.64	0.99	0.43
SDG8	277	-0.50 (0.1)	0.45	0.27			-0.64	0.99	0.43
SDG9	277	-0.56 (0.1)	0.45	0.14			-0.64	0.99	0.43
SDG10	242	0.11 (0.1)	0.42	-0.27 (1)	0.22	-0.50*	-0.67	0.88	0.53
SDG11	277	0.66 (1)	0.45	0.24			-0.64	0.99	0.43
SDG12	277	-0.47 (0.1)	0.45	0.10			-0.64	0.99	0.43
SDG13	135	0.37 (0.1)	0.85	0.03	0.79	0.46	-0.71	0.89	0.46
SDG14	211	-0.21 (0.01)	0.31	-0.03	0.49**	0.56 (1)	-0.69	0.93	0.43
SDG15	277	-0.23 (0.1)	0.45	0.42	-0.18	-0.05	-0.64	0.99	0.43
SDG16	277	0.34 (0.1)	0.45	-0.60 (1)			-0.64	0.99	0.43

*Reciprocal square root transformation

** Square root transformation

Regression model 1 without imputed data

	Dependent variable:								
	SDG research share (log)								
	all SDGs	SDG4	SDG5	SDG6	SDG8	SDG9	SDG11	SDG12	SDG16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KCI score	0.001***	0.011***	0.007***	-0.002	0.003**	0.007***	-0.001	0.002	0.013***
	(0.0002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
RD regions-SDG	-0.0004***		0.003***	0.005***	0.005***	0.005***	0.004***	0.006***	0.003***
	(0.0002)		(0.001)	(0.001)	(0.001)	(0.0005)	(0.001)	(0.001)	(0.001)
Population (log)	-0.009	0.108*	0.128***	0.205***	-0.096**	0.049	-0.010	0.059	0.546***
	(0.007)	(0.064)	(0.036)	(0.052)	(0.046)	(0.042)	(0.047)	(0.043)	(0.071)
GDP	-0.00000***	-0.00001***	-0.00000	-0.00001***	0.00000	0.00000	-0.00000**	-0.00001***	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	-0.0003	-0.005	-0.003	0.005*	-0.004*	-0.002	0.001	0.002	0.001
	(0.0003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
Constant	0.178***	-1.102***	-1.546***	-1.334***	0.105	-0.992***	-0.160	-0.872***	-5.106***
	(0.043)	(0.408)	(0.232)	(0.333)	(0.294)	(0.270)	(0.298)	(0.276)	(0.449)
Observations	277	277	277	277	277	277	277	277	277
R^2	0.237	0.149	0.400	0.279	0.293	0.406	0.098	0.325	0.412
Adjusted R^2	0.223	0.136	0.389	0.265	0.280	0.395	0.082	0.312	0.402
Residual Std. Error	0.039 (df = 271)	0.372 (df = 272)	0.205 (df = 271)	0.298 (df = 271)	0.267 (df = 271)	0.244 (df = 271)	0.272 (df = 271)	0.248 (df = 271)	0.401 (df = 271)
F Statistic	16.797*** (df = 5; 271)	11.860*** (df = 4; 272)	36.088*** (df = 5; 271)	20.934*** (df = 5; 271)	22.471*** (df = 5; 271)	37.015*** (df = 5; 271)	5.918*** (df = 5; 271)	26.046*** (df = 5; 271)	38.033*** (df = 5; 271)
Note:	*p<0.1; **p<0.05; ***p<0.01								

Regression model 2 without imputed data

	Dependent variable:								
	SDG research share (log)								
	all SDGs	SDG4	SDG5	SDG6	SDG8	SDG9	SDG11	SDG12	SDG16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
KCI score	0.002***	0.007***	0.007***	-0.002	0.002**	0.005***	-0.001**	0.001	0.007***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RD regions-SDG	-0.001	0.006***	0.003***	0.504***	0.006***	0.006***	0.003***	0.007***	0.075***
	(0.0004)	(0.001)	(0.001)	(0.071)	(0.001)	(0.0005)	(0.001)	(0.001)	(0.024)
Population (log)	-0.044**	-0.144***	0.081**	0.050	-0.096**	0.075*	-0.026	0.053	0.149***
	(0.018)	(0.055)	(0.034)	(0.046)	(0.042)	(0.041)	(0.023)	(0.041)	(0.034)
GDP	-0.173***	-0.464***	-0.138	-0.521***	0.014	0.057	-0.148**	-0.040	-0.067
	(0.049)	(0.151)	(0.093)	(0.129)	(0.117)	(0.114)	(0.065)	(0.116)	(0.093)
Education	-0.0002	-0.003	-0.001	0.003	-0.0001	-0.001	0.002	-0.001	0.001
	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Constant	2.407***	2.273***	-0.602	1.257**	0.152	-1.183**	0.962***	-0.547	-1.610***
	(0.210)	(0.638)	(0.396)	(0.559)	(0.498)	(0.487)	(0.274)	(0.493)	(0.395)
Observations	277	277	277	277	277	277	277	277	277
R^2	0.195	0.266	0.436	0.273	0.341	0.424	0.173	0.368	0.412
Adjusted R^2	0.181	0.252	0.425	0.259	0.329	0.414	0.158	0.356	0.401
Residual Std. Error (df = 271)	0.102	0.313	0.192	0.267	0.242	0.236	0.133	0.238	0.193
F Statistic (df = 5; 271)	13.171***	19.609***	41.869***	20.334***	28.070***	39.941***	11.328***	31.525***	37.987***
Note:	*p<0.1; **p<0.05; ***p<0.01								

Regression model 3 without imputed data

	Dependent variable:							
	SDG research share (log)							
	SDG1	SDG2	SDG3	SDG7	SDG10	SDG13	SDG14	SDG15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KCI score	0.007^{**}	0.003^{**}	0.129^{***}	-0.001	0.006^{***}	0.003	0.012^{***}	0.001
	(0.003)	(0.001)	(0.026)	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)
RD regions-SDG	0.096	0.005^{***}	-0.114^{***}	0.006^{***}	0.070^{***}	0.006^{***}	0.006^{***}	0.007^{***}
	(0.068)	(0.001)	(0.016)	(0.001)	(0.022)	(0.001)	(0.001)	(0.001)
Regional characteristic 1	0.042		-25.928		-0.671	-0.075		-0.023
	(0.097)		(33.169)		(0.450)	(0.046)		(0.076)
Regional characteristic 2	-0.212^{***}		-0.234^{***}	0.044	0.001	-0.766	-0.665^{*}	0.628
	(0.068)		(0.057)	(0.082)	(0.014)	(0.768)	(0.395)	(0.481)
Population (log)	0.671^{***}	0.002	1.281	0.102^{**}	0.116^{***}	0.045	0.357^{***}	-0.090
	(0.115)	(0.052)	(0.981)	(0.050)	(0.036)	(0.069)	(0.124)	(0.060)
GDP	0.00001^{**}	-0.00001^{***}	-0.0001^{***}	-0.00000	-0.00000	0.00001^{***}	-0.00001	-0.00000
	(0.00001)	(0.00000)	(0.00005)	(0.00000)	(0.00000)	(0.00000)	(0.00001)	(0.00000)
Education	0.012^{**}	0.005^{*}	0.078	-0.001	0.004^{**}	-0.005	0.002	0.001
	(0.006)	(0.003)	(0.055)	(0.003)	(0.002)	(0.004)	(0.006)	(0.003)
Constant	-7.623^{***}	-0.556^{*}	34.951	-0.656^{**}	-1.725^{***}	-0.539	-2.639^{***}	-0.073
	(0.775)	(0.324)	(34.781)	(0.325)	(0.232)	(0.480)	(0.857)	(0.542)
Observations	277	277	219	277	249	114	211	277
R^2	0.341	0.101	0.326	0.285	0.398	0.315	0.215	0.234
Adjusted R^2	0.324	0.085	0.304	0.269	0.381	0.270	0.192	0.214
Residual Std. Error	0.656 (df = 269)	0.295 (df = 271)	4.610 (df = 211)	0.290 (df = 270)	0.190 (df = 241)	0.269 (df = 106)	0.623 (df = 204)	0.343 (df = 269)
F Statistic	19.904^{***} (df = 7; 269)	6.100^{***} (df = 5; 271)	14.573^{***} (df = 7; 211)	17.894^{***} (df = 6; 270)	22.786^{***} (df = 7; 241)	6.956^{***} (df = 7; 106)	9.332^{***} (df = 6; 204)	11.752^{***} (df = 7; 269)

Note: *p<0.1; **p<0.05; ***p<0.01

Some variables in the Regional Characteristics were log or sqrt transformed, see Appendix O

Regression model 4 without imputed data

	Dependent variable:							
	SDG research share (log)							
	SDG1	SDG2	SDG3	SDG7	SDG10	SDG13	SDG14	SDG15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KCI score	0.002	0.001	0.090^{***}	-0.0001	0.005^{***}	0.002	0.009^{***}	0.002^{*}
	(0.001)	(0.001)	(0.024)	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)
RD regions-SDG	0.197^{***}	0.007^{***}	-0.117^{***}	0.004^{***}	0.108^{***}	0.006^{***}	0.006^{***}	0.008^{***}
	(0.040)	(0.002)	(0.019)	(0.0004)	(0.029)	(0.001)	(0.001)	(0.001)
Regional characteristic 1	0.030		0.473		-0.687^{*}	0.048	0.0002^{***}	-0.066
	(0.035)		(0.546)		(0.400)	(0.030)	(0.0001)	(0.070)
Regional characteristic 2	-0.046	-0.235^{*}	-0.358^{***}	0.091^{*}	-0.012	-0.004	3.527^{***}	0.370
	(0.030)	(0.119)	(0.078)	(0.050)	(0.012)	(0.012)	(1.151)	(0.416)
Population (log)	0.325^{***}	0.044	0.408	0.041	0.124^{***}	-0.219^{***}	0.136	-0.045
	(0.054)	(0.053)	(0.998)	(0.029)	(0.036)	(0.065)	(0.114)	(0.057)
GDP	0.00000^{*}	-0.00001^{***}	-0.0001^{***}	0.00000	0.00000	0.00001^{*}	-0.00001	-0.00000
	(0.00000)	(0.00000)	(0.00004)	(0.00000)	(0.00000)	(0.00000)	(0.00001)	(0.00000)
Education	0.006^{**}	0.004	0.122^{**}	0.0002	0.002	-0.004	0.007	-0.002
	(0.003)	(0.003)	(0.052)	(0.001)	(0.002)	(0.004)	(0.006)	(0.003)
Constant	-4.136^{***}	-0.917^{***}	19.791^{***}	0.118	-1.674^{***}	1.292^{***}	-2.246^{***}	-0.034
	(0.355)	(0.343)	(6.639)	(0.180)	(0.224)	(0.439)	(0.733)	(0.493)
Observations	277	251	236	277	242	135	220	277
R^2	0.320	0.135	0.239	0.301	0.396	0.307	0.220	0.279
Adjusted R^2	0.303	0.114	0.216	0.285	0.378	0.269	0.195	0.261
Residual Std. Error	0.312 (df = 269)	0.290 (df = 244)	5.199 (df = 228)	0.163 (df = 270)	0.190 (df = 234)	0.259 (df = 127)	0.588 (df = 212)	0.325 (df = 269)
F Statistic	18.117^{***} (df = 7; 269)	6.361^{***} (df = 6; 244)	10.240^{***} (df = 7; 228)	19.363^{***} (df = 6; 270)	21.883^{***} (df = 7; 234)	8.031^{***} (df = 7; 127)	8.555^{***} (df = 7; 212)	14.907^{***} (df = 7; 269)

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix P: Regression models control variables

Regression model only control variables (2010-2014)

	Dependent variable:																
	SDG research share (log)																
	All SDGs	SDG1	SDG2	SDG3	SDG4	SDG5	SDG6	SDG7	SDG8	SDG9	SDG10	SDG11	SDG12	SDG13	SDG14	SDG15	SDG16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
Population (log)	-0.035***	0.115***	0.030	0.005	-0.081***	-0.007	0.051**	0.020	-0.048**	-0.013	0.016	0.006	0.036*	-0.060***	-0.030	-0.057**	0.100***
	(0.007)	(0.023)	(0.019)	(0.006)	(0.023)	(0.017)	(0.020)	(0.013)	(0.020)	(0.021)	(0.015)	(0.009)	(0.019)	(0.021)	(0.042)	(0.024)	(0.030)
GDP	-0.00000	0.00001**	-0.00001***	-0.00000	-0.00000	0.00000	-0.00001***	-0.00000	-0.00000	0.00000	0.00001***	-0.00000***	-0.00000	0.00000	-0.00000	-0.00000	0.00001**
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	0.0003	0.010***	0.003*	0.001	0.004*	0.008***	-0.001	-0.0003	0.002	0.003	0.007***	0.001	-0.001	0.001	0.011**	0.001	0.016***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
Constant	1.953***	-3.629***	-0.574**	1.333***	0.806**	-0.556**	-0.426	0.230	0.603**	0.044	-1.162***	0.267**	-0.492*	0.957***	-0.168	1.150***	-2.868***
	(0.101)	(0.324)	(0.274)	(0.091)	(0.334)	(0.241)	(0.292)	(0.181)	(0.283)	(0.304)	(0.209)	(0.134)	(0.275)	(0.302)	(0.598)	(0.351)	(0.429)
Observations	333	333	333	333	333	333	333	333	333	333	333	333	333	333	333	333	333
R^2	0.080	0.240	0.031	0.007	0.047	0.158	0.088	0.012	0.022	0.017	0.246	0.035	0.026	0.039	0.025	0.026	0.232
Adjusted R^2	0.072	0.234	0.022	-0.002	0.038	0.151	0.080	0.003	0.013	0.008	0.239	0.027	0.018	0.030	0.016	0.018	0.225
Residual Std. Error (df = 329)	0.107	0.343	0.290	0.096	0.353	0.255	0.309	0.191	0.299	0.321	0.221	0.142	0.291	0.320	0.632	0.371	0.453
F Statistic (df = 3; 329)	9.526***	34.717***	3.517**	0.825	5.420***	20.613***	10.636***	1.334	2.446*	1.908	35.806***	4.028***	2.979**	4.472***	2.853**	2.979**	33.059***
Note:																	*p<0.1; **p<0.05; ***p<0.01

Regression model without regional characteristic variables (2010-2014)

	Dependent variable:							
	SDG research share (log)							
	SDG1	SDG2	SDG3	SDG7	SDG10	SDG13	SDG14	SDG15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KCI score	0.002*	-0.001	0.001***	-0.001*	0.004***	0.002**	0.007***	0.0002
	(0.001)	(0.001)	(0.0003)	(0.0005)	(0.001)	(0.001)	(0.002)	(0.001)
RD regions-SDG	0.004***	0.006***	-0.001***	0.004***	0.002***	0.007***	0.006***	0.008***
	(0.001)	(0.001)	(0.0002)	(0.0004)	(0.0004)	(0.001)	(0.001)	(0.001)
Population (log)	0.135***	0.009	0.013**	0.005	0.037***	-0.046**	0.061	-0.033
	(0.022)	(0.020)	(0.006)	(0.011)	(0.013)	(0.019)	(0.040)	(0.022)
GDP	0.00000*	-0.00000***	-0.00000**	0.00000	0.00000**	0.00001***	-0.00000	-0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	0.005*	0.004*	0.001	0.001	0.001	-0.002	0.001	-0.003
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)	(0.002)
Constant	-3.875***	-0.501*	1.246***	0.229	-1.458***	0.359	-1.701***	0.525*
	(0.311)	(0.276)	(0.089)	(0.160)	(0.189)	(0.279)	(0.582)	(0.317)
Observations	333	333	333	333	333	333	333	333
R^2	0.356	0.088	0.127	0.284	0.444	0.265	0.187	0.277
Adjusted R^2	0.346	0.074	0.113	0.273	0.435	0.254	0.174	0.266
Residual Std. Error (df = 327)	0.317	0.282	0.091	0.163	0.191	0.280	0.579	0.320
F Statistic (df = 5; 327)	36.118***	6.339***	9.491***	25.894***	52.140***	23.600***	15.027***	25.028***
Note:								*p<0.1; **p<0.05; ***p<0.01

Regression model only control variables (2015-2020)

	Dependent variable:																
	SDG research share (log)																
	All SDGs	SDG1	SDG2	SDG3	SDG4	SDG5	SDG6	SDG7	SDG8	SDG9	SDG10	SDG11	SDG12	SDG13	SDG14	SDG15	SDG16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
Population (log)	-0.037***	0.225***	0.020	-0.001	0.042	0.033*	0.101***	0.069***	-0.054***	-0.025	0.020	0.011	0.034*	-0.028	0.022	-0.050*	0.179***
	(0.008)	(0.045)	(0.020)	(0.006)	(0.359)	(0.017)	(0.023)	(0.022)	(0.020)	(0.020)	(0.015)	(0.019)	(0.020)	(0.023)	(0.045)	(0.026)	(0.032)
GDP	0.00000	0.00003***	-0.00000**	0.00000	-0.00000	0.00000***	-0.00001***	-0.00000	0.00001***	0.00001***	0.00001***	-0.00000	-0.00000**	0.00001**	-0.00000	-0.00000	0.00001***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	-0.0002	0.014***	0.004*	0.0004	0.001	0.007***	-0.002	-0.001	-0.001	0.001	0.007***	0.0001	0.0005	0.003	0.012**	0.003	0.016***
	(0.001)	(0.005)	(0.002)	(0.001)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)	(0.003)	(0.003)
Constant	1.866***	-6.291***	-0.588**	1.373***	-0.494	-1.216***	-1.265***	-0.768**	0.458	-0.041	-1.274***	-0.227	-0.637**	0.273	-1.055	0.848**	-4.187***
	(0.118)	(0.648)	(0.287)	(0.092)	(0.953)	(0.247)	(0.323)	(0.314)	(0.286)	(0.287)	(0.211)	(0.271)	(0.281)	(0.331)	(0.648)	(0.370)	(0.452)
Observations	333	333	333	333	333	333	333	333	333	333	333	333	333	333	333	333	333
R^2	0.058	0.279	0.019	0.011	0.001	0.154	0.118	0.049	0.052	0.064	0.234	0.009	0.036	0.047	0.027	0.014	0.263
Adjusted R^2	0.049	0.273	0.010	0.002	-0.009	0.147	0.110	0.040	0.044	0.055	0.227	0.0003	0.027	0.038	0.018	0.005	0.256
Residual Std. Error (df = 329)	0.125	0.687	0.304	0.098	0.400	0.261	0.342	0.332	0.303	0.304	0.224	0.287	0.298	0.351	0.687	0.393	0.479
F Statistic (df = 3; 329)	6.734***	42.514***	2.161*	1.272	0.062	20.037***	14.625***	5.644***	6.060***	7.455***	33.492***	1.033	4.054***	5.428***	3.017**	1.538	39.064***

Note: *p<0.1; **p<0.05; ***p<0.01

Regression model without regional characteristic variables (2015-2020)

	Dependent variable:							
	SDG research share (log)							
	SDG1	SDG2	SDG3	SDG7	SDG10	SDG13	SDG14	SDG15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
KCI score	0.006***	0.002*	0.002***	-0.002**	0.004***	0.002*	0.008***	-0.002
	(0.002)	(0.001)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
RD regions-SDG	0.003***	0.006***	-0.002***	0.006***	0.002***	0.007***	0.006***	0.007***
	(0.001)	(0.001)	(0.0002)	(0.001)	(0.0004)	(0.001)	(0.001)	(0.001)
Population (log)	0.254***	0.009	0.012*	0.050**	0.033**	-0.022	0.110**	-0.042*
	(0.045)	(0.020)	(0.006)	(0.020)	(0.013)	(0.021)	(0.043)	(0.024)
GDP	0.00002***	-0.00001***	-0.00000	-0.00000	0.00000**	0.00001***	-0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Education	0.004	0.002	-0.0002	-0.0001	0.001	0.0002	0.002	0.002
	(0.005)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.005)	(0.003)
Constant	-6.686***	-0.712**	1.194***	-0.798***	-1.478***	-0.207	-2.579***	0.494
	(0.657)	(0.289)	(0.087)	(0.284)	(0.190)	(0.311)	(0.626)	(0.341)
Observations	333	333	333	333	333	333	333	333
R^2	0.335	0.081	0.185	0.288	0.448	0.240	0.192	0.236
Adjusted R^2	0.325	0.067	0.172	0.277	0.439	0.229	0.180	0.225
Residual Std. Error (df = 327)	0.662	0.295	0.089	0.289	0.190	0.314	0.628	0.347
F Statistic (df = 5; 327)	32.933***	5.780***	14.833***	26.419***	53.061***	20.706***	15.549***	20.232***

Note: *p<0.1; **p<0.05; ***p<0.01