LUISS

Corso di laurea in Data Science and Management

Cattedra Data Visualization

Academic impact on climate change and extreme weather events: a thematic-geographical analysis applying graph theory

Sinaimeri Blerina

RELATORE

Martino Alessio

CORRELATORE

753361

CANDIDATO

Anno Accademico 2022/2023

Contents

Abstract

Climate change always poses new challenges to the nations of the globe, in terms of mitigation and adaptation but also in terms of cooperation. The role of academic research is crucial to embarking on a path that is victorious in defending the global population from the threats posed by extreme climate disasters. It is shown how developing countries are not only less mentioned in the literature, but also have different challenges, related to the defence of their agricultural sector, than developed countries, on which most academic research focuses. Through the collection of papers related to climate change and extreme weather events and the exploitation of NLP and Graph Theory techniques, this thesis undertakes the task of analysing the thematic and geographical connections existing between the published papers by processing the texts of their abstracts. The results highlight the disparities between the percentage of papers dedicated to certain geographical areas, and their level of vulnerability to climate change. It also illustrates the different topics being addressed in different countries, identifying patterns and discrepancies at global and local levels.

1 Introduction

In the days when science could not yet explain the phenomena, natural disasters were seen as the sign of some divine punishment, or whim. In short, when the sea was stormy or the earth shook, there was good reason to conclude that, for some matter, Poseidon was angry. From the moment mankind became capable, through experience and science, of explaining natural phenomena, of knowing why, and even being able to predict when, water comes down from the sky, the approach, but not the relationship, that mankind has with the natural world changed. Often buried within the concrete walls of our cities, we fail to imagine a collective effort to take care of what is really our home: our planet. And our planet is home to a myriad of other species that, together with us, have managed to develop and flourish over the past millions of years precisely because of a unique factor, as far as we know, that characterises our planet: climatic equilibrium. This unique factor is now being threatened by the very mankind it helped to develop. Climate change is a faithful companion with which the new generations are growing, aware that they have inherited a sick planet that needs its often delayed or ineffective cures as soon as possible. In a world where economic interest often does not match the mitigation and adaptation policies needed to heal the earth, there is hope that lies in the values pursued by the scientific community, which for years has been shouting about the danger and the need to take action, now. But the diversity that resides in both the natural and social-economic elements of different areas of the world pose challenges stemming from inequalities, often the scourge of the most underprivileged, which the scientific community must be able to balance. The aim of this thesis is to provide as broad overview as possible of how far academic research is able to respond to the demand for adaptation to the effects of climate change in a manner that cuts across economic and political logics, how independent and linked to territorial needs, rather than economic interests, it is.

To do this, cutting-edge tools are used, with an extensive collection of academic papers and their abstracts, processed to extract both summaries of the topics and geographical locations targeted by the research. The research is organised in such a way that no manual intervention on the texts is required, but techniques are used and algorithms implemented to extract information autonomously. A visual representation of the data is favoured in order to have results that are as readable as possible. All this is implemented thanks to the power of Python in data processing and visualisation, but also with the help of tools such as Javascript and Gephi, for more complex visualisations. All the tools used, starting with the platform from which the papers are collected, are open source, and therefore freely usable. The techniques used made it possible to enrich data that was poor in information, except for abstracts, with many features that made it possible to analyse insights that otherwise would not have emerged. A number of challenges were faced in drafting the analysis, most notably the limitation of the available computing power with respect to both the large number of data to be processed and the high computational effort required for text processing. Different techniques were used for each problem of extracting information from the data, such as Natural Language Processing and Graph Theory.

The academic contribution that is attempted in this thesis is related to building a bridge between a quantitative research analysis, a thematic, and thus qualitative, and a geographical analysis. Indeed,

academic research often moves in one of these directions, without taking the whole into consideration. Obviously, however, this type of analysis undermines a detailed examination of the topics covered by this thesis, some aspects of which are then only touched. But the aim is not to go deep, for example, in constructing a perfect algorithm in the extraction of keywords or geographical locations from abstracts, but rather to build a general overview capable, however, of reaching conclusions, taking a picture of the situation of academic research with respect to climate change and extreme climate events. The results reflect what other research has identified, namely that developing countries, which are often the most vulnerable to climate change, are often left behind, even though they need interventions often related to population survival issues.

In the section 2, the current state of the art on literature research is analysed, which considers a comparison of published papers on the topic of climate change and the needs of different nations. In particular, the paper "Climate change: Does international research fulfil global demands and necessities?" by Klingelhöfer et al. [20] is analysed which was the starting point and source of numerous insights for this thesis. A brief excursus is then made on the origins and effects of climate change and extreme weather events, and some of the tools used in the course of the analysis are introduced, such as graph theory, which are later explored in more detail.

In the 3 section, the dataset on which the entire analysis is based is built, downloading the data from the Semantic Scholar platform, chosen because it is one of the open-access platforms with the largest selection of papers. An initial visualisation of the data statistics is then made, both to provide initial considerations and to assess the success of the process, analysing, for example, the categories in which the downloaded papers fall by topic

In the 4 section, one of the most important steps of the research is carried out, i.e. features are extracted on the keywords relating to the abstracts and the geographical locations mentioned in the abstracts. The techniques used, their advantages and limitations are explained. Finally, an initial visualisation of the results is made by cross-referencing thematic with geographical results.

In the 5 section, graph theory is exploited to identify communities of papers that share the same topic. Six different communities are detected, for each of which a theme is defined, common to each paper.

Finally, in the 6^{th} section, all the results of the previous sections are gathered, cross-referenced and represented in order to reveal the research evidence. Data from different sources, concerning the vulnerability and readiness of countries to climate change, and macroeconomic indicators are also used, so that the results can be enriched with valuable insights. All methodologies used are explained in detail.

In the 7 section, the room for improvement of the entire analysis is emphasised, and bottlenecks are highlighted. Some possible future developments are then hypothesised for which what has been done and stressed in this thesis can be a starting point.

All code related to data cleaning, manipulation, extraction, processing and visualisation used in this thesis can be found in the following GitHub repository: https://github.com/cosimopoccianti/Master_thesis/tree/main

2 State of the art

Although from a common knowledge perspective, the thematic urgency of climate change has only become more central to public discussion in recent years, tending to be in the new millennium, the climate change debate has much deeper roots. While the international community began some form of organisation on the subject of at least awareness, if not prevention, in 1979, with the First World Climate Conference, of the potential of climate change, and of the fact that massive industrialisation had some side-effects, it began to be talked about much earlier [7]. As written by Evans et al. [7] Jean-Baptiste Fourier, a French scientist, identified the effects of greenhouse gases as early as 1827, and by 1930 there was already a perception in the public imaginary that the planet was warming, with children having milder winters than their fathers. In 1938, scientist GS Callendar demonstrated that it was possible to influence the earth's climate through the emission of CO^2, but found that: "the idea that man's actions could influence so vast a complex is very repugnant to some" [7]. Some even hypothesised that a warmer climate would improve living conditions in colder regions and help crops, but by the mid-1950s, with the increasingly lethal pollution of the big cities, as in the case of the "killer smog" in London in 1953, people began to become aware, at least on a scientific level, of the problem, and came to the conclusion of something that, a few decades earlier, was unimaginable, namely that: "we started worrying less about what nature can do to us, and more about what we have done to nature" as Anthony Giddens argued [7].

The international community, since it began to mobilise, has generated a series of agreements with the main aim, usually, of stabilising or reducing CO^2 emissions, the only real means of contrasting climate change. The first important agreement reached was the Kyoto Protocol. It stems from the objective of the United Nations Framework Convention for Climate Change (UNFCCC) to stabilise CO^2 emissions at a level that would prevent dangerous interference with the climate system. For the implementation of the convention the Conference of Parties (COP) was created and during the third of these conferences the Kyoto Protocol was signed on 11 December 1997 [41]. The Kyoto Protocol stipulated that, on the 1990 baseline, CO^2 emissions should be reduced by approximately, on average, 5%. On average because there were some special clauses for a number of specific countries. The time target was 2008-2012, so the effectiveness of the now completed protocol was debated for a long time. Among the various opinions, Maamoun Nada's study [21] points out that there is a hidden success of the protocol, which lies in the question: what would have happened if it had not been there? Empirical results are varied, but Maamoun's work shows that nations have on average reduced emissions by 6 to 7 percent compared to what would have been expected without the implementation of the protocol. Building confidence in international cooperation in solving the climate problem is not only important, but necessary, as it is the only way to avoid the most disastrous effects.

When thinking about the effects of climate change, there is a fundamental bias inherent in the human brain, accustomed by millennia of evolution to adapting and reacting to its direct environment to defend against immediate dangers. Moreover, the human mind tends to think optimistically about the future [31]. This creates an asynchrony with the rhythms of the climate, where the cause-effect relationship is not so immediate, but characterised by a strong, compared to the average life expectancy

of a human being, lag between the emission of greenhouse gases into the atmosphere and the effects on the climate. Even if, hypothetically, we were able to reduce emissions to zero today, the effects of what we have emitted so far would last for hundreds of centuries, with an increase of 0.1 degrees Celsius on average per decade, based on emissions in 2000 [31]. The fact that climate change is happening today is something that has already been happening for a long time, certainly before much of the human population started to worry. The idea that this is a problem often relegated to the future resides only in the most dramatic aspects of change, but already we are witnessing ever higher average temperatures, rising sea levels, changes in precipitation patterns and increasingly frequent extreme weather events. The impacts are wide-ranging, including on human health, and require urgent mitigation and adaptation action [31]. However, if evolution has not prepared man to be responsive to and afraid of long-term changes, it has led him to build an organized society capable, through scientific research, of casting its gaze beyond the limits and necessities of survival. There is thus the ability to analyze the phenomenon and predict its dangers and intensities. What is perhaps present to a lesser extent is the political capacity to put in place effective preventive actions. But regarding climate change, the academic output is vast, addressing the topic from many different disciplines, from environmental to psychological. However, the impact of change is uneven across the earth's territory, the question that arises is whether scientific research is able to overcome even these spatial differences in order to be effective. Forecasts of country inequality rarely take into account the impacts of climate change, which are of two kinds: on the one hand, the effects of changes are suffered more by the countries least ready to cope them, usually developing countries; on the other hand, the costs of mitigation slow down their own development process [40]. Developing countries, however, have less funds to invest in scientific research on climate and its effects, creating a vicious cycle. Just as vicious is the circle within the countries themselves in social inequality. Indeed, as shown by Islam et al [16], within the same country, the poorest groups of the population suffer more from the negative effects of climate change, just think of the costs related to the reduction in food security, reduced access to drinking water or the negative effects on health.

2.1 Does international research fulfill global demands and necessities?

Regarding the ability of academic research to reflect the real needs arising from climate change both socially and geographically, various studies have been done. Interesting for the purposes of this thesis is the paper by Klingelhöfer et al.: "Climate change: Does international research fulfill global demands and necessities?" [20]. The authors start from the assumption of the diverse impact that different countries suffer from the effects of climate change, highlighting the enormous exposure of developing countries, taking into consideration not only natural, but also socio-economic factors. They then consider as a starting point the increase in natural disasters such as floods, melting glaciers, droughts and heat waves, which lead to dramatic consequences for the food and water supply chain as well as human health, and their inextricable connection to human responsibility for the release of CO^2 into the atmosphere. From these initial considerations they collect from The Core Collection Indices of Web of Science all those articles containing the words "climat* change", "global warming", and

"greenhouse effect". In total, they collect 40,062 papers of which 38,917 can be used for geographical analysis. Their study is based not only on associating the number of papers and their citations with the relative countries in which the papers are written, thus analysing the numbers of academic production per nation, but also on cross-referencing these results with several macroeconomic, environmental, and specific indicators related to climate change and academic research. More precisely, total population, gross domestic product, number of researchers in FTE[1], expenditures on research and development and gross expenditures for R&D[2] are taken into account as general indicators, but relating specifically to climate, carbon dioxide emissions, The Global Climate Risk Index, sea-level rise, Readiness and vulnerability index are considered. Of these, The Global Climate Risk Index [42] takes into account the extent to which a country has suffered extreme weather events, with also a forecast of future development, while the Readiness and vulnerability index [28], as the authors say [20], covers "countries' vulnerability to climate disruption and their readiness to improve resilience by 'leveraging of private and public sector investments'". This last index in particular is important both in the analysis of the results of this paper, but also because it will be incorporated into the analysis of the results of this thesis. The results of the paper see a series of maps like those in figure 1, where the distribution of papers by country is shown, with the evolution of academic productivity for the highest performing countries. With together a series of tables, graphs and evidence linking academic production with the above indices, as in the case of the table below (Table 1).

The findings of the paper show USA, the UK, China, Australia, and Germany to be the most active countries in publishing papers related to climate change. However, the U.S. as a nation is not the best example of climate virtuousness, having also pulled out of the Paris Agreement[3]. In contrast, Europe, taken as a whole, shows both a robust research infrastructure on the topic, but also a conrect attention to mitigation and adaptation policies. Of note is the Chinese economic effort in R&D, which in fact is lately closing the gap with other nations in terms of publications. Regarding socioeconomic indicators, and the ratio of emissions to published papers, the Scandinavian countries are leading, as can be seen in Table 1, having long since begun a massive investment program in renewable energy. The paper also shows a positive correlation between readiness, i.e., countries' preparedness and prevention of the effects of climate change, and the number of papers published in the countries, highlighting how, arguably, investment in research leads the country to be ready for adaptation policies. While, on the other hand, it identifies a negative correlation between the number of publications and a nation's vunerability, i.e., the extent of the impact of the effects of change, indicating how research is done more in relatively less vulnerable countries, mainly in the northern hemisphere of the globe. Developing countries in particular show, on average, high vulnerability without adequate investment in research. These observations lead (quote) to the following conclusions: given the disparity in effect size and ability to do research, careful consideration should be given to better cooperations among different nations to share greater research results, indeed, in the authors' words, "all scientists have only one planet to take care of." Cooperation efforts especially should be directed from developed

[1]FTE: Full-time equivalents

[2]R&D: Research and Development

[3]The Paris Agreement call for a 30% reduction in 2030 emission levels compared with the 1990 baseline [46]

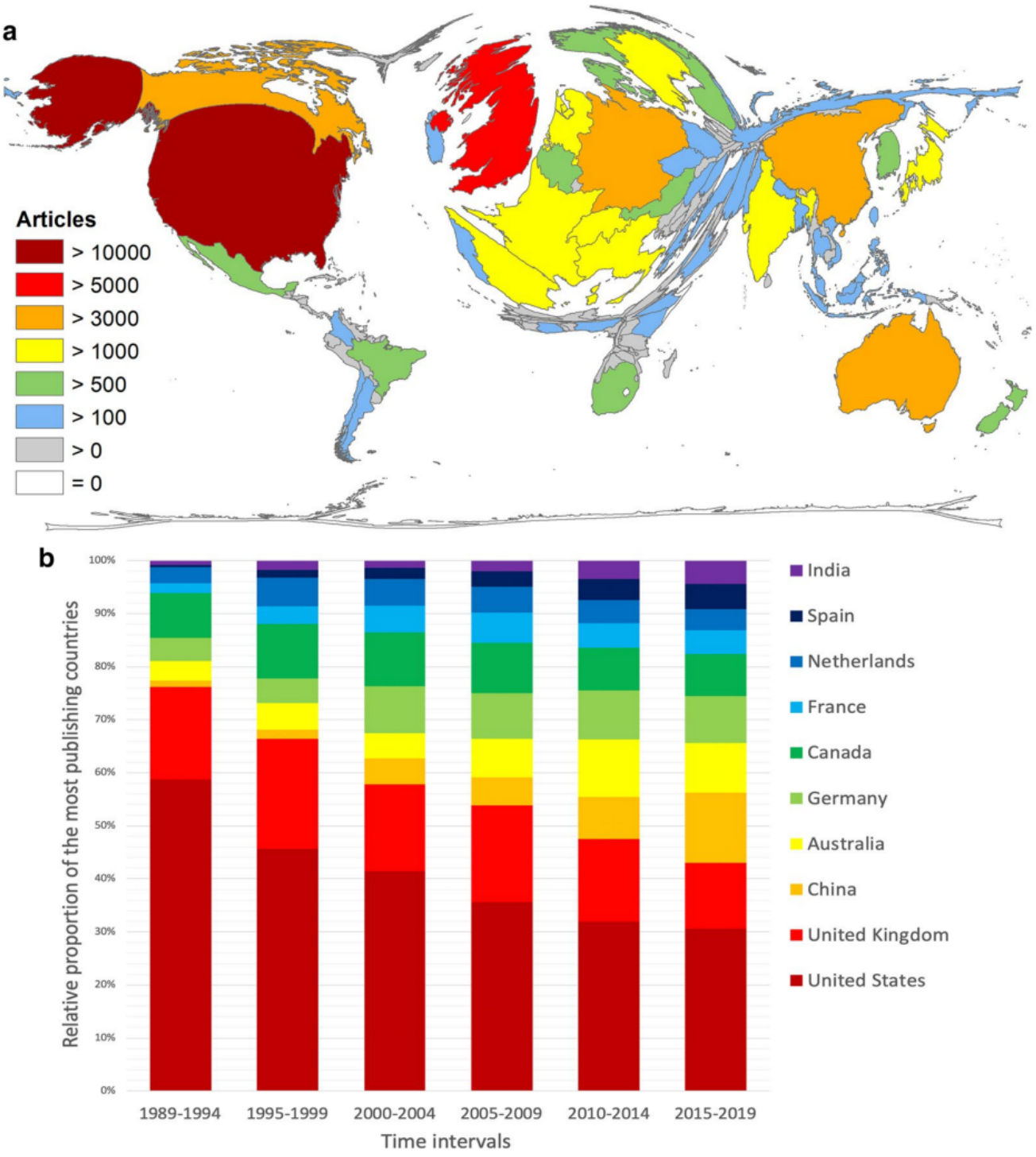


Figure 1: The most publishing countries. **a.** Density equalizing map projection of the number of articles. **b.** Relative share of the most publishing countries in 5-year intervals from 1998 to 2019. Source: Klingelhöfer et al. [20]

countries to developing ones, which will both suffer more from the effects of climate change in an absolute meaning, but also have a greater deficit in the tools needed to tackle them.

The paper comprehensively examines country publications and effectively identifies, by cross-referencing the data, the areas for improvement that the international community should implement to bring some level of equity to developing countries, which have contributed the least to the causes, and suffer the

Table 1: Number of articles on climate change related to the countries' CO^2 emission in billion tons (threshold 300 articles), countries are ranked by R_{CO^2}=number of articles/CO^2 emission in billion tons [45]

Country	Articles	CO^2 in bn	R_{CO^2}
Sweden	1215	41.50	29.28
Switzerland	1126	40.07	28.10
Denmark	795	34.55	23.01
Norway	917	44.79	20.47
New Zealand	523	36.01	14.52
United Kingdom	5524	384.71	14.36
Finland	637	45.96	13.86
Netherlands	1595	164.05	9.72
Austria	668	69.94	9.55
Australia	3349	413.09	8.11
Portugal	429	54.86	7.82
Canada	3126	572.78	5.46
Spain	1458	281.42	5.18
Belgium	503	100.12	5.02
France	1724	356.30	4.84
Greece	316	76.00	4.16
Germany	3238	799.37	4.05
Italy	1300	355.45	3.66
United States	12,637	5269.53	2.40
South Africa	733	456.33	1.61
Brazil	688	476.07	1.45
Mexico	501	490.29	1.02
South Korea	621	616.10	1.01
Japan	1106	1205.06	0.92
Turkey	305	447.90	0.68
India	1300	2466.77	0.53
Iran	339	672.31	0.50
China	3508	9838.75	0.36
Russia	488	1692.79	0.29

most from the effects. There is also a thematic analysis, with the division of papers into different areas, identifying the most recurring keywords in the papers, however, what could be implemented based on the research is to perform a cross-analysis not only with the various indicators used, but also with the topics covered across countries in the papers regarding climate change. In the writing of this thesis, an attempt will be made to extend the research in this direction as well, expanding the qualitative analysis to be able to identify not only geographic, but also thematic patterns in academic research. However, extending the analysis thematically and qualitatively is a challenging task. Luckily, academic production comes to help, with the example of various applications.

2.2 Bibliometric analysis and Graph Theory

Quantitative researchers usually tend to avoid qualitative research considering its results insufficiently robust due to their dependence on the subjectivity and ability of the researcher to identify and analyse the results, relying on his or her ability to also consider possible alternative interpretations [43]. With the development of AI, however, a series of new possibilities and horizons have been created, thanks to the techniques of Natural Processing Languages (NLP), which aim to tackle one of the greatest challenges in computer science, that of being able to generalise irregularities and diversity of human language, in order to be able to build models on it [43]. On the other hand, graph theory has emerged, which often makes it possible to solve, or just understand, complex relationship mechanisms between qualitative data classes. Graphs theory is an ancient science, which began in 1736 with Euler's solution to the Königsberg, nowadays Kaliningrad, bridge problem [6]. A graph is a set of nodes, or vertices, and edges. Each edge connects a pair of vertices, indicating a relationship between them. This relationship could be anything - a social media connection, a road connecting two cities, or even a hyperlink between two web pages. Graphs can be undirected, where relationships are bidirectional, or directed, where they're one-way. They can also be weighted, with each edge carrying a value representing the strength or cost of the relationship. Now imagine applying this concept to NLP. Suddenly words become nodes, sentences become paths, and a piece of text becomes a complex, interconnected network. This network can then be analysed, manipulated and even visualised, providing valuable insights into the structure and semantics of the text. Relationships between nodes, however, are not necessarily represented by words; rather, they are often witness to other kinds of connections. For example, Singhania et al. [35], base their Scientometric analysis on sustainability reports by constructing networks where the edges can be, for example, collaborations between authors. In their work, they analyze 1 434 papers published between 1992 and 2022, creating some visualizations and studying graphs to explain the relationships among sustainability reports. the conclusions of their work are not relevant for the purposes of this thesis, but some of the methods they used may be useful in building the framework on which the analysis in this research will be based. Scientometric is a narrowing of the term bibliometric, more referring to science and technology. A bibliometric analysis is about collecting papers and ranking them based on citations, authors, geographic areas, and other possible connections between research articles. While in the past the purpose was more to give an overview of the literature in a particular field, perhaps by analyzing the most cited papers, as technology has advanced it has paved the way for a whole new set of possible applications, from collecting keywords, to identifying patterns that the human mind, with its limited information storage capacity, can not bring to surface [5]. An example of the use of bibliographic techniques is described in the paper summarised in the section above 2.1. There are thousands of types of graphs, they can be tree, clickque, direct or indirect. In the case study of this thesis, a subset of graphs will be used, namely networks. A network is nothing more than a graph in which, however, each node belongs to a component. In fact, graphs may be unconnected, i.e. a node, or a group of nodes connected by edges, is not connected to other nodes in the graph. In a network this is not possible. That is, there will always be a path, i.e. a set of nodes and edges, through which two nodes in the same

network can be connected. The power of networks, however, lies in a set of algorithms capable of finding, through the structure of connections between nodes, communities, that is, clusters of nodes that are particularly interconnected with each other and particularly disconnected from the rest of the network. It means many of the edges of a node join nodes in the same cluster, few of them nodes outside the cluster. Graphs are objects where order often coexists with disorder [8], but in a graph representing a real situation, recurring patterns, or some form of order and hierarchies are often found, when compared to one where the generation of edges between nodes is random, which would lead to no possibility of finding cluster logic. The identification of communities has many practical revolves that can lead to the identification or efficiency of systems, bringing evidence that would not have emerged from a different analysis of the data. The identification of communities allows the subdivision of nodes into clusters according to their structural position within the graph [8], and according to their connections with other nodes, which represent real relationships with other elements of the case study. Communities can also be composed of smaller communities within them, with a hierarchical structure. However, the problem of how to find communities is computationally relevant, given the large amount of data that often makes up a network. Furthermore, there is no universally accepted definition of a community within a network, but various studies have led to the formation of different algorithms, which will however somehow yield a different result. The choice of which algorithm to use from among the many developed is therefore not a trivial task. Defining what is or is not a good breakdown of a graph in a community is an open problem, but the concept of Modularity as identified by Newman and Girvan [11] has quickly become the most widely used and well-known quality function [8]. Modularity is a measure of the goodness, of the strength of the communities found within a network, and various algorithms have been developed with the aim of maximising modularity, which is then used as a criterion for stopping algorithms, which are not always able to find the optimum. One of the first algorithms built to maximise modularity is the greedy method proposed by Newman [25]. An agglomerative hierarchical clustering [8] that starts with as many communities as nodes, with therefore modularity equal to zero, and gradually merges the nodes, trying to maximise modularity. Both the concept of modularity and the greedy modularity algorithm will be further explored in the empirical community detection section of this thesis 5.1.

2.3 Shifts in climate change statistics

The theme of climate change can be approached mainly from two perspectives: either the causes can be investigated, or the effects. While much has been done from the point of view of causes, which are attributed with certainty to the actions of mankind and its responsibility for atmospheric emissions of gases harmful to the climatic balance, there are still some uncertainties as far as the effects are concerned, since they have been residing for a long time in a hypothetical future. The causes indeed now belong to the past, and there is an effort by the international community to mitigate emissions. While the effects, although they are already in place, reside mainly in the future, and it is therefore necessary to rely on predictive models and theoretical studies to assess their impact. Such a violent surge in the amount of greenhouse gases in the atmosphere has never occurred throughout

the history of our planet, so there are no examples in the climatic past of what is about to happen. There have certainly been many changes in climate patterns, but rarely have they been as fast as they are now, and so they can be investigated to predict what will happen. But the prediction of patterns is made even more difficult by the infinite geographical facets that the climatic balance has across the earth's planet [22]. There is thus a relative chaos in the predictive ability of future climate events, where some findings contradict each other and where some expectations are not matched in certain geographical areas [22]. There are certain effects, which emerge globally and which, despite more or less pronounced territorial variability, are reproduced in common trends and which tend to lead the statistics of climate metrics to accentuate the tails associated with extreme events. In particular, the changes in average temperatures that are already, and will be, recorded in the planet's climate stations are quite common: there is and is predicted to be a further shift in temperature distribution, towards milder temperatures, but also an increase in its variability [32]. Although it appears that the variability will be greater in the summer than in the winter seasons, specifically in the northern hemisphere, and the extreme temperatures recorded will be more at the high end of the thermometer than at the low end. Furthermore, when considering temperature variation on a daily level, temperatures will rise more at night, reducing the day/night range [22]. Indeed, there will likely be a reduction in the number of days of winter frost, and a lengthening of the growing season, especially in the colder areas of the planet, along with an increase in heat waves that [32], together with changes in rainfall patterns, will lead to a dryening of inland areas of the continents [22]. From a precipitation point of view, along with temperatures it is the area that has the most common trends across the globe. In fact, there will be a general decrease in rainfall, but an increase in extreme events [32], as well as a decrease in the number of storms, but an increase in the violence of those that occur. This is tied in with the increase in temperature, which increases the availability of vapour ready to be discharged into the atmosphere [22]. As for droughts and extreme winds, they are more dependent on regional climatic characteristics, so it is more difficult for models to predict with an acceptable degree of certainty what will occur [32]. The change will therefore be twofold: on the one hand there is a shift in the averages of the statistics, such as a reduction in precipitation or a general increase in temperatures, but on the other hand there is a sharpening of extreme events, which are only partly affected by the direction of the shift in the averages. That is to say, a decrease in average precipitation does not presume a reduction in floods or monsoons, and an increase in temperatures does not lead to a reduction in the frequency of extreme frost peaks. It will therefore be a more extreme climate that will present challenges, both from the point of view of preparing territories to best adapt to the arrival of extreme events, and in the necessary rapid recovery after they have occurred, with effects that cut across a myriad of different sectors, as already seen, ranging from human health, to the food supply chain, ecosystems, and so on. In addition, the disparity of the wounds that these extreme events leave in the different countries on which they strike will increase. Indeed, as pointed out by [20], the distribution of vulnerability is far greater in developing countries, which, however, find themselves in the situation of chasing after extreme climate events, without playing in advance, and investing more in the recovery of territories after disasters have occurred rather than creating adaptive capacities that can help mitigate the most disastrous effects of climate events. And investments are often made in

debt, thus creating vicious circles that lead developing countries to find themselves in increasingly disadvantaged situations [23]. In light of what has been written so far, it will be the aim of this thesis to start not with the causes, but with the effects that climate change already has and will continue to have, in an ever increasing form, on the globe. An attempt will then be made to take into account the shift towards extreme events in order to assess the extent of the suffering of each nation. The analysis will therefore be restricted to the branch of those who suffer, in all respects, the disastrous effects of change.

3 Data Collection and Starting Dataset Definition

The objective of this section is to collect sufficient data to be able to implement an analysis based on the factors emerging from the above-mentioned litterature review. In particular, multiple studies are highlighted from a qualitative point of view on the causes and effects of climate change, and bibliometric studies on the ability of international research to meet the demand arising from the needs of individual nations. What is therefore desired, in order to contribute to the research, is to include in a quantitative analysis a qualitative part that can be derived from capturing the themes covered in the published papers, in order to create a bridge with the geographical areas, understanding both the needs and assessing whether the effort of the academic community is capable of aligning with territorial necessities. The negative aspects of climate change, which result from the most extreme manifestations of climate change, are to be taken into account in order to have a broader view of the effects rather than the causes. For these reasons, the starting point for the gathering of data from the papers will be the abstracts, in order to have material on which to implement algorithms capable of extracting the most accurate thematic information possible concerning the research objects of the papers.

3.1 Target and Tools

The final aim of this section is to create a dataset comprehensive of information on papers, their authors and the relative abstracts. The theme that links all the papers is obviously that of climate change, with a particular emphasis on the impact it generates in terms of extreme weather conditions. To do this, several possible solutions that would provide an API for large-scale research in the vast sea of published papers were analysed.

3.1.1 Semantic Scholar API

In the initial phase, several possible APIs were explored that would allow connection to databases containing scholarly works (Figure 2). Of these, the choice fell on Semantic Scholar [34], mainly for these reasons:

- It is open and free: it is only needed to apply for an API key, which is granted for research purposes
- It has a large database containing around 213 million papers from 79 million different authors
- It allows the entire database to be downloaded
- It contains abstracts of papers
- Contains a lot of metadata, e.g. the academic field the paper falls into

Looking at the table in figure 2 , Semantic Scholar appears to be the best choice for the case under consideration in this thesis, and furthermore, the way in which the dataset download is structured

Resource	URL	Article Count	Access	Services
Aminer	aminer.org	321.5M	open	D*
arXiv	arxiv.org	2M	open	D**,F,S
BASE	base-search.net	180.5M	open	S
CORE	core.ac.uk	207.3M	open	D*, S
Dimensions	app.dimensions.ai	123.8M	subscription	D, F, M, S
Google Scholar	scholar.google.com	?	-	-
The Lens	lens.org	240.4M	subscription	D, M, S
Meta	-	-	terminated 3/31/22	-
Microsoft Academic	-	-	terminated 12/31/21	-
OpenAlex	openalex.org	205.2M	open	D, F, M
PubMed Central	ncbi.nlm.nih.gov/pmc/	7.5M	open	D**,F,P,S
ResearchGate	researchgate.net	135.0M	-	-
Scopus	scopus.com	84.0M	subscription	F, M, S
Semantic Scholar	**semanticscholar.org**	**205M**	**open**	**D, F, M, P, S, T**
Web of Science Core	webofknowledge.com	83.2M	subscription	F, M, S

Key: D=data download; F=field-of-study classification; M=advanced metadata;
P=semantically parsed text; S=title and abstract search; T=natural language summarization
*=data more than a year stale; **=restricted fields of study
Article count does not include patents or datasets.

Figure 2: Representation of possible alternative sources for papers collection. Source: [19]

ensures that it can be manipulated even if the memory and capacity of the laptop used is limited. Semantic Scholar allows the download of 30 or 31, depending on the reference database, different files representing the entire block, thus also allowing them to be processed consequently, saving memory. The other alternatives with open access are still valid but do not share the features of Semantic Scholar, either in terms of the amount and variety of data that can be obtained per paper or in terms of the size of the database.

Semantic Scholar was released in 2015 as a solution to gain access to research and thus help it grow and progress. As the authors write in "The Semantic Scholar Open Data Platform", it has become a key tool especially because: "The need for timely and comprehensive scholarly data has become more imperative since the 2021 sunsetting of the Microsoft Academic Graph (MAG) [...], which was long a standard source for scholarly data in the community" [19]. The Semantic Scholar API is the way to access the Semantic Scholar Academic Graph, which is a "disambiguated, high-quality, bibliographic knowledge graph" [19]. It is a sophisticated graph in which the nodes are papers, authors, venue and academic institutions and the edges represent what can connect these nodes, i.e. papers written by the same author, cited by another paper, published in the same venue or connected to the same academic institution. So for instance, for each author, there will be only one node that will be connected through the edges to all the papers related to that author, which will in turn be connected to all the other authors involved in their writing.

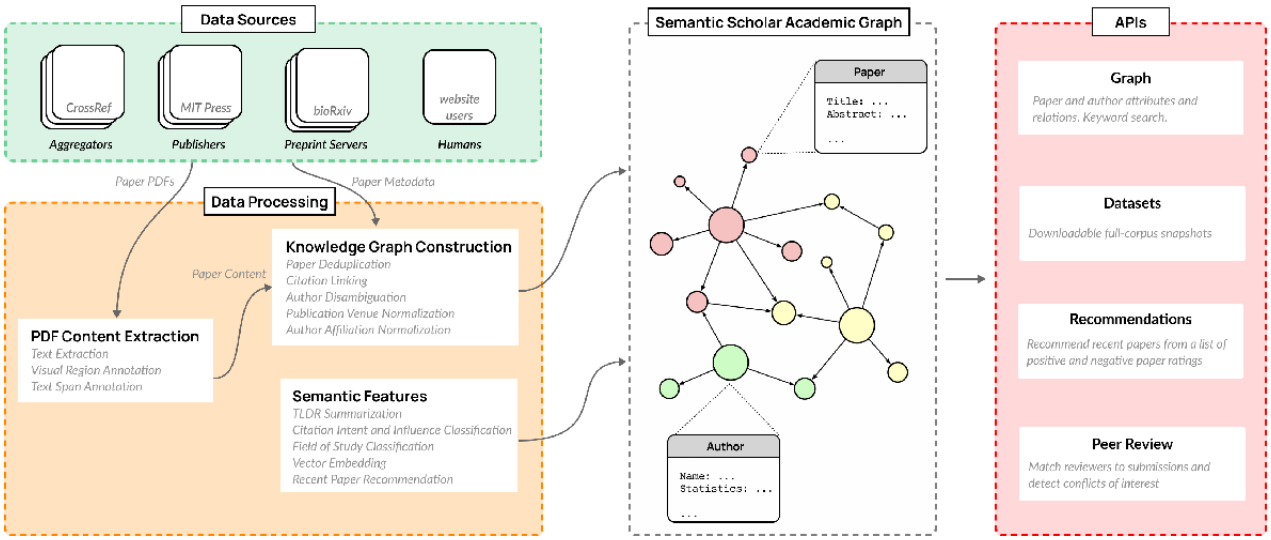


Figure 3: Illustration of the Semantic Scholar platform. Source: [19]

The graph structure builds on the use of a variety of different inputs, more than 50, which are combined and harmonised within the pipeline. There are both direct sources, such as non-profit platforms, from which the information on the papers is taken directly, as well as technologies for extracting the information directly from the pdfs, which ultimately are the official source of the papers, and there is also an involvement of the platform's user community. Human effort is therefore not underestimated in the process, e.g. an author may report the intellectual property of a paper, or a normal user may report an error, which is promptly fixed by a team member. Such a complex, and above all varied structure, and subject to sophisticated technologies, as the extraction of information from PDFs, is prone to possible errors, but in the case of this thesis, given the large volume of papers collected, the error is negligible.

3.1.2 Datasets API

The Semantic Scholar platform gives the possibility to use two different APIs: Academic Graph API and Datasets API.

- Academic Graph API: offers the possibility to search papers by keywords, using the same engine that manages searches directly on the Semantic Scholar website, and to download information related to them, including author details and abstracts. However, it places a limitation on the number of papers that can be downloaded: 10 MB and 1000 papers at a time. Furthermore, for each search, the sum of papers and offsets[4] must be less than 10000. So, it is possible, by making several requests in sequence, to download the first 9999 resulting papers per search key.

- Datasets API: It allows periodic downloads of snapshots from the entire graph present up to date on Semantic Scholar, e.g. for this research the information present on the platform as a

[4]The offset is the position where the papers start to be downloaded from. For example, if the offset is 10, and the limit 15, papers will be downloaded from the tenth result in search order, up to the twenty-fifth. Thus, by combining the limit and offset values, the first 9999 papers of the search can be downloaded.

result of the snapshots taken on the 11th of July 2023. Three different types of datasets can be downloaded: papers, authors, and abstracts.

For the scope of this thesis, the use of Datasets API was preferred. Indeed, not only does it allow for a broader search, and access to all the papers corresponding to a given search key, but it also gives more control over the search itself. Although the Semantic Scholar search engine certainly works well, it is complex, and figuring out according to which parameters and combinations it finds results that match the search would require separate work. By actually downloading the entire dataset, there is then more freedom in choosing how to organise the search for the desired papers, which phrases to include and how to combine the words, and also in which fields of the papers search for these words, for instance by excluding the authors' names. Three different types of datasets can be downloaded with the Dataset API:

1. Papers dataset: contains all the metadata related to the papers, such as title, year of publication, authors, citations, etc.

2. Abstracts dataset: contains the abstracts related to each paper

3. Authors dataset: contains information about the authors, such as number of published papers, aliases for the name, affiliation institute, etc.

The three datasets have two keys to bridge them: the papers are linked to the abstracts through the corpus id, while they are connected to the authors through the author id. With these keys, it is possible to create a unique dataset with the relevant information for each paper, including abstracts.

3.2 Datasets Download and Processing

Downloading the entire dataset involves retrieving approximately 30 files for each type of dataset among papers, authors and abstracts. Each file arrives zipped, with a size of roughly 1.5 GB, for a total of about 45 GB each dataset. In total, therefore, 135 GB are downloaded through the API. Below is an example of the query used to download the papers, as far as authors and abstracts are concerned, the differences are few.

```
import requests
import wget
from tqdm import tqdm
import os

url = 'https://api.semanticscholar.org/datasets/v1/release
    ↪ /2023-07-11/dataset/papers'

header = {'x-api-key': 'Semanti Scholar API Key'}
```

```
response = requests.get(url, headers=header, verify=False)
response = response.json()

for file in tqdm(response['files']):
    present = False
    for j in os.listdir('.'):
        if j in file:
            present = True
    if present == False:
        wget.download(file)
```

The for loop is used to check whether the file has already been downloaded, as the loop was sometimes interrupted due to connection failures. Therefore, since it takes about two hours to download the entire dataset, this avoids re-downloading files that have already been downloaded. Each 1.5 GB file, becomes 5 when unziped. Unzipping all the files together would mean, posing to divide the process for each of the three datasets, at least 150 GB free, and handling files of this size on a laptop requires caution. Also, considering that there are 3 datasets, there would need to be 450 GB of free space. Therefore, it was decided to proceed with a more sustainable process given the limited computational capabilities. Indeed, the code was programmed to extract and process one file at a time. Moreover, having as the ultimate goal to search for all those papers with a common theme "Climate change and extreme weather conditions," there is also the need to combine the dataset of papers, particularly by taking their titles, with that of abstracts, in order to create a search as refined as possible. This poses a further challenge in managing the files, because in order to combine the titles to the papers, with the aim of being able to enter both into the function that searches for keywords, there is a need to keep in memory all the titles extracted from the papers dataset.
The first step is therefore to process the paper dataset files one at a time, extracting the titles and saving them in a structured dictionary with a key equal to the paper's corpus id, and the title as the value. The code runs by unzipping a file, saving all the paper titles linked to their corpus id, and then deleting the unzipped file to save space, moving on to process the next file. The following code repeats the logic of processing the files, this time of the abstracts dataset, one at a time, and gradually eliminating the unpacked version:

```
# Path to the folder containing the input gzip files
folder_path = "Zipped_abstracts"

# Path to the output folder for the unzipped files and the final
    ↪ result of the research
output_folder = "Unzipped_abstracts"
output_abstract = "abstracts_results2"
```

```
count = 0
# Iterate over the files in the folder
for filename in tqdm(os.listdir(folder_path)):

    count += 1
    abstract = {}
    #extracting the files
    if filename.endswith(".gz"):
        input_file = os.path.join(folder_path, filename)
        output_file = os.path.join(output_folder, filename[:-3])  #
            ↪ Remove the .gz extension

        with gzip.open(input_file, 'rb') as gz_file:
            with open(output_file, 'wb') as out_file:
                shutil.copyfileobj(gz_file, out_file)

        #reading the extracted files
        df = open(f'{output_folder}/{filename[:-3]}', encoding='utf
            ↪ -8')
        file = df.readlines()
        dict_abs = {}
        #Storing all the abstract of the file in a dictionary, whit
            ↪ as key the corpus id of the paper of each abstract
        for i in (range(len(file))):
            j = json.loads(file[i])['corpusid']
            dict_abs[j] = json.loads(file[i])

        #searching for keywords: checking if the combination of
            ↪ abstract + paper title contain the given combination
            ↪ of keywords
        for i in (dict_abs):
            string_p = dict_abs[i]["abstract"].lower() + '␣' + dict
                ↪ [i].lower()
            string = re.sub(r'(?:[^\w\s]|_)+', '␣', string_p).strip
                ↪ ()
            if ("climate" in string or "environmental" in string or
                ↪  "atmospheric" in string) and ("change" in string
                ↪  or "changes" in string or "shift" in string or "
                ↪ shifts" in string or "alteration" in string or "
```

```
            ↪ alterations" in string or "transformation" in
            ↪ string or "transformations" in string) and ("
            ↪ extreme" in string or "severe" in string or "
            ↪ harsh" in string or "unusual" in string) and ("
            ↪ weather" in string or "climate" in string or "
            ↪ environmental" in string or "temperature" in
            ↪ string or "temperatures" in string or "
            ↪ atmospheric" in string or "meteorological" in
            ↪ string) and ("conditions" in string or "condition
            ↪ " in string or "event" in string or "events" in
            ↪ string or "pattern" in string or "patterns" in
            ↪ string or "phenomena" in string or "episodes" in
            ↪ string):
              abstract[i] = dict_abs[i]
    df.close()


    #writing the result of the resarch for each zipped file
    js = json.dumps(abstract)
    fp = open(f'{output_abstract}/abstracts_{count}.json', 'a')
    fp.write(js)
    fp.close()

    #remove the unzipped file to save store space
    os.remove(f'{output_folder}/{filename[:-3]}')
```

The file containing the abstracts is unzipped, the abstracts are extracted from the elements of the file and associated, in a dictionary, with the corpus id of the paper to which they belong. Then, for each abstract in the dictionary, its string is concatenated with that of the title that has the same corpus id as its key, and the concatenated string is searched for words related to climate change and extreme climate events, with a logic that will be explored in more detail later. Concatenations of abstracts and titles that do not fit the search keys are discarded, while those that do match are saved in a dictionary. In the end, there will thus be 30 dictionaries, corresponding to the 30 files in the dataset of abstracts, with the search results. After shortlisting, the resulting papers are 72 355.

3.2.1 Research Keys

The key point in this phase of the research is to carefully select the papers to be retained and to fit into what will become the source dataset of the analysis. It is therefore crucial to choose precisely the logic and words used to define the boundary beyond which a paper is discarded. As already stated, the aim is to collect those papers that have "climate change and extreme weather conditions" as their topic.

It is obvious, however, that by seeking an exact match with these words, many of the papers that are related in topic but which, either in nuance or meaning, differ, even slightly, from these exact words would be discarded, unfairly. Furthermore, the intention was to keep as much control as possible at this stage, setting up the search with a simple IF followed by the words and logical operators, rather than using some external library, probably more efficient. Going into the details of the code structure, the first step is to clean the abstract and title strings of all punctuation and special characters, setting the lowercase and adjusting the string so that each word is divided by a simple space. The notions of punctuation and sentence organisation are then lost, but are retained in the original abstract, which will then eventually be saved in the dictionary. For greater clarity on the choice of search words, please refer to the diagrams (Figure 4 & 5). The code consists of a series of AND and OR conditions which, if satisfied, lead to the inclusion of the paper in the final dataset. To understand how it works, it is useful to imagine the code as a filter consisting of 4 levels. In order to pass the filter, the combination of title and abstract must contain at least one of the words on each level:

- Layer 1: weather, climate, environmental, atmospheric, meteorological
- Layer 2: change, changes, shift, shifts, alteration, alterations, transformation, transformations
- Layer 3: extreme, severe, harsh, unusual
- Layer 4: condition, conditions, event, events, pattern, patterns, phenomena, episodes

The choice of words is geared towards the effects of climate change, precisely because the purpose of this work is to understand what patterns exist between climate change research and the actual effects in various locations around the world. Therefore, all those words more simply related to environmental issues, such as sustainability, which are often found in papers resulting from this research, are left out. Two examples are given below to show how the papers are filtered by the algorithm: words that match those found in the various layers are underlined, and in the diagrams below (figure 4 & 5), corresponding one to the first and one to the second example, is it possibible to see how these words are at least one per layer. Each diagram can be read in the following way: if, by highlighting the underlined words in the papers, it is possible to move from top to bottom on the diagram while remaining on a green path, then the paper passes the filter logic and is included among those selected for analysis.

Example 1: "A Hyper-Integrated Mobility as a Service (MaaS) to Gamification and Carbon Market Enterprise Architecture Framework for Sustainable Environment". [30]

Various human activities emit greenhouse gasses (GHGs) that contribute to global climate change. These include the burning of fossil fuels for energy production, transportation, and industrial uses, and the clearing of forests to create farmland and pasture, all for urban and industrial development. As a result, temperatures around the world are rising, extreme weather events are occurring more frequently, and human health is suffering because of these changes. As a result of massive traffic,

agriculture, and urbanization, the natural environment is being destroyed, negatively affecting humans and other living things. Humanity plans to live in smart cities within this ecosystem as the world evolves around these mutations. A smart city uses technology and data to improve the quality of life of its citizens and the efficiency of its urban systems. Smart cities have the potential to be more sustainable because they use technology and data to improve the efficiency of urban systems and reduce the negative impact of human activities on the environment. Smart cities can also use technology to improve green transportation and waste management and reduce water consumption, which can help conserve natural resources and protect the environment. Smart cities can create livable, efficient, and sustainable urban environments using technology and data. This paper presents a new Enterprise Architecture Framework for reducing carbon emissions for environmental sustainability that combines gamification and green behavior with blockchain architecture to ensure a system that is trustworthy, secure, and scalable for shareholders, citizens, service providers, and technology vendors. In order to achieve this, the hyper-integrated framework approach explains a roadmap for how sustainability for reducing carbon emissions from transportation is based on an optimized MaaS approach improved by gamification. As part of this study, a computational model and a formulation are proposed to calculate the activity exchange values in the MaaS ecosystem for swapping, changing, and bartering for assets within the integrated system. This paper aims to propose the framework and a module interoperability approach, so numerical values for computation parameters are not included as they may belong to other research studies. In spite of this, a case study section has been provided as an example of a calculation approach.

Example 2: "The Effects Atmospheric Changes Have on Runoff". [44]

Over the last century, the earth has seen unprecedented atmospheric concentrations contaminate our ecosystems due to human activity. Predictions state the introduction of carbon dioxide, methane, nitrous oxide, and chlorofluorocarbons (CFCs), will increase temperatures, and change the amount and location of precipitation causing more runoff. This could potentially result in disturbance events such as floods, to be more frequent and severe. This study aims to perform an assessment of the effects of a range of hypothetical climate changes on runoff in the North-east Pond River watershed, located in Newfoundland. To carry this out a watershed runoff model simulates runoff in the basin for current climatic conditions and for hypothetical climatic conditions that represent a range of possible climate changes (Bobba et al., 1997). The hypothetical changes in climate will showcase the effects of a 2oC increase in temperature on the total annual precipitation. This will then be compared to flood forecasting models to analyze how runoff will be affected by various climatic conditions, inducing unusual flooding events (Wijayarathne & Coulibaly, 2020). Previous studies have indicated the runoff sensitivity in watersheds to changes in temperatures which raises concerns as to the adverse effects this may cause in limiting water resources in the semi-arid regions in parts of Canada and the U.S. Thus, there is a need to increase the understanding of the sensitivity of water resources in Canadian watersheds to climate variability and climate change as effects of this magnitude on the North-east Pond River could have significant environmental implications.

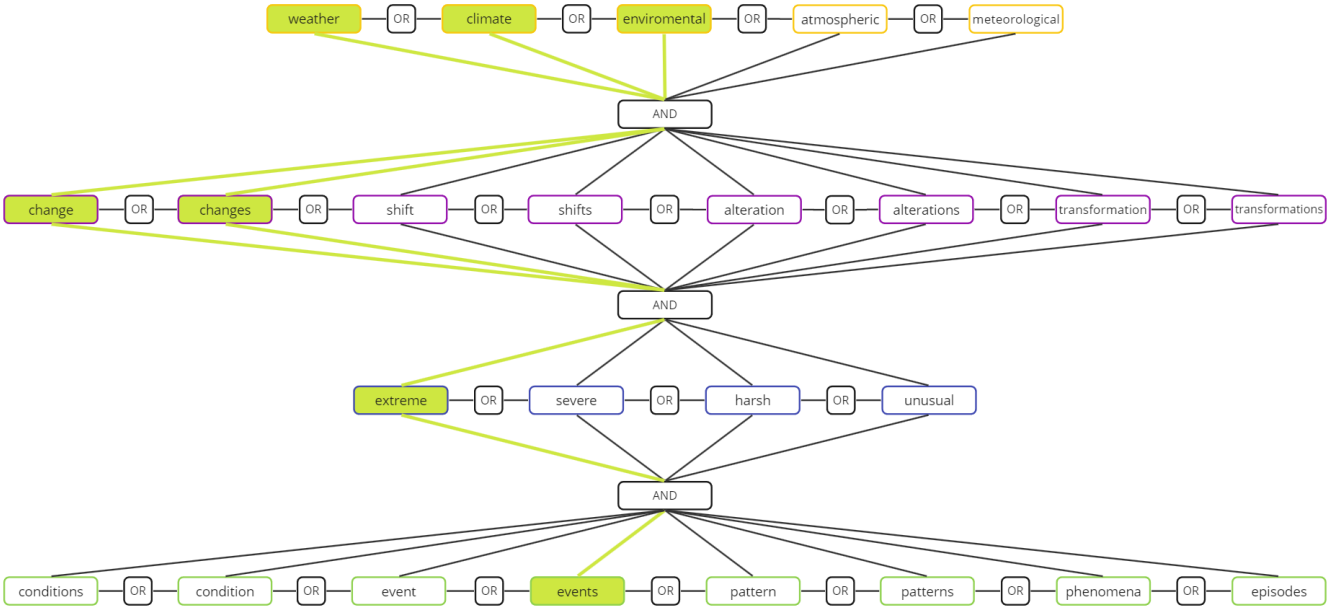


Figure 4: Diagram of which words made the paper "A Hyper-Integrated Mobility as a Service (MaaS) to Gamification and Carbon Market Enterprise Architecture Framework for Sustainable Environment" [30] pass the search filter.

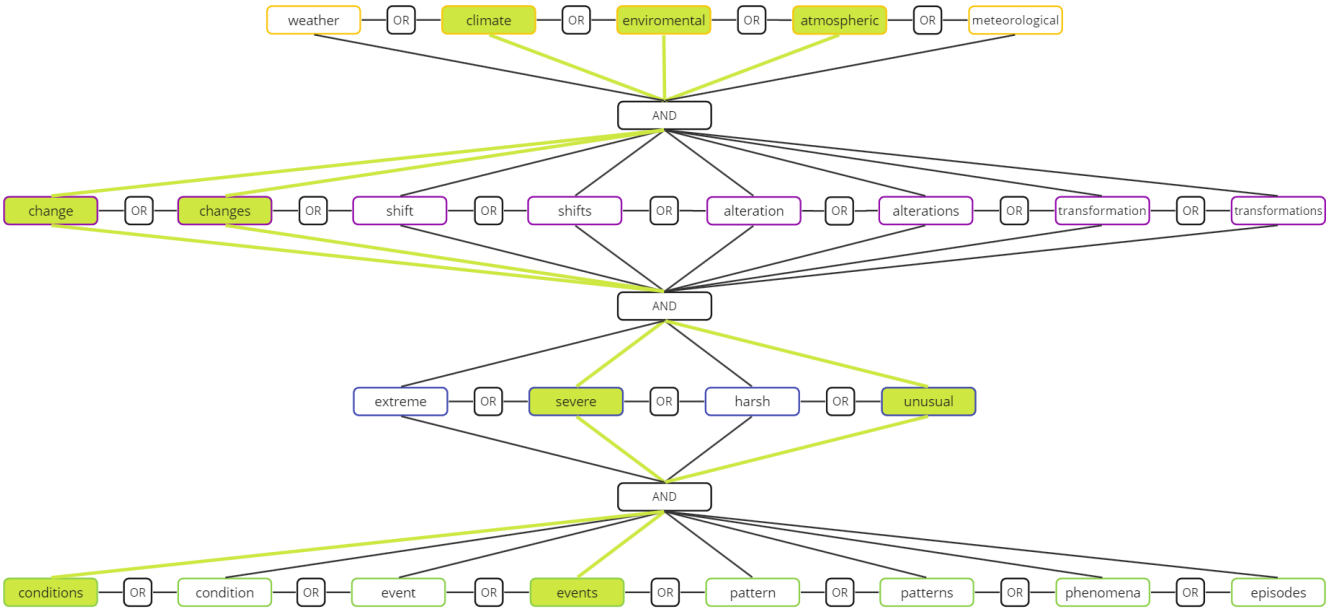


Figure 5: Diagram of which words made the paper "The Effects Atmospheric Changes Have on Runoff" [44] pass the search filter.

3.2.2 Merging results with information on papers and authors

The next step is to merge the thirty dictionaries created with the search, and associate each abstract with metadata regarding the referring paper and the authors of that paper. To do this, the 30 json files containing the dictionaries are first merged into a single file which then contains the 72,355 papers of the search. Then, starting from the key of each dictionary element, i.e. the corpus id, we search whether that particular corpus id is present in one of the thirty files representing the metadata of

the papers. The process is then always the same, extracting the information from one file at a time, checking which of the search result keys are present in the extracted file, saving the elements that match and then deleting the extracted file to save memory. Once the abstracts are attached to the paper metadata, it is time to collect information on the authors, for whom a further 30 datasets were downloaded. Among the paper metadata, there are both the names and the author id, thanks to which it is possible to have a bridge to find the information in the author datasets. So, as before, we proceed to extract one file at a time, and search in it for the author ids that emerged with the search results. The final result of the entire search, and thus the resulting dataset, the starting point for the analysis, can better be visualised in the image below (Figure 6).

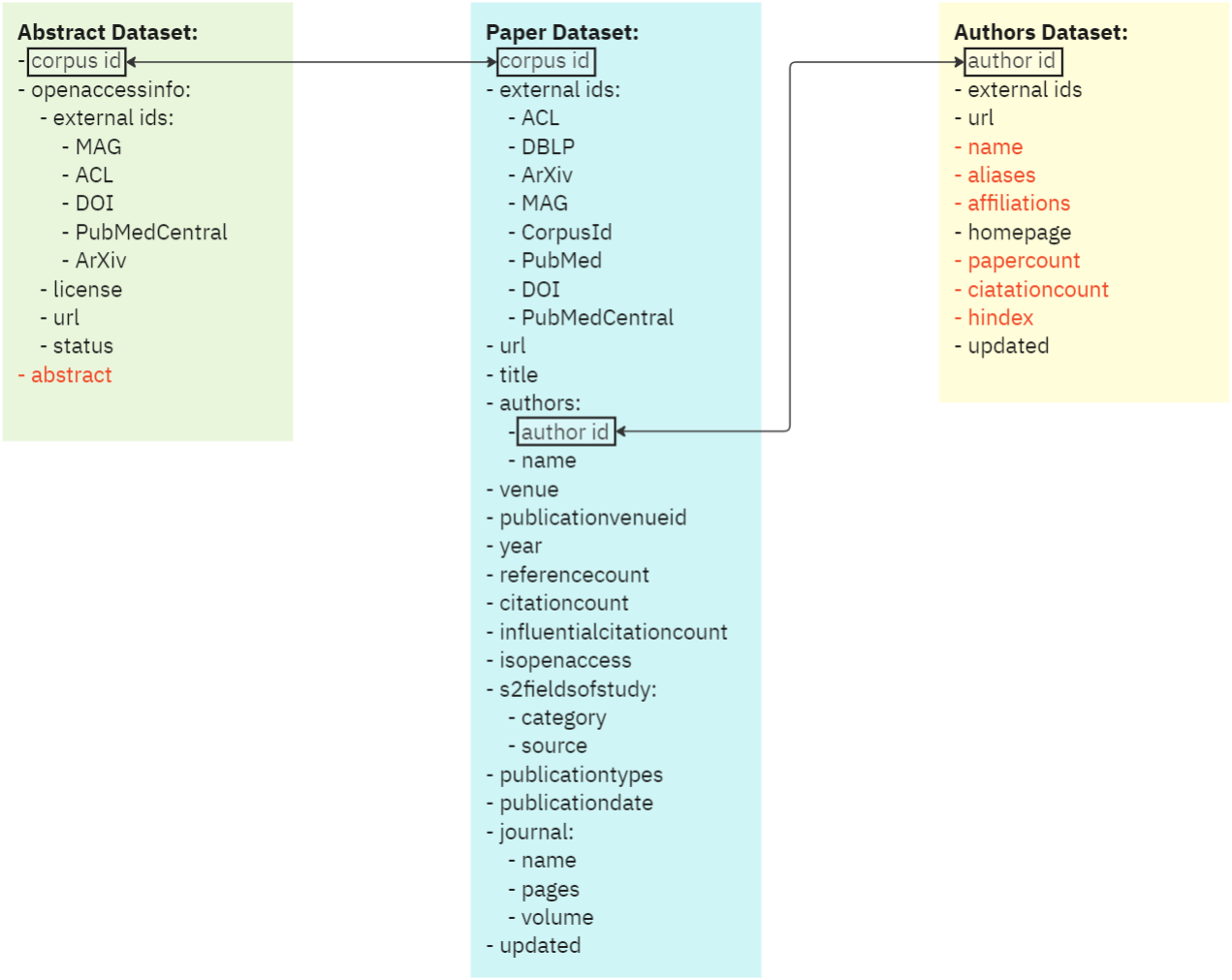


Figure 6: Representation of the final columns in the dataset. In red are represented fields taken from the abstracts dataset and the authors dataset respectively, on the join keys corpus id and author id.

3.3 Data Exploration

With the implemented method, 72 355 papers were collected. A careful analysis of the results is necessary, together with a check on the fairness and consistency of the data, in order to be able to read the results in the light of this starting point once the analysis has been completed. The dataset

contains the following columns: "corpusid", "externalids", "url", "title", "authors", "venue", "publicationvenueid", "year", "referencecount", "citationcount", "influentialcitationcount", "isopenaccess", "s2fieldsofstudy", "publicationtypes", "publicationdate", "journal", "updated", "abstract". Of these, some contain data in pure form, i.e. one entry per paper, while others are represented by dictionaries. For example, the author column contains, for each row, a dictionary with information taken from the authors dataset for each author who has contributed to the paper. The number of different authors, i.e. the number of unique author ids, is 196 053, with an average of 2.71 authors per paper. There are no missing values for abstracts in the dataset, because that was the search key. However, there are some missing values relevant to the exploration made in this section. First of all, there are 622 missing values in the "year" column and 15 542 in the "s2fieldsofstudy" column, which represents the category, or field of study, in which the paper is classified. Later in the analysis, when necessary, the entries with missing values in these fields will be dropped. The collected papers range from 1812 to 2023. However, it should be noted that some dates are misclassified. For example, the only paper from 1812, which is a book review, actually mentions events from 1890 in the abstract, and molecular biology, which dates back to at least 1938 [38], has a name. The book in question could therefore be from 1812, but it is not clear. In any case, the analysis will certainly focus on more recent periods, and it is assumed that most of the papers are correctly classified. Analysing the distribution of papers with respect to year of publication, it is clear to observe a left-skewed distribution, as in the figure 7. Delving into the detailed count of papers per year, we see how, in 30 years, we go from 179 papers in 1990 to 5 637 in 2020. Moreover, in the period between 2010 and 2020 alone, the number of papers has more than doubled. Furthermore, when considering the number of papers published before 2010, they are 17 116, whereas, when considering from 2010 onwards, they are 54 617, a number more than three times as large. It should be noted that, for 2023, the figure is not applicable, as the dataset's last update is on 11 July 2023. It can be noted, however, that there seems to have been a decrease in the number of papers from 2021 to 2022. This brief, cursory analysis might lead the unwary reader to conclude that the increase in the number of papers published over the years is due to the increase in attention, awareness, and needs related to climate change and its consequences. However, this is not correct, or at least, there is not enough evidence to draw such conclusions at present. Indeed, not only is there evidence that the number of papers in general published per year is steadily increasing, but also that the means of digitally cataloguing papers certainly did not exist in 1950. According to "Wordsrated" [3], the number of papers published between 2018 and 2022 increased by 22.78%, in the case being explored the increase, considering 2018 and 2021, since 2022 shows an anomalous decrease for the reasons explained below, is more pronounced but similar, at 32.77%. But without having to go looking for confirmation of this hypothesis, it is sufficient to look at the data of the downloaded dataset, taking into account not only the number of papers published per year after applying the search filters, but the number of papers published per year on the entire downloaded dataset. Remember how the original dataset has more than 200 million entries, and therefore this comparison can only be made in a limited number of cases, due to the limitation posed by computational power. In this case, fortunately, collecting a dataset with only the years of publication is not prohibitive.

The comparison of these results can be seen in figure 8. In blue is the line representing the total

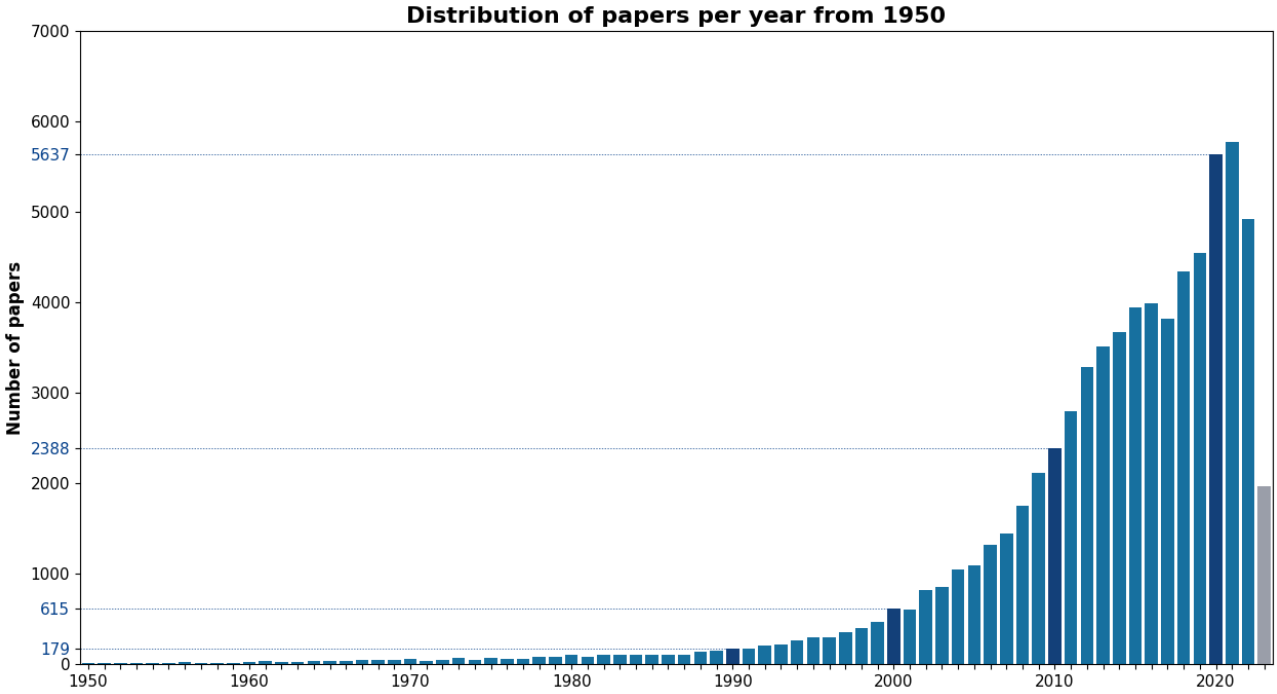


Figure 7: Papers distribution per year from 1950. The 2023 bar is grey because data are up to the 11th of July 2023.

number of papers in the initial dataset per year, in red the line for the dataset after shortlisting of papers. Comparing two such different magnitudes of numbers is not meant to be misleading, but it can help to understand the phenomenon of the dramatic increase in the number of papers published on environmental topics in recent years. From the figure it can be immediately noted, confirming what was stated above, that the number of papers on the Semantic Scholar platform has been increasing at a good rate since 1950, and that it has undergone a vertiginous surge after the 2000s. As far as the "climate change etc." themed papers are concerned, it is immediately apparent that the increase does not follow the general number of papers in perfect photocopy, but is practically flat until the 2000s. If we look at the area between the two lines, representing the difference in growth between the two cases, we can see how it starts to increase significantly around the 1950s, reaches its maximum amplitude around 2000, and then declines significantly. This means that, in terms of percentage growth, papers on the subject of climate change were initially an insignificant part of the total papers, but that, from the 2000s onwards, the number of papers published grew at a faster rate than that of papers in general. Again, one can see an anomalous behaviour of the lines, compared to what one might expect, after 2020. However, for a more accurate analysis it is useful to rescale the two distributions by taking the year-to-year percentage changes cumulatively. It makes sense to do this from a not too far back year. Indeed, environmental-themed papers published in the years around 1900 are few in number, around the order of units, so we would have a percentage increase of 5000/6000% in 2022. It is therefore good to start from the time when the curve of filtered papers starts to diverge slightly from the 0-line, around the 1970s. The results of the cumulative percentage changes can be seen in figure 9. It is already clear at first glance that the growth of environmentally themed papers has been much

more pronounced than that of papers in the entire Semantic Scholar database, although it also meets expectations. Analysing the two percentage trends in more detail, we see that from 1980 to 2020, the growth in the number of total papers published per year has more or less remained constant: from 1980 to 2000, there was a growth of 106.53%, while in the following two decades, it decreased slightly to 102.66%. If we look instead at the numbers of shortlisted papers, they grew by 199.2% between 1980 and 2000 and by 242.62% between 2000 and 2020. It is thus confirmed that the publication of papers related to "climate change and extreme weather conditions" per year grew on average at a double rate. Precisely between 1970 and 2020, the ratio is on average 2.29 : 1. In conclusion, it can be said that the presence of environmentally themed papers on climate change has been increasing over time, gradually accounting for a larger share of the total number of papers published. And that it has done so in an ever-increasing manner, as can be seen from the divergence of the two curves in figure 9, where the distance between them steadily increases. A careful reader might have noticed a discrepancy between the total number of papers resulting from the search of 72,355, shown above, and the 69,653 in figure 8. But two points should be noted: firstly, only the period from 1900 to 2020 is considered in the graph, and secondly, those 622 papers lacking the year information have obviously been excluded.

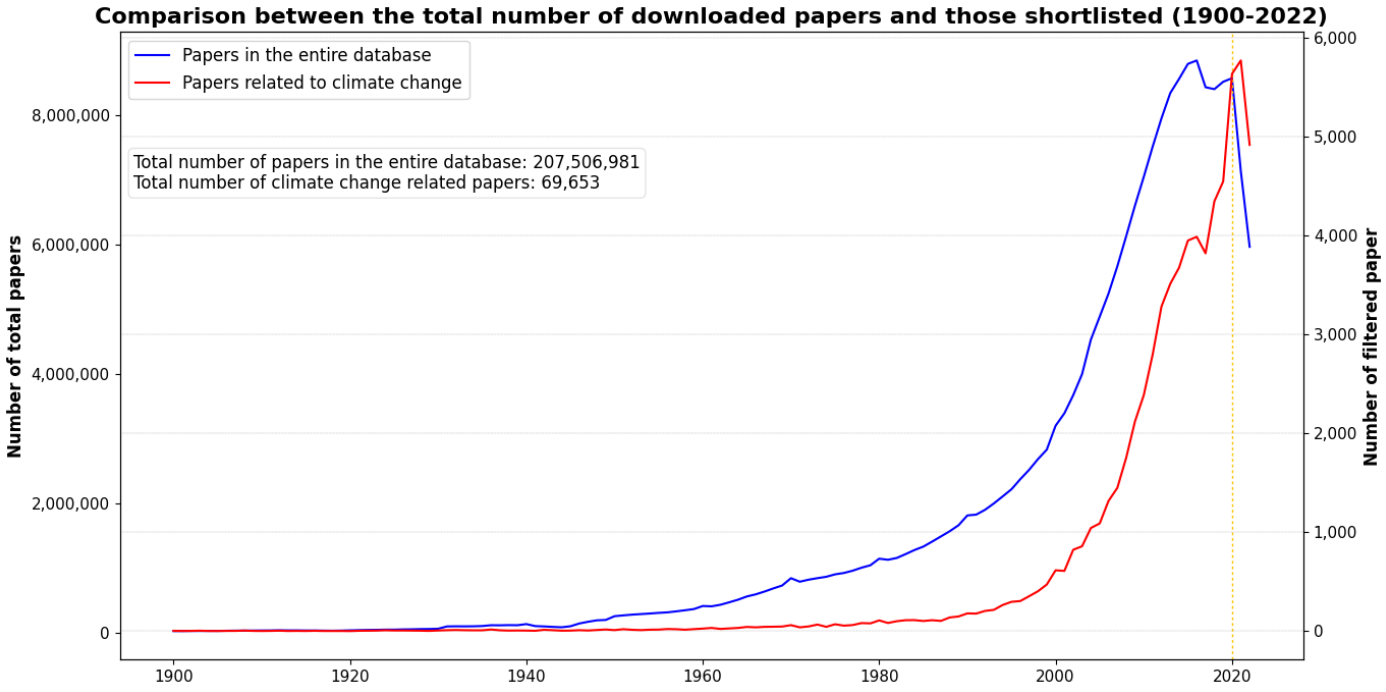


Figure 8: Comparison between the total number of papers in the Semantic Scholar database and those resulting after filtering, by year between 1900 and 2023

3.3.1 Dataset familiarisation: categories analysis

Analysing the result of the shortlisting of the papers is not solely important to familiarise with the starting data for the purpose of this thesis, but also to validate the results of the shortlisting itself, to assess whether the process has taken place correctly and whether indeed the remaining papers are

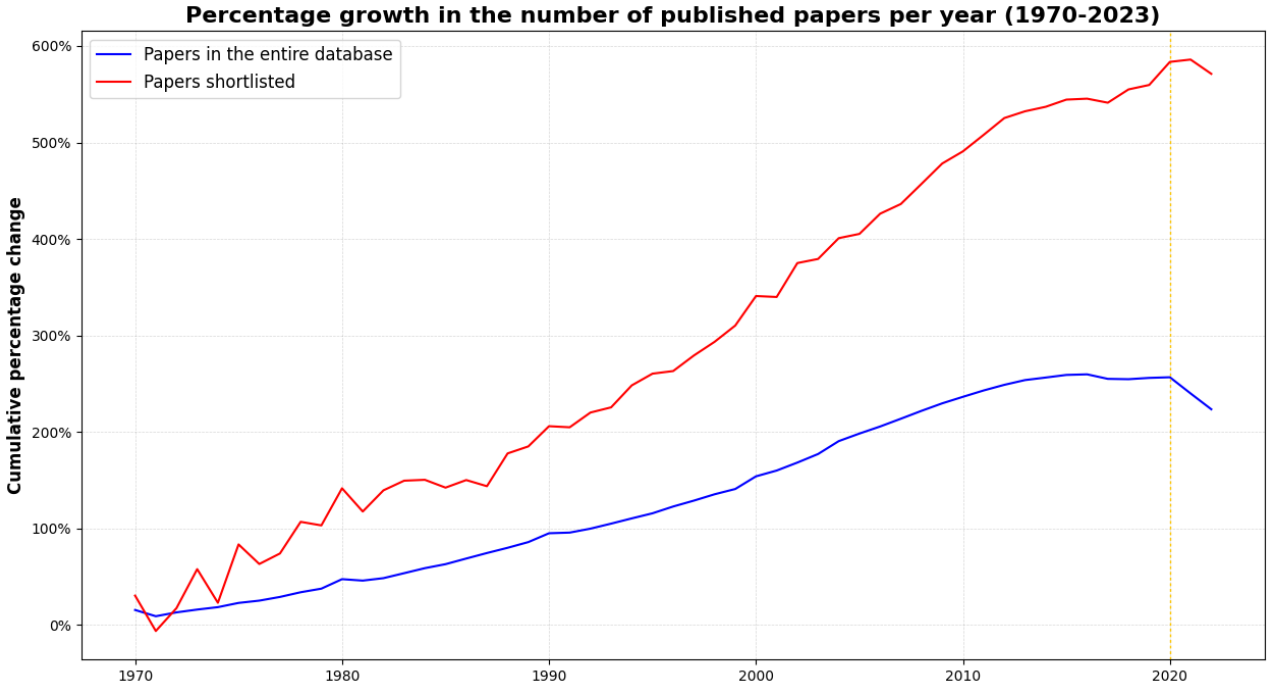


Figure 9: Cumulative year-to-year percentage increase of papers in the Semantic Scholar database and those resulting after filtering, between 1970 and 2023

consistent, as content, with the objective of this research. Obviously, having to deal with more than 70,000 papers, it is impossible to do this check "by sight", as it would take an unnecessary amount of time. Nor would it make sense to rely on a random check for the conclusions. The need, therefore, to find a quick yardstick for evaluating the results makes it necessary to study the contents of the "s2fieldsofstudy" column, which contains, for each paper, the categories, the fields of study, into which the paper is classified. It can then be assessed whether, at the very least, the papers concern fields of study that can be associated with "climate change and extreme climate conditions". If, for example, most of the papers were to be found classified as "Linguistics" or "Art", it is obvious that, however these fields may find their contribution to the cause, there is some problem in filtering the papers. Now, this column is populated for most of the papers, although there are 15 542 missing values, but not uniquely, i.e., each paper may have more than one classification. Each entry in this field has such a structure: for each paper, it can be seen that there is both the classification proposed by the Semantic Scholar platform, but also other classifications, which are taken directly from the sources or from the paper itself. It was decided, for greater data consistency and quality, to take into account only those classifications made by the Semantic Scholar platform algorithm. However, even in this case the classification is not necessarily unique, but there may be papers classified in two or more different categories. From a qualitative point of view this is not a concern given the purpose of respecting the nature of the papers, which may well be multidisciplinary, and keep track of this information. The analysis initially focused on a mere calculation of the number of occurrences for each category, but then the type of relationships existing between the different categories was also explored for a deeper understanding of the nature of the papers in the dataset. To give an insight, a total of 65 659 categorisations were found. The number of unique categories is 23. In the graphs in

figure 10, an initial result of the analysis can be seen, the figure is structured as follows:

1. In the first graph, the number of occurrences for each category can be viewed: it is clear that the leading category is "Environmental Science" with 34 068 papers falling into it, followed with a large gap by the group containing "Agricultural and Food Science", "Biology" and "Medicine", a slightly more detached "Engineering" and then the rest.

2. In the second graph, the aim was to analyse, within the subset of papers falling under the classification "Environmental Science", which other categorisations the papers fell into. It can be seen that the most frequent association of categories is in this case, that of "Environmental Science" and "Geography", in which 2357 papers fall.

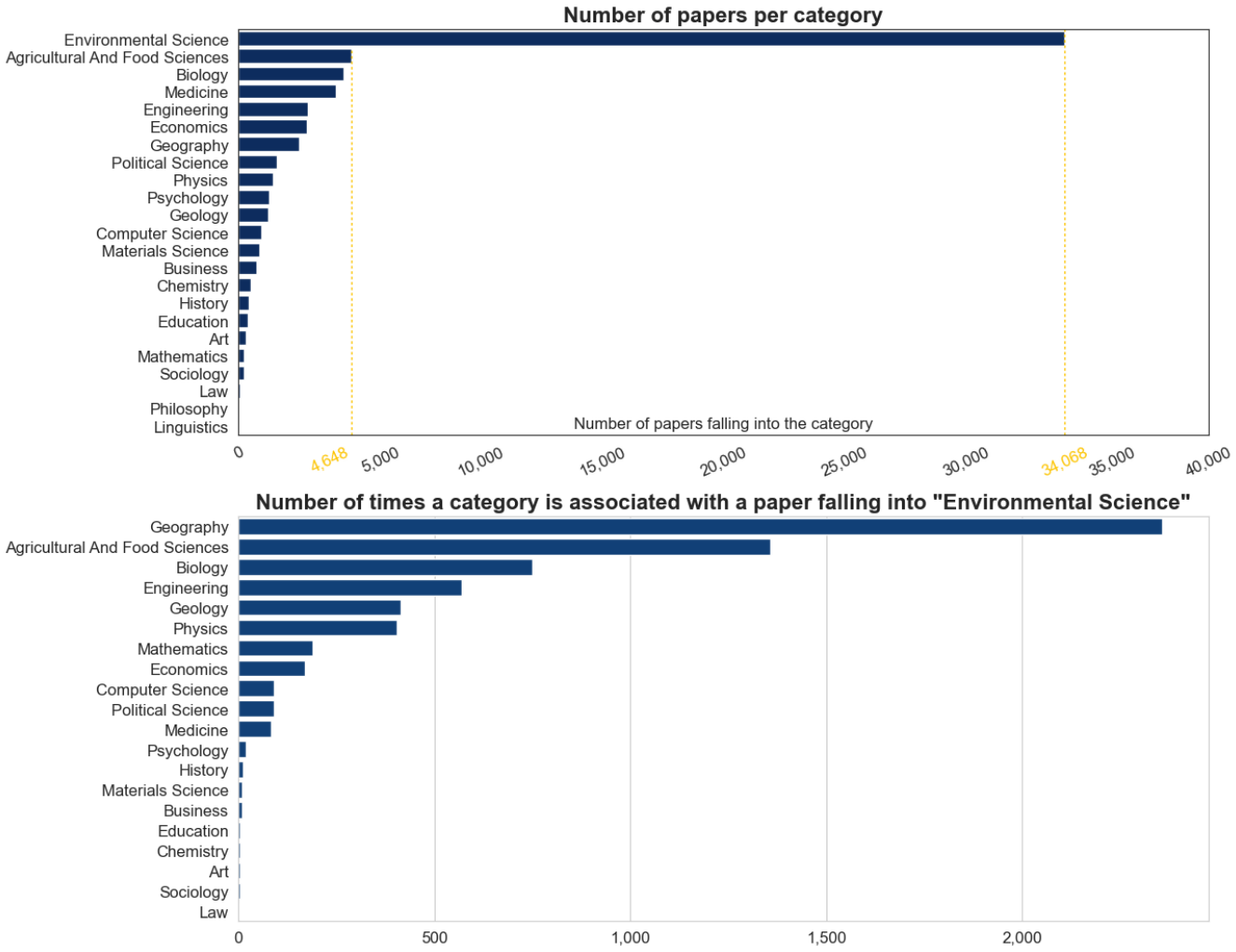


Figure 10: Representation of the occurrence of the categories in which the papers are classified.

From this first and certainly cursory analysis, it can be said that the objective of filtering the papers, at least to the extent of their topics, was achieved. For a more comprehensive exploration of the fields of study concerning the papers, a relational matrix was created for all 23 categories in the dataset. As a process, a dataset was created with equal indexes and columns representing the categories, and populated such that, at the intersection of each pair of categories, the number of times those categories

were associated in the classification of a paper is given. Then the matrix was plotted, the result of which is in figure 11.

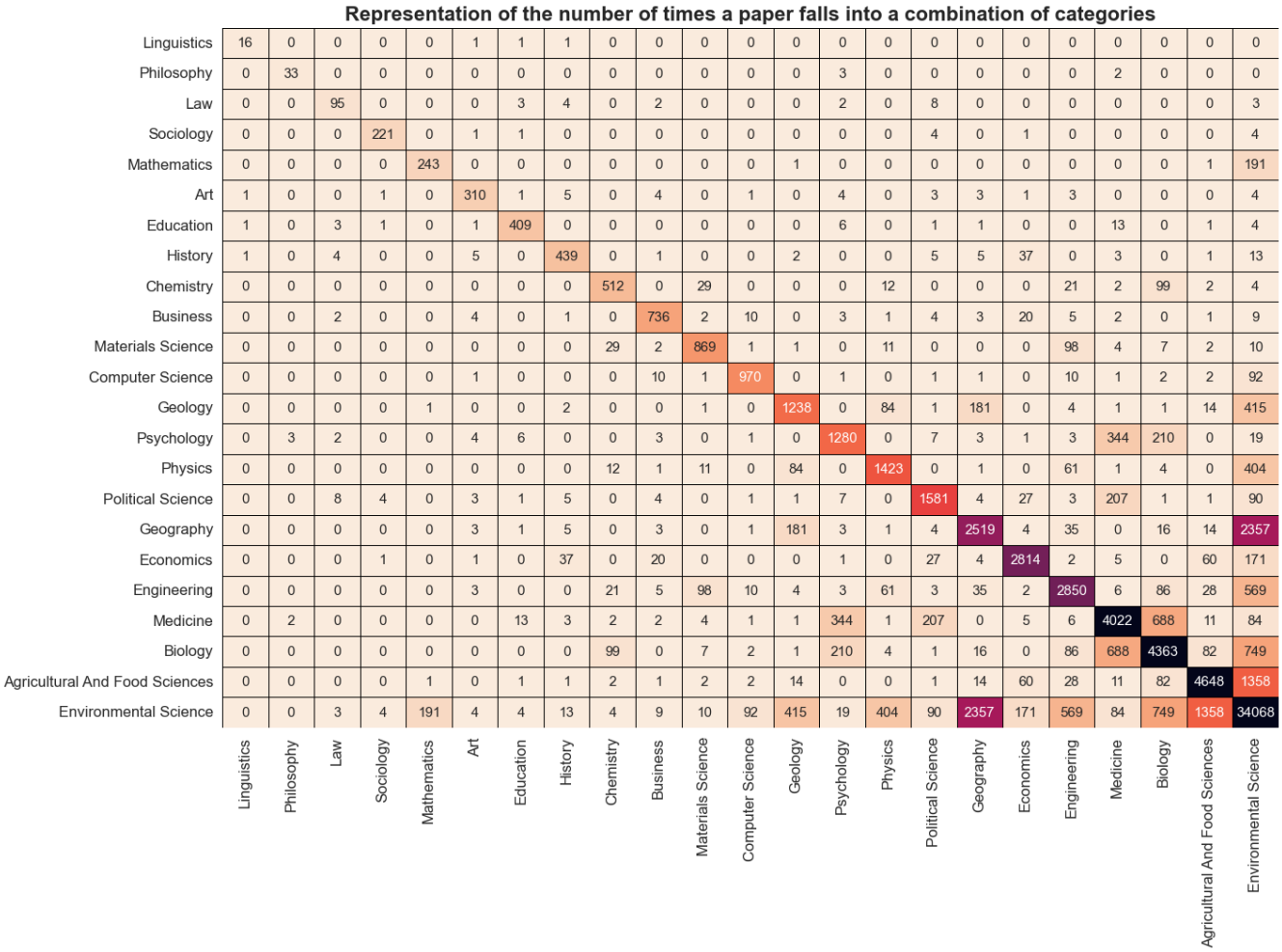


Figure 11: Matrix of relations between categories, showing the number of times a paper is classified with a given pair of fields of study.

Even analysing this plot, it can be seen that most papers fall in the bottom right-hand corner of the matrix, where categories that fall within the scope of this thesis can be found. Indeed, several categories related to the environmental sciences can be seen to interrelate, with the presence of the category "Engineering", and "Agricultural and Food Sciences", both of which are inextricably linked to aspects of climate change, due to the impacts they have on research to find innovative solution and food production. The purpose of this matrix is to confirm once again that, at least with regard to the topic, the initial objective of collecting papers related to "climate change and extreme weather conditions" has been achieved.

The last step in this category analysis, but perhaps the one that can offer the most insights, is to take the four categories most present in the dataset, and observe their distribution over the years. The main objective is to assess whether, for the papers in our dataset, certain categories had a particularly significant presence in a specific time period. Again, the primary task was to reorganise the data in such a way as to be able to associate the year of occurrence with each occurrence of each category

entry. The results can be seen in figure 12.

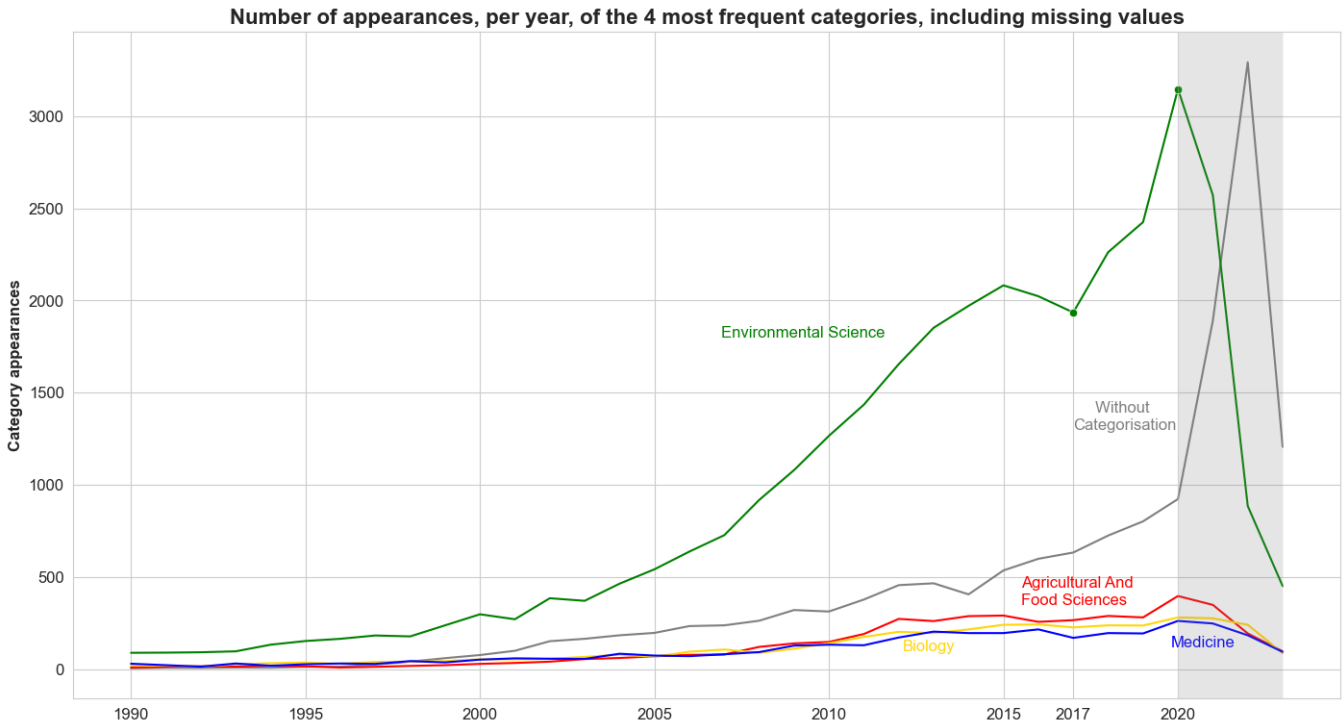


Figure 12: Line plot representing the number of occurrences per category in the years 1990 to 2023.

In the graph, it can be seen that the growth in the number of papers published per year is largely related to the growth in the number of "Environmental Sciences" papers published. Furthermore, comparing the trend of the "Environmental Science" category to the others, it can be noted that between 2017 and 2020 there has been a very noticeable increase in papers so classified, while the slope of the curves of the other categories remains relatively stable. It can be concluded that this category has always been more present in the classification of the papers covered by this research, but that the spread between its presence and that of the other categories has steadily increased over time, reflecting an increasing focus on the topic.

The unusual catch, however, that immediately stands out, partly because it is evidenced by the gray area in the graphs, is that, as of 2020, the drastic reduction in the total number of entries for the category does not reflect an equal decline present in the number of papers per year, which, on the contrary, continues to increase. After careful analysis, it appeared that, and is represented in the graph, after 2020 a large proportion of papers do not possess a categorization. Hence the choice to include, in both line plots, also the fictitious category "Missing Values," which can be seen to skyrocket enormously after 2020, surpassing "Environmental Science" between 2021 and 2022, and becoming the most represented. This, together with other events highlighted above in the analysis, leads us to conclude that after 2020 the data are not fully reliable, probably due to a delay in the categorization of the papaers of the database management systems, which have to process a large amount of data. Certainly the dataset is reliable on the content of abstracts, but sometimes there is a need to use related metadata as well, it is more prudent not to consider 2021 and 2022, the latter year for which, anyway, the number of papers decrease. It is not the intention of this thesis to analyse

why the semantic scholar database after 2020 has dirtier data and, it could be argued, even less data in general, but the purpose of this section is to set a baseline, to deepen the knowledge of the data in such a way that a solid, and well known, foundation can be laid for the analysis and models that will be built later.

3.4 Known Dataset Limitations and Biases

It is therefore worth considering and highlighting the limitations of the constructed dataset, so that the results of the analysis can be as knowledgeable as possible.
Firstly, the source is unique, so the papers present depend on the collection methods of Semantic Scholar, which cannot be all-inclusive of all published papers. Secondly, of the papers present on Semantic Scholar, due to the way the search was set up, only papers with an abstract were selected, so all those papers that do not contain an abstract on the platform were excluded.
It cannot therefore be considered an omni-comprehensive representation of reality, but a subset of that population which approximates, given the large number of papers collected, to a good extent the academic production on the topic in question.
Moreover, it is necessary to emphasise that the general theme of this research, and thus the way in which the papers were collected, is not that of climate change, but a subset of it, i.e. applied to extreme climatic events, which justifies the predominant presence of papers categorised as "Environmental Sciences". The decision to take a more specific theme also stems from the decision to make a geographical analysis of the most evident effects of climate change.
Lastly, it is evident that a tool such as Semantic Scholar is much more representative of papers published since the advent of the internet, and less so of those published earlier, dependent primarily on digitisation, and secondly on whether or not they are included on the platform. Furthermore, as this first exploratory analysis has shown, one can see the anomalous behaviour, both in the number of papers present, and in the missing categorisation of many of them, from 2020 onwards, a reason that could depend, especially regarding the defect in number, on the lack of uploading of abstracts. In any case, this aspect will be taken into account in the analysis below.

4 NLP - Feature Extraction

Moving on with the analysis, it is essential to focus on processing the most valuable information that the created dataset contains: the abstracts. To work with the texts, it is necessary to perform NLP[5]. The objectives of this section are twofold, because there are two types of information to be obtained: on the one hand, the abstracts will be processed to obtain keywords that can represent the general theme of the paper, and on the other hand, geographical information will be extracted from the abstracts themselves on the location, when present, that is the object of the paper's research.

4.1 Key-words Extraction

The first step is to associate keywords with the various papers. Alternative methods have been analysed for this section. It is not the purpose of this thesis to find the best keyword extraction method, but rather to attempt to do it quickly, thus avoiding computationally heavy models. For this reason, the idea of training an LLM[6] or using a pre-trained one such as those that can be found on HuggingFace [14] was ruled out, of which the possibility of using "H2-keywordextractor" [13] in particular was explored. The latter, however, while extremely effective in the quality of the results, takes about 30 seconds per abstract to extract the keywords, something less with the use of the hugging face API, which however has limitations. Assuming we had around 70 000 abstracts to process, the total time would have been excessive. This is why the search for alternative solutions proceeded, eventually building a model based on the KeyBERT [12] library for NLP and keywords extraction, plus NLTK [27] and SpaCy [36] for the lemmatisation of the extracted keywords.

4.1.1 KeyBERT

KeyBERT is a library created with the aim of having an easy-to-use pre-trained model, so that it is possible, with merely three lines of code, to extract keywords from a text quickly. It is based on the embeddings extraction model known as BERT [47]. BERT is a model, initially called Transformer, developed, starting in 2017, by researchers at Google [47], seeking a lighter model to train than RNN[7] and CNN[8]and with more parallelisation capabilities, while maintaining, if not increasing, the quality of the output. The functioning of this model can be visualised in figure 13, where it can be seen that the output is reused as model input before text generation, and that there is the addition of a softmax function before output, as explained by Chi Sun et al. in "How to Fine-Tune BERT for Text Classification?" [39]: "BERT-base model contains an encoder with 12 Transformer blocks, 12 self-attention heads, and the hidden size of 768. BERT takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The sequence has one or two segments that the first token of the sequence is always [CLS] which contains the special classification embedding and another special token [SEP] is used for separating segments. For text classification tasks, BERT

[5]NLP: Natural Language Processing
[6]LLM: Large Language Model
[7]RNN: Recurrent Neural Networks
[8]CNN: Convolutional Neural Network

takes the final hidden state h of the first token [CLS] as the representation of the whole sequence. A simple softmax classifier is added to the top of BERT to predict the probability of label c: $p(c|\mathbf{h}) = \text{softmax}(W\mathbf{h})$, where W is the task-specific parameter matrix."

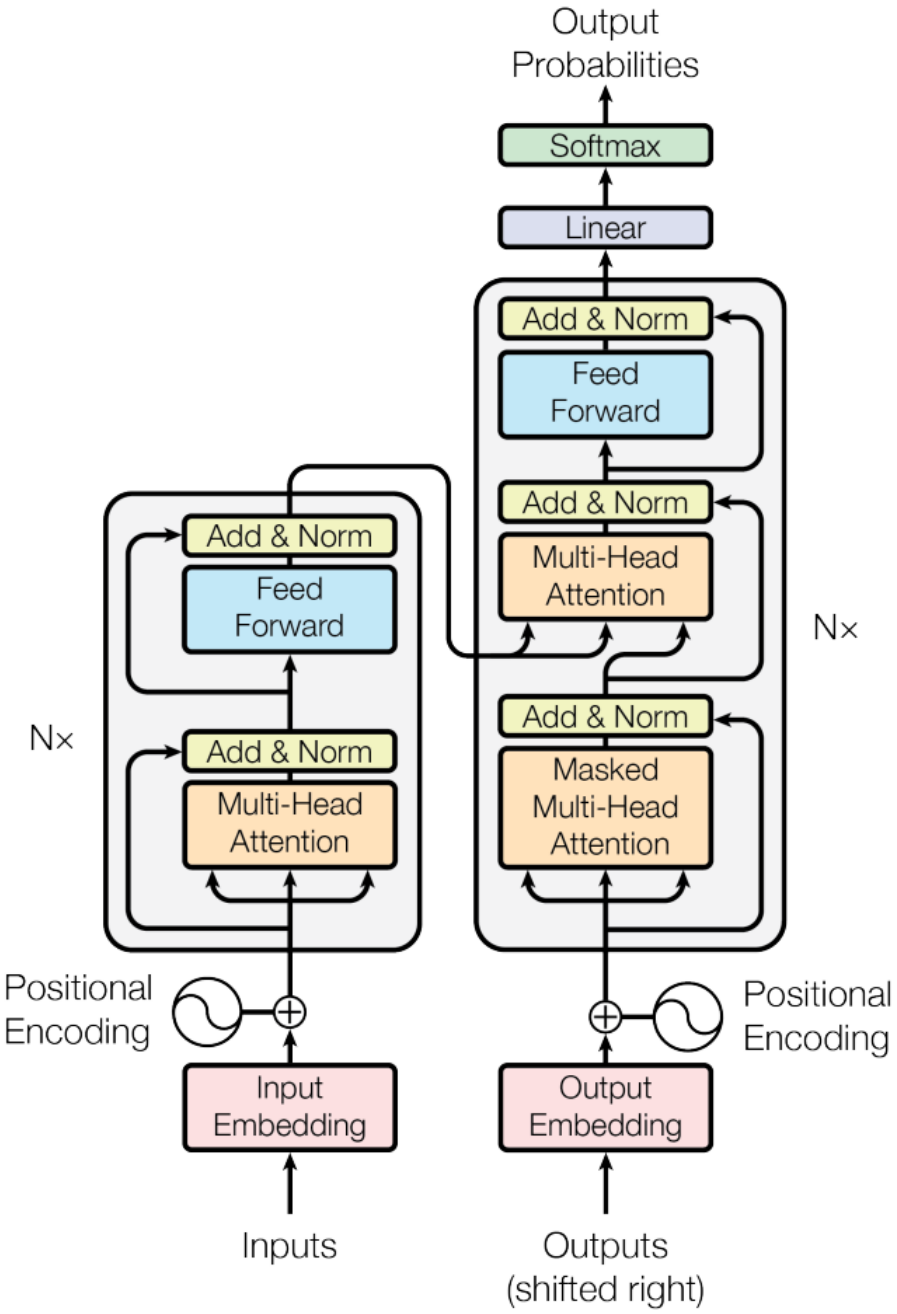


Figure 13: The Transformer - model architecture.[47]

The name BERT originates from the paper that formalised it [4] and stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, and is defined by the authors as a "conceptually simple and empirically powerful" method. Compared to the first model, the Transformer, BERT is optimised in its fine-tuning phase, using an MLM[9], "randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context" [4].

[9]MLM: Masked Language Model

The algorithm used by KeyBERT is built on top of BERT, from which it first extracts document ebbeddings and then word embreddings, which are extracted for N-gram words/phrases. Then cosin similarity is used to identify which words/phrases are most similar to the document. The most similar ones will be the best possible candidates to represent the document, and then extracted as keywords. KeyBERT is a simple and fast method for extracting keywords, but it is certainly not the only one, nor the most powerful, but the best one for the purposes and necessary compounds of this research.

4.1.2 Extraction process

KeyBERT makes it possible to highly customise the type of extraction required, such as, for example, setting a parameter to determine the degree of diversity desired in the extracted keywords, or choosing the embedding model to be used, being able to pick between different frameworks, such as SpaCy or Gensim. However, in this case, the basic KeyBERT model worked very well, the only parameters defined being two: the choice to extract only single keywords and not phrases, in order to make the papers more easily comparable, and the number of keywords to be extracted set at seven. The decision to choose seven keywords was empirical: taking a subset of papers with a length similar to the average length of 2 585 characters, seven was the number that best captured all the nuances of the abstract, but without creating redundancy in the type of words extracted. It is therefore the number that, on average, best retains all the information contained in the abstract, without generating superfluous information. A close inspection of the dataset shows how, however, the number of keywords is often less than seven. This happens mainly with short abstracts for two reasons, either the number of words that can be extracted is just under seven, or the same words are extracted, and the algorithm is set to only keep unique, non-repetitive words. Another reason why this may happen is an opportunity to explain the next step that occurs after word extraction: lemmatisation. Lemmatisation is that process by which each possible inflection of a word is brought back to its lemma. So that we can compare, for example, the word "better" with "good", or "events" with "event", and so on. In this case, a lemmatizer from the nltk library is used: WordNet, which returns the lemma of a word if it can be identified, otherwise the word itself. A final reason why the keywords may be less than seven in number is the fact that all those words that were used to filter the papers at the beginning of the process ("weather", "climate", "environmental" ,"atmospheric" ,"meteorological" ,"change" ,"changes" ,"shift" ,"shifts" ,"alteration" , "alterations" ,"transformation" ,"transformations" ,"extreme" ,"severe" ,"harsh" ,"unusual" ,"conditions" ,"condition" ,"event" , "events" , "pattern" ,"patterns" ,"phenomena" , "episodes"), which would otherwise surely be predominant, are intentionally excluded. The flow of the process is therefore as follows: first the words are extracted from the abstract. Of these, those that were part of the word list with which the papers are filtered are removed. Then a lemmatisation of the set of words is done, and finally the unique values for each word are kept.

Figure 14 shows an initial analysis of the keyword extraction results. In particular, a distribution of the density of the number of occurrences per keyword is shown. That is, the number of keywords that are present in only one paper out of the total number of papers, the number of keywords that are present in two papers out of the total, and so on. It can be clearly seen that the distribution is right

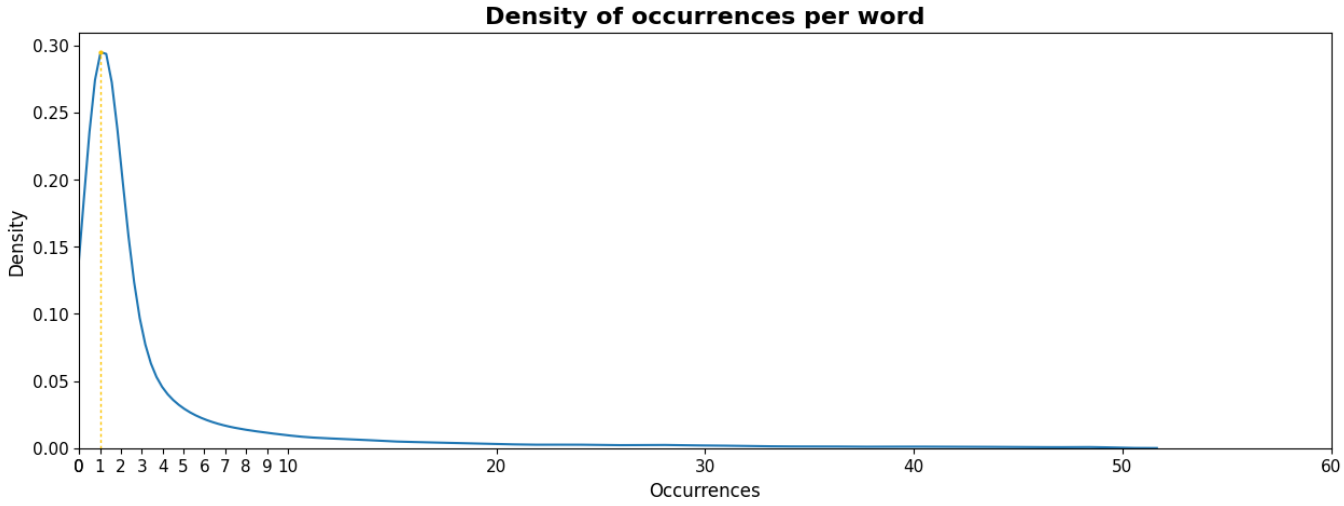


Figure 14: The graph shows the distribution density of occurrences per keyword found. That is, the number of papers that have that keyword

skewed, most words appear only once, but that still a good number of keywords are shared by at least four papers. The distribution is cut off at 50 occurrences, even though there are words appearing in more than 8000 papers, because the tendency is to squeeze more on zero, as can easily be seen from the graph. Approximately 30 per cent of the keywords are unique, i.e. related to only one paper, but this means that 70 per cent of the keywords are shared by at least two papers, creating the possibility of thematic connections between papers.

After an analysis of the results, as can be seen in figure 15 in the left graph, where the number of occurrences per word are represented, it can be seen that the lemmatisation needs to be more accurate, because for example the word 'agricultural' should have 'agriculture' as a lemma. In general, words ending in -al, such as hydrological, ecological etc.. do not add to the count of hydrology or ecology, as is intended. Furthermore, words such as 'warming' have not been reduced to the lemma 'warm', which is another desired goal. So in the first place, a lemmatiser other than nltk was used to refine the result. The SpaCy library is then used to perform another lemmatisation, which improves the quality of the result but does not solve the problem of words ending in -al. Therefore, after having explored various alternatives, such as Gensim, Pattern, TextBlob, and Stanford Core NLP, and not solving the problem with any of them, a manual adjustment is carried out, but only acting on those words that appear more than 500 times, about 4000. The final result can be seen in figure 15 in the right graph.

4.2 Geographical Information Extraction

The second step is to associate each paper with information, if any, on the geographical location to which they refer. Firstly, an attempt was made to go along the way of analysing authors' affiliations in order to determine the places where research on extreme climate events is carried out. However, by moving in this direction, two problems immediately arose: the first is that, by definition, on the Semantic Scholar database [19] an author's affiliation is only the last one available temporally for that author, or the last one that was entered on the platform. This therefore implies that, for each author, the

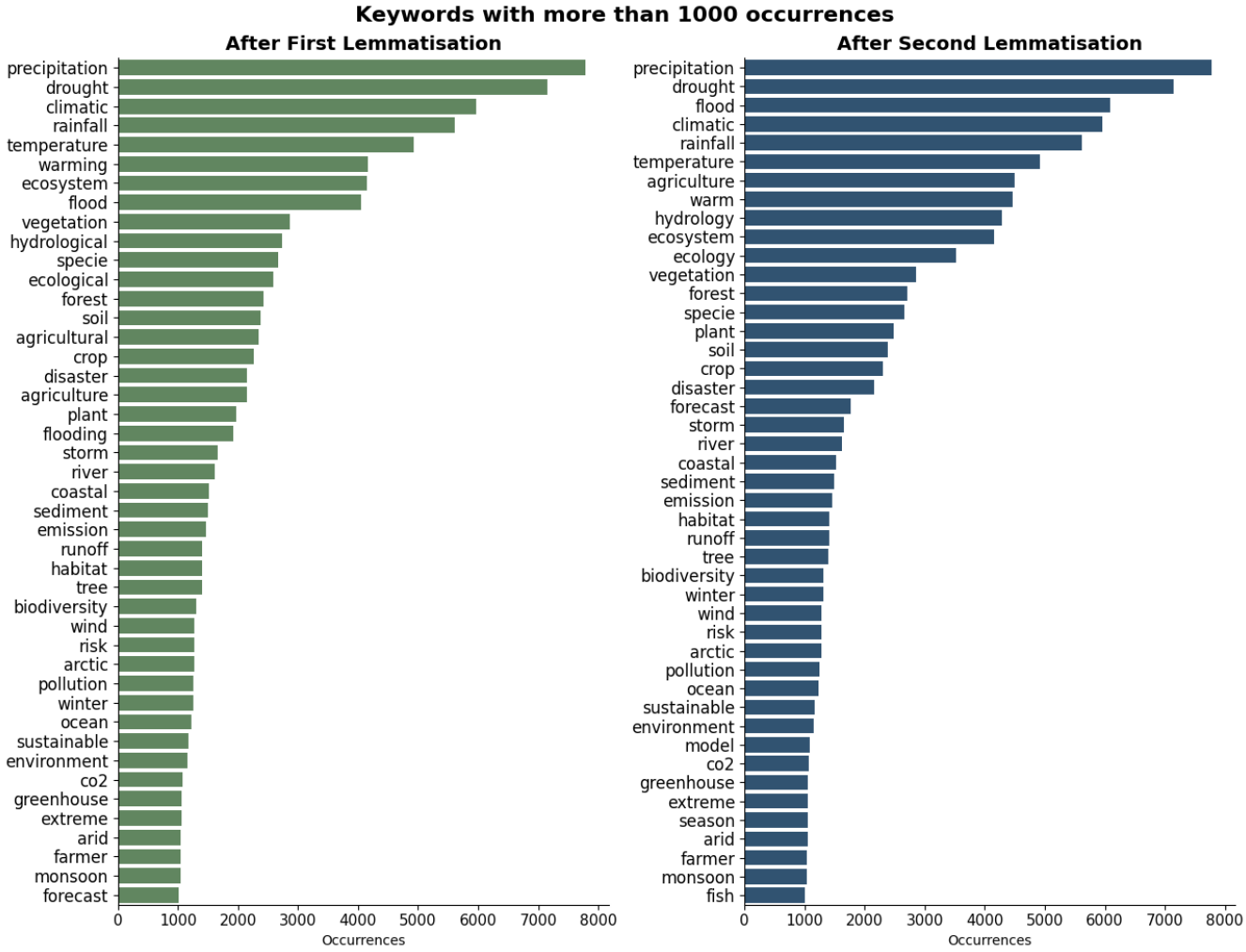


Figure 15: The image shows, on the left, the keywords that have more than 1000 occurrences, before the second lemmatisation is performed to clean the results. On the right, again the keywords with more than 1000 occurrences but after cleaning, and therefore belonging to the final dataset.

affiliation information would be taken without considering whether the published paper was actually written by the author at the time he or she was affiliated with that institution, or perhaps the paper was written and published under a different affiliation, possibly therefore in a different place. The second reason why taking the affiliation route is unreliable is the fact that only 1 060 papers are linked with an author having affiliation information. This therefore leaves the road to affiliation impracticable. Texts are therefore processed again to extract geographical information from the abstracts. In this case, therefore, it will not be an analysis of the places where the research is carried out, but the places that the academic research concerns. The process of extracting geographical information involves three stages, with different tools used:

1. Using the SpaCy [36] library, NLP[5] is performed on the abstracts and terms that have a geographical reference are identified.

2. Thanks to the Geopy library, the geographical places identified with Spacy are located, their countries, when possible, and coordinates are extracted.

3. The results are cleaned and thanks to the Pycountry library, the results are classified into nations, continents, oceans/sea/bays.

4.2.1 Identifying geographical terms

The first step is to process the abstracts to be able to identify which words contained in them can help identify a geographical area to which the paper can be linked. SpaCy offers an NLP tool that allows each word to be associated with a label that identifies whether that word is a noun, or a verb, etc... The number of different terms that can be identified is extensive depending on how the algorithm is set up. Among its basic functions, however, it has the detection of those terms that refer to a geographical location. In particular, the label "GPE" identifies a term representing countries, cities and states, while the label "LOC" identifies Non-GPE locations like mountain ranges and bodies of water. Using only the label GPE would lead to very clean and easily classifiable results in countries. However, it can be assumed that the researched papers often speak not of cities, but of natural places such as protected areas, rivers or seas. Hence the decision to also include those terms that have "LOC" as a label, which will require further effort to clean and accurately classify places that will also have very different natures, as the granularity of the data to be obtained is in any case at the level of nations. Moreover, this approach creates a whole series of problems concerning which nation to associate with, for example, a river that crosses several states, or a sea that washes several of them. However, all the extra work, some of it manual, that this entails benefits the completeness of the analysis. This provides geographical information for 75.55% of the papers in the dataset, immensely more than those papers that have information on the affiliation of at least one of their authors. The requirement to keep geographical locations at country level stems from the fact that the purpose is again to compare different papers. Keeping therefore a level of granularity equal to that of the places found directly in the papers would increase further entropy, with many places being unique values in the dataset, thus rendering them useless for comparison. This is why it was necessary to find a tool to standardise the results.
This was done using GeoPy [9], a library that allows the user to search through APIs of geolocation services, such as Google Maps, for the addresses of a given place, including its coordinates. Or alternatively, given the coordinates, it allows the location to be traced back. As can be seen in figure 16, GeoPy is a bridge that connects its codes to these services that are carried by different platforms.

Among the APIs to which GeoPy connects are the most popular ones from Google and Bing, which, however, offer commercial services, and thus charge a fee. Nevertheless, the API of OpenStreetMap [29] is also available, which is a tool that offers online maps held together by a community of mappers and engineers, and is non-profit orientated. The API is therefore free and the data on the platform is free to access, but given the non-commercial nature, there is a rate limiter of the API that poses challenges, as in this case there is a large amount of data to process. The limit set by the API depends on the load borne by the servers at the time of the request. In particular, the API does not tolerate repetitive requests on the same location, and those made in a serial manner without an interval between requests. Moreover, there are some cases where the location is peculiar, or is not a location

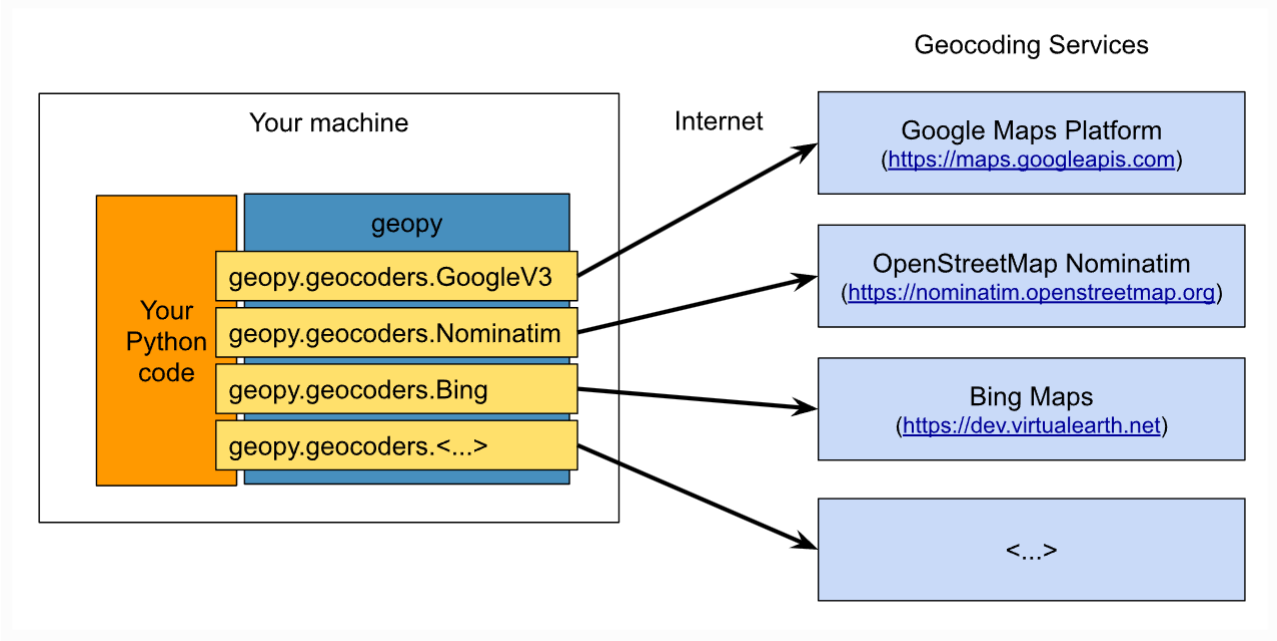


Figure 16: Mechanism of GeoPy's functioning. Source: [9]

well identified by SpaCy, so GeoPy fails in its identification, and the code must also be written to handle these errors. The objective is to extract, where possible, the reference country of each location. GeoPy as a result of the query provides a dictionary containing, among other things, information on the type of address, so whether the location is a road, rather than a mountain or a continent, the address itself and the co-ordinates. From the address, the country can be derived for all those places that are within one. However, there are places, such as continents, seas, oceans, Antarctic places etc., that cannot be identified with a nation. For these reasons, the code is set as follows:

- A function is written that invokes the OpenStreetMap API to find the address of each location. When an error is raised due to the rate limiter, or the location is not found, the request is repeated, spacing one second between each, 5 times. If an address is still not found at the 5th time, that location will be discarded.

```
def do_geocode(address, attempt=1, max_attempts=5):
try:
    location = geolocator.geocode(address, language='en
    ↪ ')
    stringa = None
    return location, stringa
except GeopyError:
    if attempt <= max_attempts:
        time.sleep(1.1)
        return do_geocode(address, attempt=attempt+1)
    elif attempt > max_attempts:
        location = 'Not Available'
```

```
        stringa = f'request failed for {address}'
    return location, stringa
```

- When searching for locations identified with SpaCy, the type of address is checked, taking the information whether it is a country, continent, ocean or sea. Four respective dictionaries are created in which the results of each query to OpenStreetMap are stored, so that if a specific location has already been searched for, the code is not repeated, obviating the problem of repetitive queries per location. All address types that do not fit into one of these four classifications are saved in the country dictionary. The data will then be cleaned.
- Finally, the locations and their coordinates are saved in the dataset. There will then be a column that for each paper has the list of unique locations, and another column that has the respective coordinates of each location.

4.2.2 Data cleaning and harmonisation

The results of the extraction described above are generally good, but there are some problems that need to be fixed. Firstly, there are some locations from which the country was not extracted because the address was not found. This is the case for some locations that represent natural places, such as "Goguryeo Hill", a Japanese location. In particular, this problem exists with many locations in Antarctica, which is not a nation and is not identified as such by GeoPy. These locations have therefore been manually adjusted and classified, as the number is not prohibitive. All locations in the South Pole have been classed as "Antarctica", so that they can be grouped together, despite not being a nation. A further clean-up performed was on the address types. In fact they are various and on a detail level not necessary for this analysis. For example, GeoPy distinguishes between "water", "waterway", "river", "canal" to indicate a river, or some locations may be ambiguous, such as San Marino indicating both city and state. Therefore, the address type has been remapped, so that country and continent are used for all those cases that are clearly identifiable, and gulf, river, sea or ocean for those cases where a country cannot be identified. A column is then added to the dataset concerning the address type, which will be useful later. The final type of information to be obtained in the dataset is to have a column containing the nations with which the paper is associated, a column containing the continents with which the paper is associated, whether the continent is extracted directly from the abstract or derived from the other locations present, and finally a column representing waterways, gulfs, seas, and oceans that are not clearly identifiable with a nation. The most complete information that can be obtained relates to continents, which is why the continent was also abstracted from the nations, thanks to the pycountry [33] library, which associates the names of countries with their ISO 3166-1 code [37] by which the continent can be traced. There are some nations, such as Kosovo, that do not have an ISO 3166-1 classification because they are only partially recognized internationally. For them, a fictitious code was created, not in conflict with existing codes, as was done for Antarctica. The dataset will thus have 6 columns in total containing for each paper the lists of nations, continents, and generally related waters, and the coordinates of these. However, information on the original locations is retained.

4.3 NLP results

Giving an ambiguous and comprehensive representation of NLP results is a complex challenge. On the one hand, there are the keywords, which by their qualitative nature pose a challenge in being able to represent their completeness, and limit the use of metrics available for analysis. On the other hand, geographic information leaves more choice of representation, but it is not guaranteed to be successful in conveying accurately and faithfully what is present in the data.

4.3.1 Keywords

A first look at the result of the keyword extraction has already been given in figure 15, where it can clearly be seen that the main theme emanating from the abstracts is meteorological, in particular relating to extreme events such as droughts or floods. But the environmental theme also emerges clearly, there is mention of ecosystems and vegetation, agriculture and crops, but also destructive events, such as storms and the related water runoff, or disasters in general. In order to represent in an unorganised manner the topics most commonly found in the abstracts, and to give a general overview, a word-cloud was created, which can be found in figure 17. This word-cloud takes the form of an atlas, but it is merely a stylistic choice, and an introduction to the type of analysis to be carried out later, also concerning cross-analysis between topics and geographical areas. But in no way does it represent, at this stage, a prevalent association of words with the geographical area in which they are placed. Nor is it intended to create associations that could lead to erroneous conclusions from words that are placed close together.

What emerges from the word-cloud is similar, of course, to what emerged earlier, i.e. clearly the theme of what concerns hydrology, in its more usual or violent forms, the natural world in its mainly arboreal guise, but also what is related to emissions and CO^2, thus indicating a focus not only on the effects, but also on the causes of climate change.
While the word-cloud is a good tool for getting a general overview and grasping the theme, it is not equally useful for extracting precise data, since one could even try to find the most recurring word with difficulty. Therefore, the need arises to at least make a more accurate classification of the number of occurrences of each keyword. Figure 18 shows the number of occurrences of the ten most frequent keywords for each continent.

Observing the graphs, it is already possible to start observing the differences in the themes of attention towards climate change in the different continents. Precipitation is the most prevalent theme in Europe, Asia and America, but if we look at Oceania, we see at the top two words that are antithetical to each other: drought and flood, which might indicate a more extreme impact of climate change in general in this area of the world. Similarly if one looks at Africa, it seems that the main concern is agriculture and drought. In general, the issue of warming and temperatures is also very present. In Antarctica, on the other hand, one can see that warming is among the major points of concern, but there is a very interesting aspect, namely the presence of the word "microbial", a theme that can be considered both in relation to the melting of glaciers and the particular research conditions in these places. The North Pole is also present in the data, fictitiously classified as a continent. But

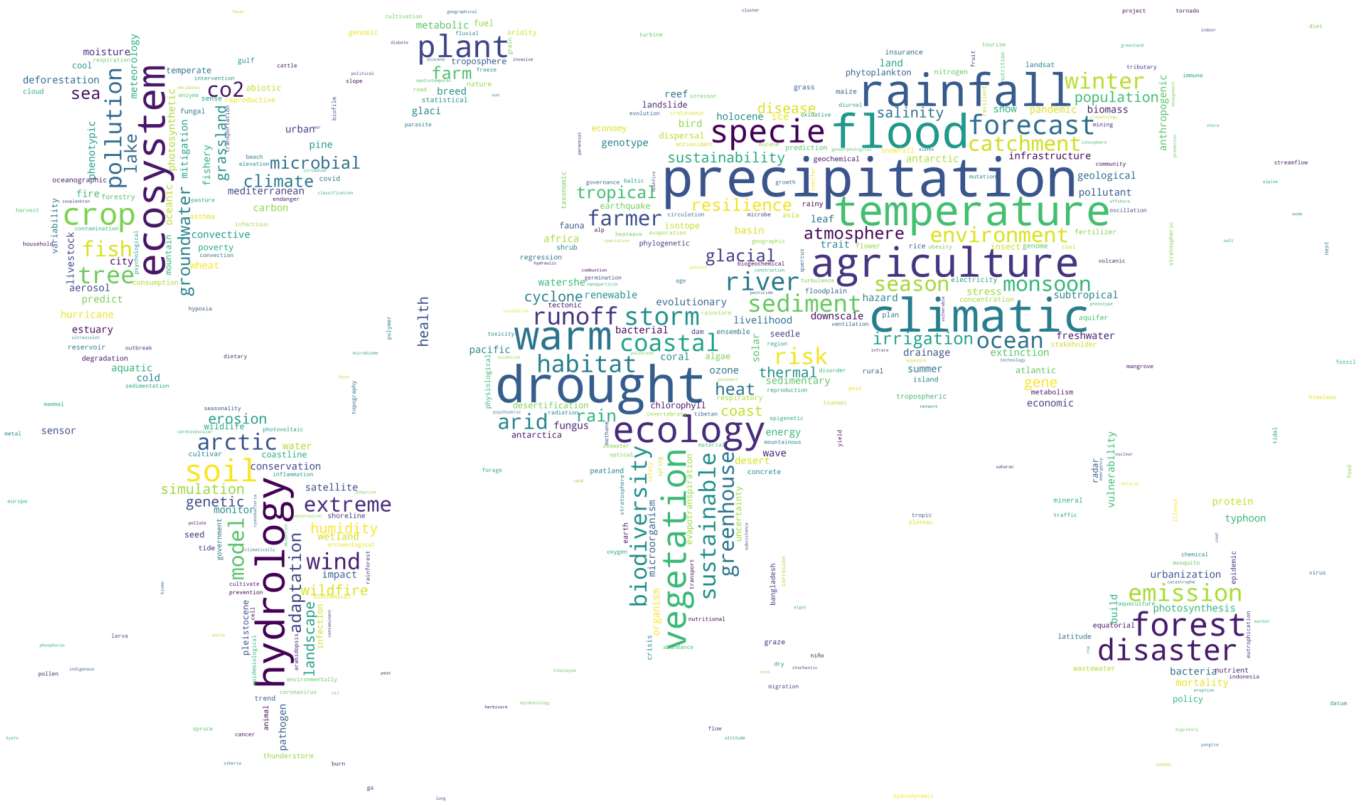


Figure 17: Atlas shaped word-cloud representing more frequent keywords

the number of papers falling into this category is only three, and the only keyword present in more than one paper is "glacier", present in two. A comparison in a barchart, which would have been essentially flat, would therefore have made no sense. For the sake of completeness, however, the keywords for the papers classified as North Pole are given: "glacier", "astronomical", "celestial", "climatic", "cryoseismology", "drought", "earthquake", "observatory", "photometry", "precipitation", "rainfall", "seismicity", "sky", "tectonic", "telescope". It is important to note that in the graphs in figure 18, information on the quantity of keywords per continent is not captured, the barcharts are not scaled according to the number of words. This type of analysis is left to the next section, in which the results of the geographical classification are presented, and in which one can see which nations are most magnetic for the accademic research attention on the theme in hand. Here, this aspect is left hidden to a not particularly alert eye, and may even be misleading, because the objective of this section is to analyse the results of the keyword extraction through the NLP done. Subsequently, the dots will be joined to make an overall analysis of the results. However, looking at word occurrences by continent, North America, Europe and Asia have similar values, followed by Africa and then South America and Oceania. The absolute value of this number, in any case, has little meaning if other metrics, such as the number of inhabitants per continent, are not taken into account.

Therefore, sticking with the need to analyse the data resulting from the NLP more qualitatively than quantitatively, another graph is constructed, which does not take into account the actual number of occurrences, but the rank of the keywords, thus putting all continents on an equal level. Common words

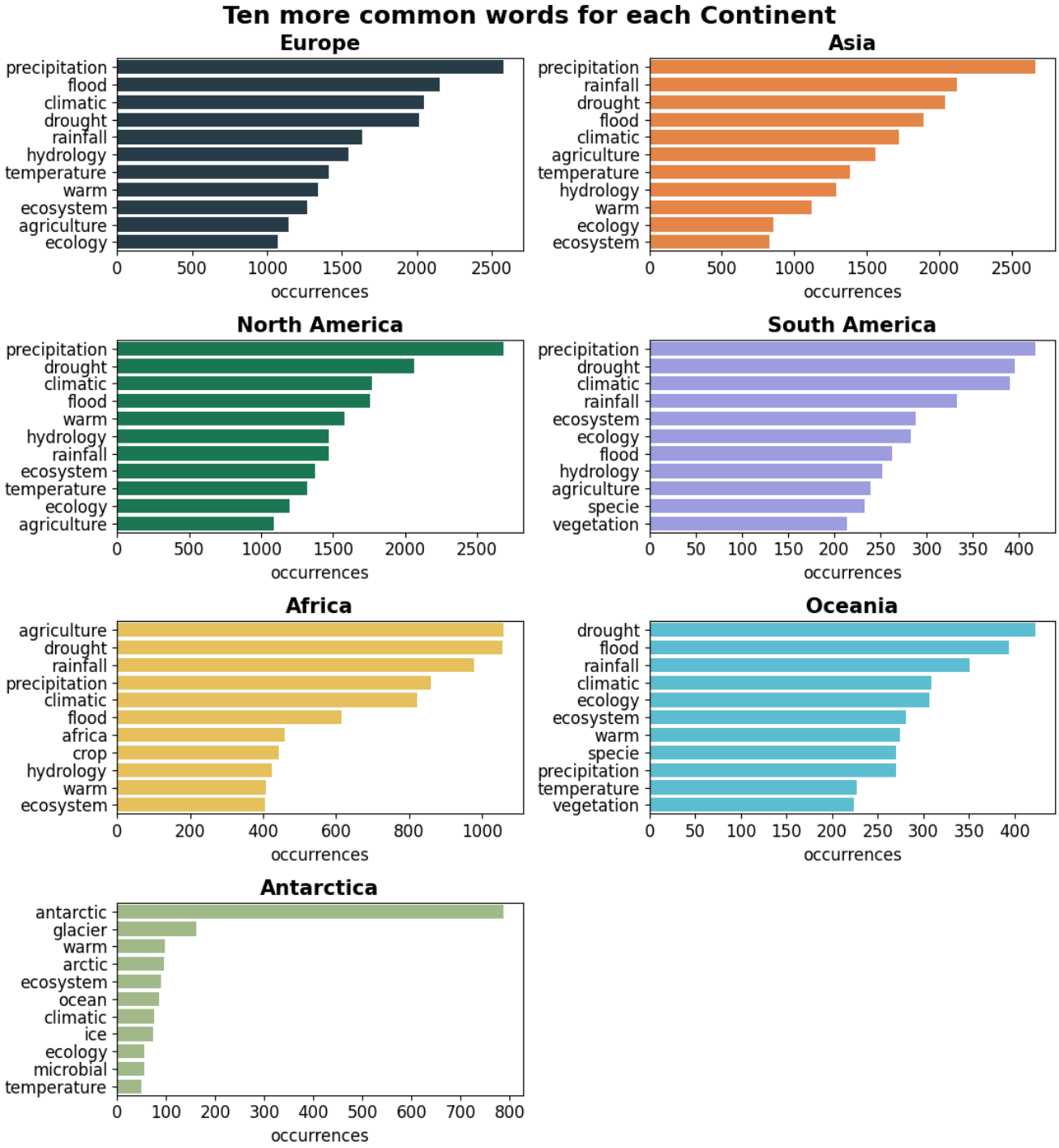


Figure 18: Number of occurrences of the top ten keywords per continent

are taken from the top 15 by number of occurrences on each continent, excluding the South and North Pole, which have a significantly different set of words, and would have made comparison tricky. The resultant common words are "warm", "temperature", "drought", "agriculture", "ecosystem", "flood", "precipitation", "hydrology", "rainfall", "climatic", "ecology", "vegetation", which are placed in a radar plot, visible in figure 19, generated equally for each continent. The keywords are placed in such an order as to attempt to respect a degree of thematic sense, i.e. words that are similar are placed close together, and words that are instead quite different, or opposed, as in the case of "flood" and

"drought", are placed far apart. In this way, the radar plot should give a sense of the thematic area most present for each continent, and the shape taken by each is a means of finding patterns across continents. The conclusions that can be drawn are similar to those derived from the barchart, i.e. that there is a certain consistency between North America, Eurpoa and Asia, while different themes emerge from the other continents. It can be seen that the radar plot for Australia has an elongated shape, again suggesting the relevance of opposing events. The same can be said of South America, where the importance of the ecological theme is also clear, creating a bridge to Australia.

4.3.2 Geographical classification

The representation of geographical results leaves much more choice than that of keywords. In this section, the intention is to analyse the geographic distribution of the papers, in particular the localities mentioned in the papers. In doing so, the text goes through the steps of the analysis, starting with the individual localities and coming to the aggregation first by country, and then by continent. It should be noted that those localities that refer to oceans or seas that cannot be directly associated with a nation are also classified in the data: these localities will be visible through the representations, but will later be omitted in favour of those localities directly identifiable with a country. For two reasons: the first is that taking these localities into account would complicate the analysis due to the absence of a common denominator with which to compare them with the identifiable localities; the second is that their number is insignificant compared to the total number of localities, so their inclusion or non-inclusion does not meaningfully influence the analysis.
The most direct and cleanest result of the extraction for geographic data in the abstracts can be found in figure 20, where each point represented is an ambassador for a location mentioned in a abstract. It is immediately noticeable, as anticipated by the number of keywords present per continent, that the highest density of locations is found in North America, Europe, and Asia. It is important to remember, at this stage, of one of the biases surely present in the data, namely the choice of the English language for the filtering of papers, which surely leads to the exclusion of local language research done by the various countries, and advantages those nations where English is an official language, first of all the United States, United Kingdom and India, which indeed are high density areas in this map. However, the high density in these places does not necessarily have to be traced to this reason, as certainly the academic research done in these countries is substantial, regardless of the language used.

The next atlas 21 shows the grouping of all these points by nation, where the size of the bubbles represents the number of papers that ended classified in that nation. Given the map in figure 21, it is not surprising that we can confirm the results discussed so far by looking at this map. The transition from the map in figure 20 to the map in figure 21 is also a representation of the cleaning that has been done on the raw data extracted from the abstracts, arriving at the result that will serve as the basis for the next analysis. From the map there is little evidence, because it is rightly small in comparison to the numbers made by other nations, but it is classified, as previously pointed out, as a state even Antarctica, which collects seventeen papers.

To convey an idea of the magnitudes in which a country is mentioned in an abstract, the last graph

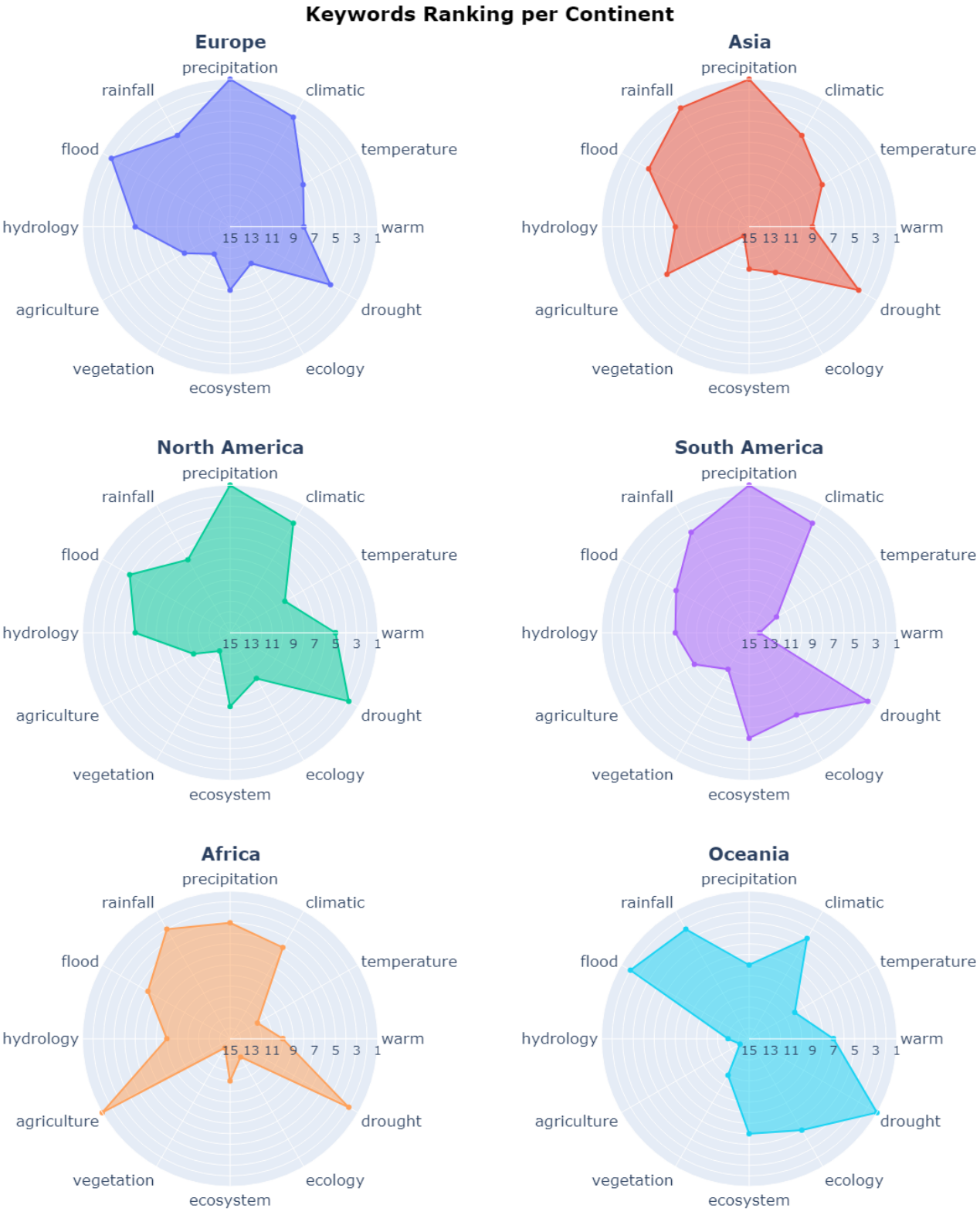


Figure 19: Representation of keyword rank by continent, taking into account words shared between continents among their top fifteen by occurrences

in figure 22 was produced. It is a Voronoi Treemap, implemented in javascript, unlike everything else. This graph was created thanks to the GitHub repository Kcnarf/d3-voronoi-treemap [18] and

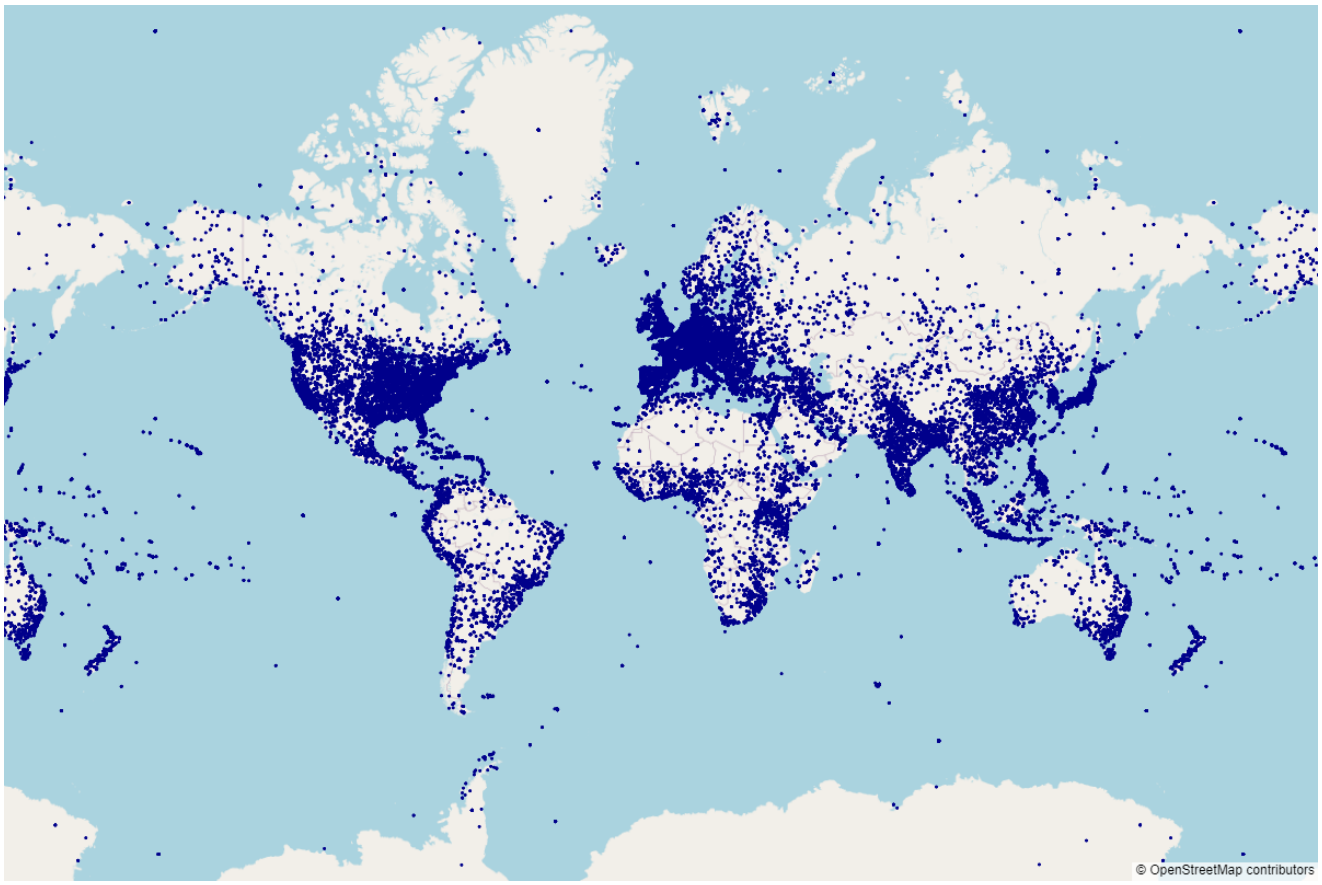


Figure 20: Each point on this map represents the coordinates of a location extracted directly from the abstracts, before any further processing.

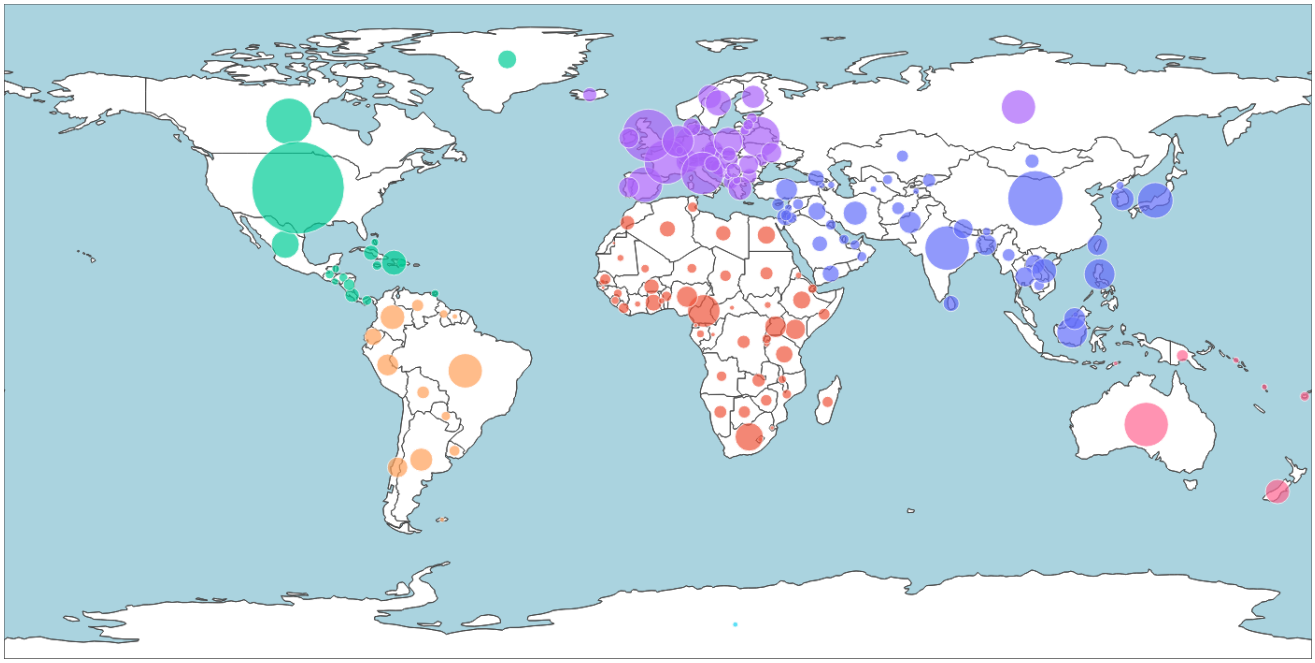


Figure 21: Each bubble in this map represents a country, and its size the number of papers in whose abstract it is cited.

specifically on the basis of the codes contained in Kcnarf/.block [17], which have been repurposed and modified in order to be able to represent the data covered by this thesis, but which represent the main skeleton behind the figure. The graph is useful because it allows to have, at a glance, information on the portion of the papers in which a nation is mentioned, but also the possibility of having a direct and immediate comparison with other countries, as well as establishing the weight of each country within its own continent. What has been concluded so far is confirmed, namely that the majority of papers speak of North America, Europe and Asia. The most important continent, in terms of numbers, is Europe, mentioned by 33 399 papers, followed immediately by North America, with 27 710. The substantial difference between the two, which reflects many other graphs of a macroeconomic, or social, nature, is the fact that in North America there is only one major player, the United States, while European research is well divided among all its states. The polarisation observed in North America is not replicated on any other continent, with the exception of Oceania, where, albeit to a lesser extent, Australia dominates. A clear evidence that emerges is the absolute dominance of the united states over all other nations, in fact they are cited four times more than the country in second place, China. At process level, it can be concluded that the result in the geographic classification of papers is satisfactory, and opens the door to a whole series of cross-analyses that can be done with the keywords, but also with other indicators, whether economic, social or environmental, enriching the simple picture that can be drawn from this brief representation.

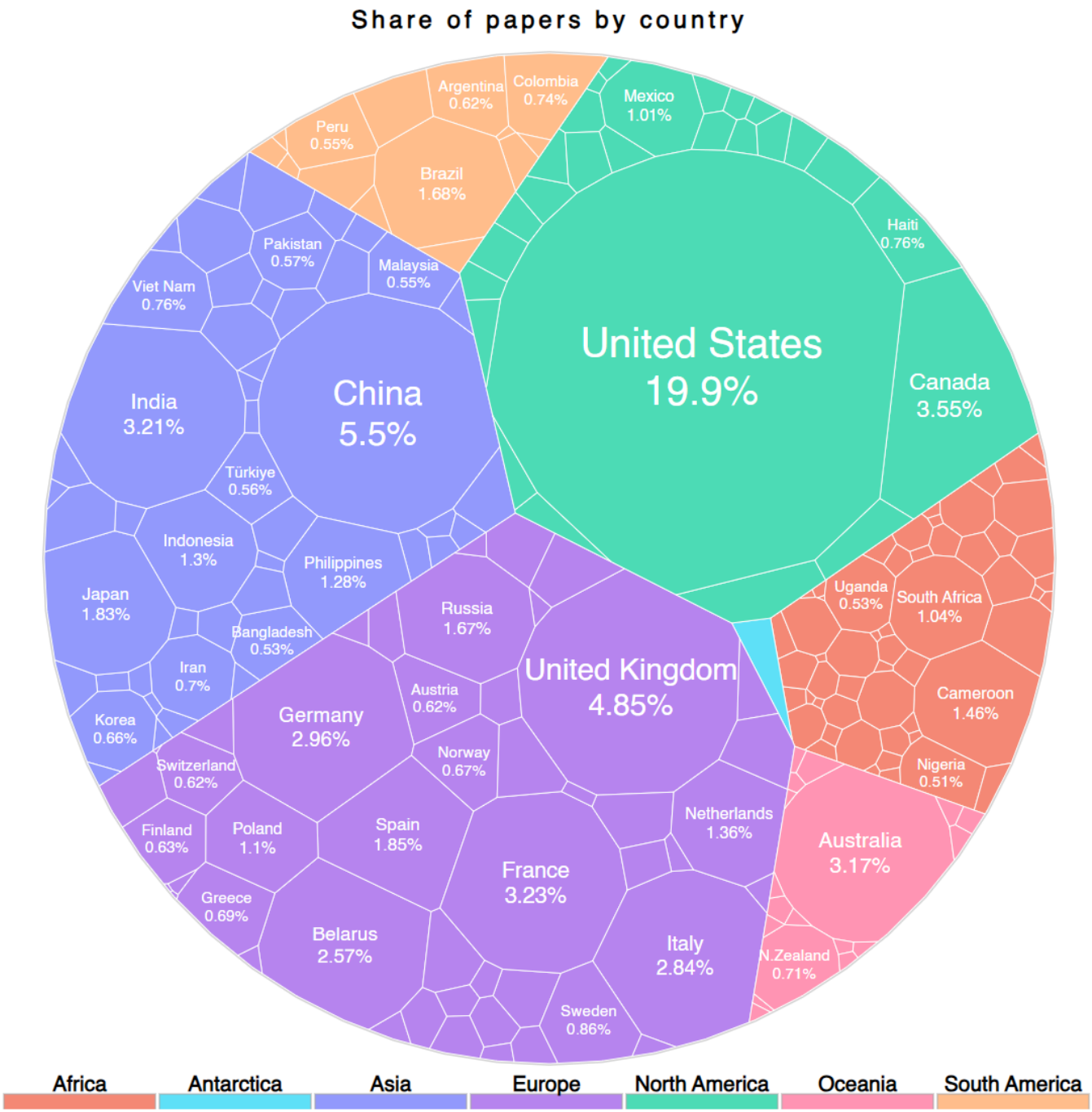


Figure 22: Voronoi treemap where the percentage of papers out of total in which a nation is cited can be compared. The nations are clustered by contient. The size of the areas is directly proportional to the number of papers.

5 Community detection

The unambiguous nature that words have by definition makes it difficult to make a quantitative analysis. The huge number of identified keywords makes a qualitative analysis impossible, if not cursory. This is why use is made of a methodology that makes it possible to attempt to cluster the kewywords in such a way as to create silos that can represent the thematic area mentioned in the paper. To achieve this, recourse is made to graph theory, and in particular network theory, already mentioned in the section 2.2. The objective of this section is to explain in an exasustive manner how the network was constructed that allowed the identification of thematically similar communities of papers.

5.1 Building the Network

The reason for processing the texts to extract the keywords associated with each paper was to have thematic representations of each paper that would allow the data to be processed in a more streamlined way, in addition to the need for abstraction. It is therefore logical that the network is constructed precisely by exploiting the keywords previously extracted. The idea is to have each paper, through its corpus id, represented by a node. It is then necessary to define the rules by which between each node, i.e. paper, an edge is created or not, i.e. there is thematic affinity. To do this, the Jaccard similarity between each possible pair of papers in the dataset is used. Jaccard's coefficient is often used to define the similarity, or dissimilarity, between two sets [26]. It is calculated using the number of elements in the union between two ensembles, i.e. the intersection of the sets, divided by the number of elements in the union between the two sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A is the set of keywords of paper a, and B the set of keywords of paper b. The first step is then to compute the Jaccard coefficient between each possible pair of papers in the dataset, creating a dictionary in which the first paper of the dataset is taken as the key, and its value is another dictionary, containing as keys the set of all the other papers in the dataset, each of which has as value the Jaccard similarity between the first paper and itself. In this way, however, with a simple combinatorial calculation formula, it becomes understandable that there are too many elements that would compose this dictionary. Indeed using:

$$C(n,k) = \frac{n!}{k!(n-k)!}$$

Where n is the number of total papers, and k is 2, since only pairs of papers are taken, this results in 2 617 586 835 possible combinations of papers. Even if the computational capacity of a laptop were capable of handling such a large object, it would still take too long. Luckily, many of the papers have Jaccard similarity of 0. Therefore only those pairs with a Jaccard coefficient greater than 0 are included in the dictionary. This considerably reduces the size of the dictionary, which is nevertheless still very large. From a few performed explorations, the number of papers involved in the analysis is still too large to create a network and effectively manage it, given the computational power constraint.

Furthermore, it will be crucial to compare the result of the community detection also geographically. For these reasons, only papers for which geographical information could be found, published from 2011 to 2021, are retained. The decision to reduce the years is due to two reasons: on the one hand certainly the necessity to contain the number of nodes in the network for computational reasons, on the other hand circumscribing the analysis temporally certainly makes it more accurate from a thematic perspective. Indeed, making a connection between two papers that have thematic affinity but were perhaps published 30 years apart could lead to inaccurate conclusions, given the evolution of both the climate topic and research techniques. It would certainly be interesting to thematically follow the evolution of a given topic over time, but this would certainly require a different approach.
Once the pool of papers on which the network will be built has been defined, it is necessary to define the rule by which an edge between two nodes is generated or not. To do this, a series of histograms have been constructed, showing that most pairs of papers have a Jaccard coefficient of less than 0.6. Therefore, this boundary is used as the cut-off beyond which an edge is generated. All pairs of papers therefore with a coefficient greater than 0.6 are linked together. Based on this, the network is constructed, which turns out to have 7 361 nodes and 17 827 edges. However, given the ultimate goal of community detection, it is necessary to verify that the graph is connected. Otherwise, there would be at least as many communities as the number of components, which can be high. In fact, it turns out that the graph is not connected, but consists of 1599 components. Most of these components are pairs of papers that have a strong similarity to each other but are not connected in any way to the rest of the papers in the dataset as can be seen from figure 23, where indeed it can be seen that the majority of the components, more than 1200, have only two nodes, in 200 they have three. The number of components with more than 10 nodes is negligible. The main component contains 3 216 nodes, the second in size only 76. Given these factors, for the purposes of further analysis, only the main component is considered.

To build the network, in concrete terms, the python library NetworkX [24] was used, which is also useful for checking the network's connection and breaking it down into its components, as well as for visualisation and future community detection. The built graph is undirected and weighted. Indeed, Jaccard similarity was kept as an attribute of the edges, so that the weight of an edge between two very similar nodes is greater than that of two connected papers, but with low similarity. In this way it is both more accurate community detection and more effective network visualization. After removing the minor components of the graph, the resulting is not only weighted and undirected, but also connected, becoming in all respects a network. The fact that it is undirected means that the relationship is biunivocal between the two nodes, that is, both node a is connected to b, but consequently b is connected to a, and the edges have no direction.

The next step is to find, in this resulting network, the possible communities of papers that compose it. The purpose of these communities, as mentioned earlier, is to systematically cluster papers according to their position in the structure of the network, which represents the thematic relationships it has with other papers. To do this, the Greedy algorithm [25] is used to maximise the modularity of the network communities. What modularity [11] is and its origins have already been discussed in the 2.2

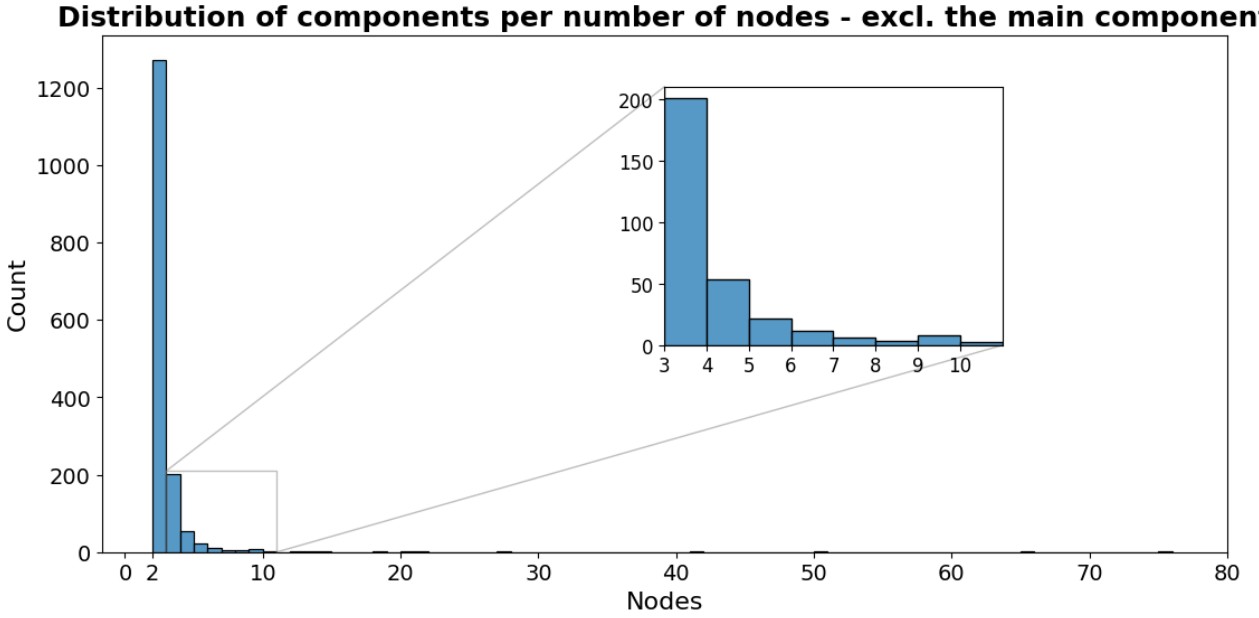


Figure 23: Distribution of components by number of nodes. The component with the largest number of nodes was excluded for a more effective visualisation, as it would have brought the x-axis up to 3216. A zoom has been inserted to better visualise the number of components in the range from 3 to 10 nodes

section, here the aim is to give a foundation on how modularity works. To do this, it is first necessary to introduce the concept of the degree of a node. The degree of a node corresponds to the number of edges incident on it, i.e., the number of other nodes with which it is connected. For example, if a node is connected, via three edges, to three nodes, it will have a degree equal to three. The degree of nodes in a random graph is distributed like a Normal, i.e. few nodes will have a high or low degree relative to the average, high instead will be the concentration of nodes with a degree similar to the average. For a non-random network, on the other hand, the distribution of node degree follows a Poisson distribution, meaning that few nodes will have a high degree, most nodes will have a low degree. This means that in a network representing some real relationship, there will be a few nodes that are highly connected to many of the other nodes, but the majority of nodes will be connected to a few other nodes, thus creating the possibility of communities clustered around the highest degree nodes, which will be the most central ones. The underlying assumption is therefore that a random network does not have a community structure. Modularity identifies a measure for when to stop the community detection algorithms, i.e. when for each community within the network the cohesion between the nodes is greater than would be expected in a random network. Modularity can thus be summarised as:

$$Q = \text{(edges inside the community)} - \text{(expected number of edges inside the community for a random graph with same node degree distribution as the given network)}$$

Given a partition in k communities $C = \{C_1, C_2, ..., C_k\}$ the modularity can be represented in a

formula as:

$$M(C) = \sum_{i=1}^{k} \left[\frac{|E(C_i)|}{|E|} - \left(\frac{k^i}{2|E|} \right)^2 \right]$$

Where k^i is the total degree of the nodes in community C_i. The objective is therefore to find the set $C = \{C_1, C_2, ..., C_k\}$ such that it maximises the modularity $M(C)$. Maximising modularity is a computationally complex task; it is a nondeterministic polynomial time problem. However, there are good approximation algorithms, such as the aforementioned Greedy algorithm, which works as follows: Initially, each node is assigned to a different community, then there will be as many communities in the network as the number of nodes. Then, for each possible pair of communities, the difference in modularity that would be obtained by merging those two communities is calculated, and the pair of communities with the greatest gain in modularity is chosen. This process is repeated until modularity is maximised, i.e. there are no more significant gains in joining more communities. Summarising all the arguments made here and in the 2.2 section, modularity is not a perfect tool, but one of the best in finding communities in networks. Among its draw backs is that given its nature it forces smaller communities into larger ones. However, for the purposes of this research, it proved to be very useful and effective. The choice of Greedy as algorithm to maximise modularity lies not only in the fact that it is still one of the best methods today, although among excellent alternatives, but it is also easy to implement thanks to the Python library NetworkX [24]. Indeed, it not only makes it possible to divide a network into communities with a single line of code, but also allows to determined the ideal number of communities to be identified within the network. The next section will explain why this is crucial.

5.2 Identifying communities

The Greedy modularity algorithm is then implemented. The aim is to find the best possible number of communities in order to have clusters that are as homogeneous as possible in terms of the number of nodes they contain. This is because it is intended to create a lowest common denominator thematically, whereby small communities focused on a specific topic would negate the effort made in trying to classify keywords. Equally, communities that are too large might risk throwing everything into the same bucket, making a comparison impossible. This is why an empirical approach is taken, identifying 6 as the optimal number of communities to divide the papers.

Indeed, as can be seen in figure 24, choosing a smaller number of communities, for example 5, would result in the first community being twice the size of the others, which are more or less the same size. Taking 7 instead, it can be seen how the seventh community is half the size of the sixth. On the other hand, looking at the first barchart, one can see how by using 6 communities, the nodes are more or less equally distributed. By using more communities than seven, or less than 5, the trend encountered is the same, i.e. an increasingly larger community encompassing the others, or smaller and smaller communities. Obviously, this graph does not take into account how the papers are distributed in the various communities, but subsequent visualisations show that six is a good number of subgroups into which to divide the network. It must also be said that six is a small enough to make a comparative analysis, both thematically and geographically, which is relevant and therefore influential

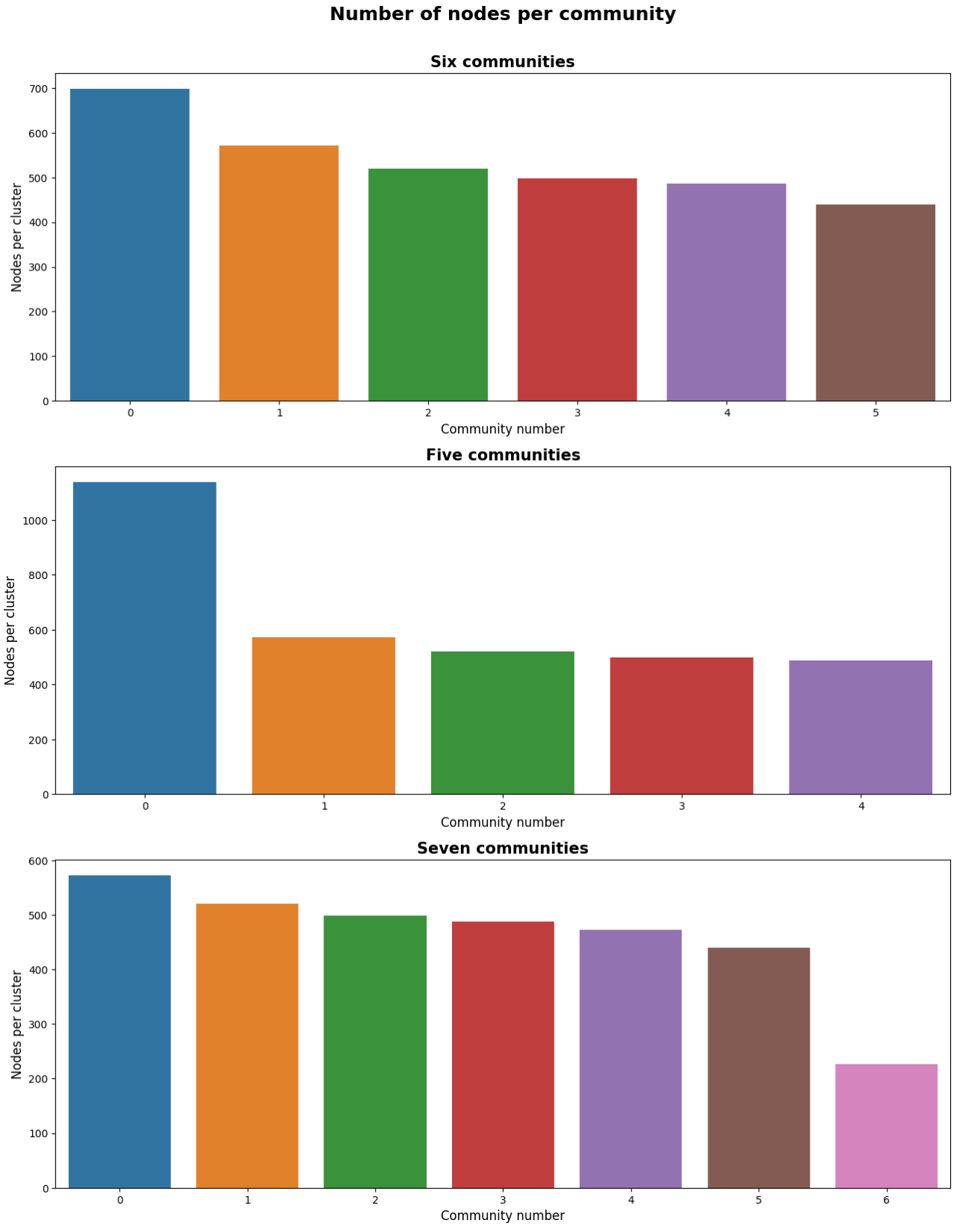


Figure 24: Barchart representing the number of nodes per community, dividing the network first by 6, then by 5 and then by 7 communities

on the number of communities chosen. Not only is an attempt made to distribute the nodes equally so as to be able to consider groups of papers in the same way, without complicating the analysis and

taking into account the number of nodes that make up a community, but also to keep this number low in order to be able to extract results that can be translated into simple and effective language. Actually, the optimal number of communities found by the algorithm without the constraint on the maximum number is much higher, meaning that from an optimal point of view the network should split the largest communities found into further subgroups. However, this would not be effective in light of the initial purpose for which community detection is done, which is to group keywords in order to make them comparable. Before trying to represent the network with a visualisation, it is worth trying to make sense of the clusters more than just the number of nodes they contain. It is known that the link between the various nodes is composed thanks to keywords, therefore it is possible to analyse which are the most recurrent kewywords in each cluster, in order to be able to find a word, or phrase, that can group the main theme of that cluster. To do this, another set of barplots is generated, where the first 7 words by number of occurrences for each cluster is displayed.

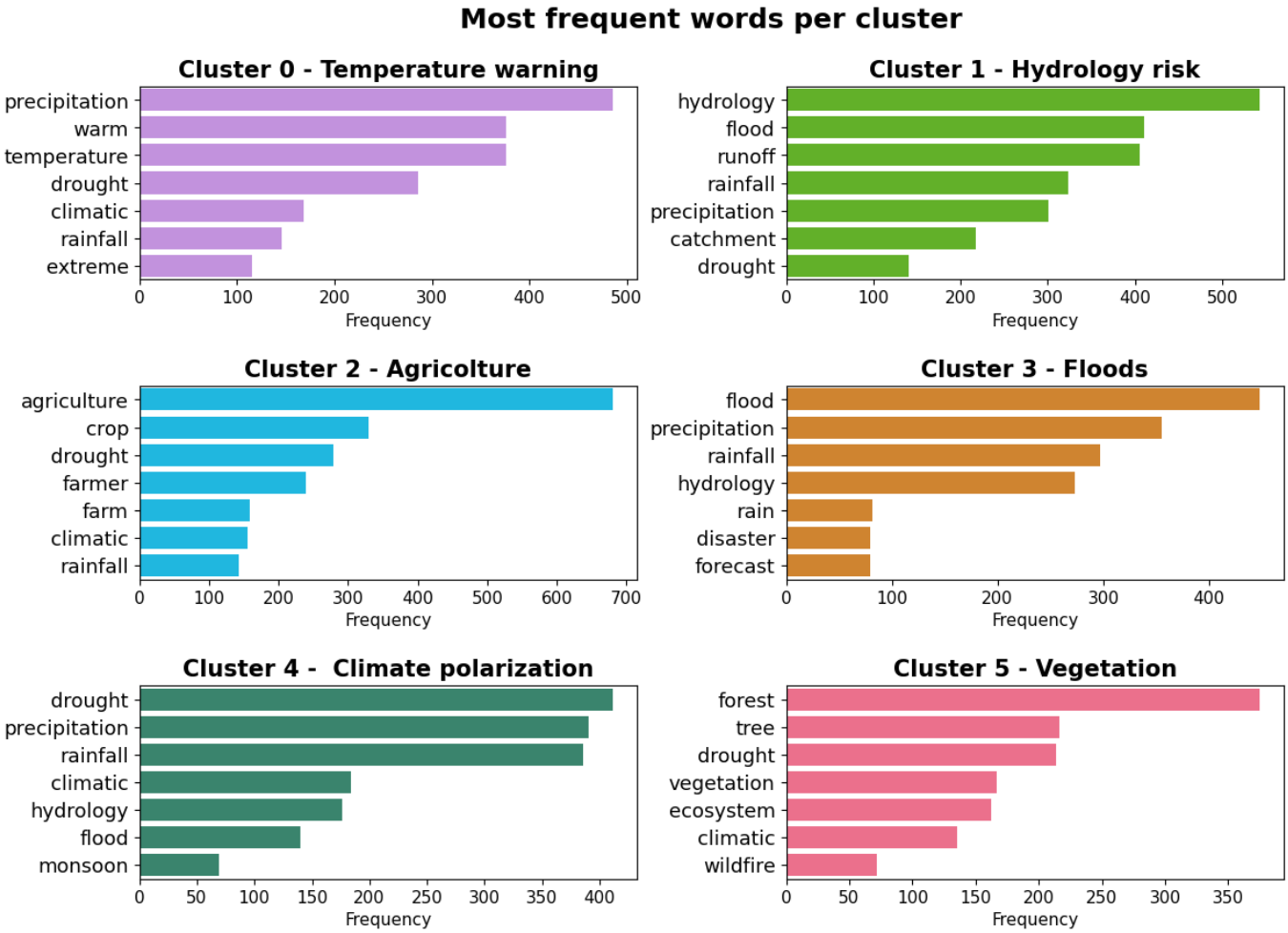


Figure 25: Barcharts representing the frequency of the seven most frequent keywords for each community identified with the Greedy modularity.

Analysing the graph in figure 25, one can become aware of the main theme of each cluster:

- Cluster 0: Although the main word in this cluster is "precipitation", it must be remembered that precipitation is the main word in the entire dataset. Analysing the others, we see that both

"warm" and "temperature" are very common, linked to "drought" and "climatic". It therefore seems that this cluster is mainly focused on concerns about temperatures, especially warmer ones. This is why it is called "Temperature warning"

- Cluster 1: The most frequent word in this community, "hydrology", already highlights the main theme. Followed by "flood", but also by water management words such as "runoff" and "catchment", it shows that the theme of this cluster is not only related to precipitation, but to the problems of managing water flows once it reaches land. This is why it is called "Hydrology risk".

- Cluster 2: The interpretation is in this case more direct, indeed the word "agriculture" is by far the most frequent, followed by "crop" and "farmer". Clearly, this cluster is called "Agriculture".

- Cluster 3: The 4th community found, similar to the second, has however a more extreme emphasis, as evidenced by the presence of the word "disaster". However, the clear theme is precipitation and its effects, which is why the cluster is named after its most frequent word "Floods".

- Cluster 4: Also similar to clusters 1 and 3, it however has the presence of the word "drought", the only case in which the word is the most frequent. It is certainly therefore an interconnected community with those related to rainfall, but it also shows the other side of the coin. Given therefore the two extremes present, this cluster is called "Climate polarisation".

- Cluster 5: The last community identified is *sui generis*, indeed it clearly concerns "Forests" and their risks, such as "drought" and "wildfire". To keep the theme generic, the cluster is called "Vegetation".

Providing a general analysis of the results, it can be seen that there are three strongly interconnected communities that have precipitation, or the absence of it, as a common theme, namely "Climate polarisation", "Hydrology risk" and "Floods". All of which, however, represent different aspects of the same theme. The remaining three, on the other hand, have themes that deviate from the main network theme, in particular the cluster on agriculture and the cluster on vegetation are particularly detached and distinct from the rest of the communities. Judging the quality of the result, the division into communities has paid off, identifying six clusters with distinct themes, which will provide useful insights on the papers. Analysing instead the distribution of nodes within communities, the number of nodes is gradually decreasing as can be seen from the first barplot in figure 24. The "Temeperature warning" cluster is the largest, with 699 nodes. The smallest is the "Vegetation" cluster, with 440 nodes. The largest cluster therefore has about 38% more nodes than the smallest. The other communities fall within this range, and their number, i.e. clusters 0,1,2, etc., is inversely proportional to their size.

5.3 Visualising the network

To effectively visualise both the network and the result of the community detection performed, a dedicated network visualisation software is used, namely Gephi [10]. Gephi is a open-source tool widely used for all kinds of network visualisation and exploration. It also offers community detection algorithms, however in our case the community division made with NetworkX is maintained, because it is more customisable, such as in the number of communities desired. Gephi however, from the point of view of graphical capability, is far superior to any other tool available in python, and far simpler and faster to implement. First of all, the datasets to be consumed in Gephi are prepared. The first concerns the nodes, which are associated in the form of a corpus id with the cluster number to which they belong. In addition, a label representing the entire cluster is assigned to the nodes with the highest degree in each community. If there are several nodes with the same degree, the choice is random. In fact, it is impossible to display the keywords of each individual node, for instance, as the result would be a black cloud. In this way it is possible to plot the name of each community within the community itself, as the nodes with the highest degree will most likely be in the centre of their respective communities. Then the edge dataset is constructed, where the corpus ids of the nodes at the end of an edge are associated, together with the weight of the edge itself, to more accurately represent the network. After that, the datasets are saved in csv and loaded into Gephi. The nodes are then linked to their edges and coloured according to the community they belong to. The labels of the nodes with the highest degree are shown, and the result can be appreciated in figure 26.

Force Atlas 2, used to visually construct this network, is an algorithm that can be used in Gephi and belongs to the class of Force directed graph layouts. The basic idea is that all nodes are assigned a repulsive force, which is inversely proportional to the distance between them, and an attractive force, which is directly proportional to the distance. Attractive and repulsive force depend on specific parameters that can be set, such as "gravity". Once both forces have been calculated for each pair of nodes, their position within the graph is updated in the direction of the resultant force. It is an iterative loop that already starts to have its limits at 3000 nodes, but it worked well in this case. The algorithm also takes into account the weight of the edges to determine the attractive force between two nodes. Observing the network, what can be deduced is that three similar clusters can be clearly seen, and the three more stand-alone clusters as well, as anticipated by analysing the most frequent keywords per cluster. The three communities "Floods", "Climate polarisation" and "Hydrology risk" are very interconnected, and are subgroups of the same cluster, which could be described as meteorological events more generally. In the drawn network, the size of the nodes represents their degree, thus their level of centrality and interconnection with other nodes in the network. It can thus be seen that the nodes with the highest degree are all in the centre, especially in the community of "Climate polarisation". On the other hand, looking at the more peripheral communities, one can see the strong interconnection of the Agriculture cluster and the Vegetation cluster, and how little they are connected to each other. The "Agriculture" cluster is also very dense, and where those nodes with a high degree that are not in the centre of the network are gathered, indicating a strong thematic connection on papers of this topic. The "Temperature warning" cluster, on the other hand, is more dispersed and less

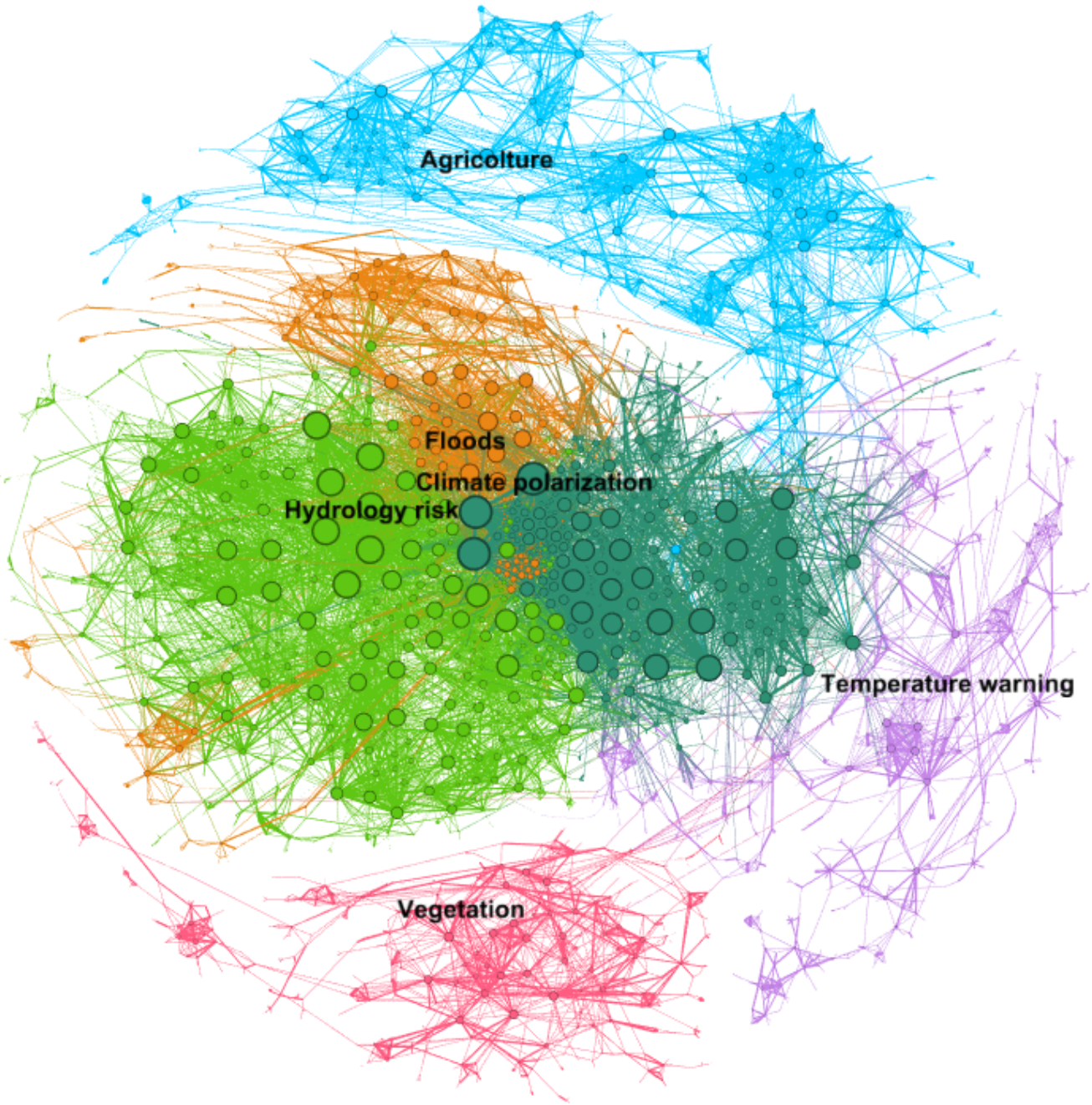


Figure 26: Representation of the network of papers by community. The visualisation algorithm used is the Force Atlas, implemented in Gephi. The size of the nodes represents their degree. Nodes coloured in the same way belong to the same community

clustered, strongly interconnected with the "Climate polarisation" cluster. It is interesting to note that the latter acts as a connector between all the different communities, indeed if the connections of the three outer clusters are observed, they are mainly created with "Climate polarisation", and less with its siblings. This indicates that there is generally more mention of "precipitation" in the papers, but that the theme of "drought", the more recurrent word in "Climate polarisation", is central.

6 Discussion

Thus, in a brief summary of what has been analysed so far, on the one hand we have the geographical distribution of the papers, or rather information on the locations discussed in the papers, while on the other there is a thematic information on the topics each paper deals with, and a division of them into clusters with common topics. The analysis therefore naturally turns to trying to extract as many conclusions as possible from the intersection of these two strands, which are parallel for the time being, but which will undoubtedly be the duty of this thesis to converge. A certainty of conclusions will not be provided, but an interpretation of the results, analysing the relationships existing between the themes and locations of the papers in question. However, since the starting dataset is poor in information that is not strictly related to abstracts, and since simply sticking to geographic and thematic clustering would certainly bring some results, but would not go in the direction of finding the answer to one of the main questions of this thesis, i.e. whether academic research is capable of addressing the actual territorial needs of the different countries, the data were enriched by taking two other sources into consideration. The first is ND-GAIN[10]'s Country Index data [15], which contains data on the "Vulnerability" of most countries around the globe to climate change, and their "Readiness", i.e. their preparedness to cope with and prevent change. The other is the World Bank's open data [2] on economic and social indicators, which also cover a large proportion of countries around the globe. How these data were obtained and processed will be discussed in more detail later. An important aspect to underline in order to make the interpretation and reading of the results more effective is the fact that, in the graphs, when information on the communities deriving from the network is present, only the 3,216 papers belonging to the main component, and on which the clustering was performed, were taken into account. Instead, in all other cases, all those papers with a geographical classification, published between 2011 and 2021, are taken into account.

First of all, so far the total number of papers in which a country is cited has been analysed, but no analysis has been made to consider the evolution of countries over time. As can be seen in figure 7, the number of published papers has increased exponentially in recent decades, but the geographical focus has not remained constant. As can be seen in figure 27, where the relative rate of appearance of each country in the abstracts is shown, compared to the other countries, considering only the nine countries that appear in the most abstracts, there is a relative reduction of the United States, which reduces its presence by about 10% from the 1990 to 2021, but also of the United Kingdom and Canada, in favour of the growth of the two Asian giants China and India, which have more than doubled their presence. As far as Italy is concerned, on the other hand, the percentage of presence remains more or less constant throughout the time span. Overall, the presence of European countries is constant. However, the nine countries chosen here, according to the criterion that they alone represent more than half of the papers, cannot represent the entire globe. The treshold was chosen to ensure the readability of the graph, further countries would have resulted in extremely small bars. However, to get a general idea of the presence of the various countries in the abstracts over time, an equal graph was created, but at the level of granularity of the continents, and can be observed in figure 28.

[10]ND GAIN: Notre Dame Global Adaptation Initiative

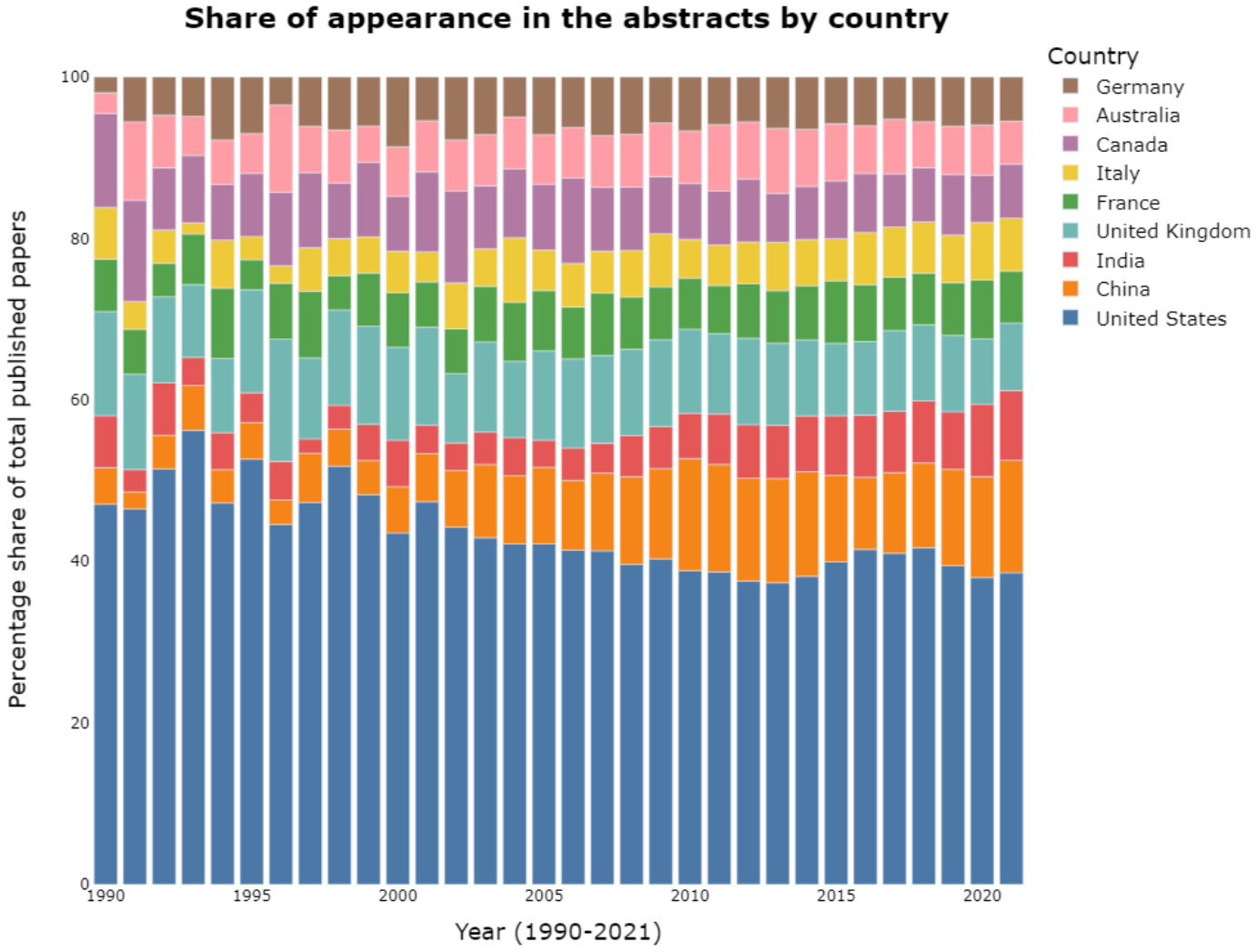


Figure 27: Representation of the relative percentage of countries presence in the abstracts, among the nine most present countries overall, from 1990 to 2021.

It can be seen how the increase in the presence of Asian countries, led by India and China, is confirmed, but also the decrease of Anglo-Saxon countries, visible both by the decrease in the presence of North America, but also Oceania, where the contibution of Australia and New Zealand is almost total, compared to the Indonesian countries. Again, this confirms that the presence of European countries remains constant. It is important to remark that the constancy in these two graphs does not imply that academic output has not increased over the years, indeed it has increased enormously everywhere, but it should be read as a comparison of different growth rates, rather than growth in an absolute sense.

As mentioned earlier, to enrich the interpretation of the results and provide a broader overview, data from two sources outside Semantic Scholar were used:

- ND-GAIN[10] Country Index [15]: provides information on the "Vulnerability" and "Readiness" of countries. As defined in the index [15], "Vulnerability measures a country's exposure, sen-

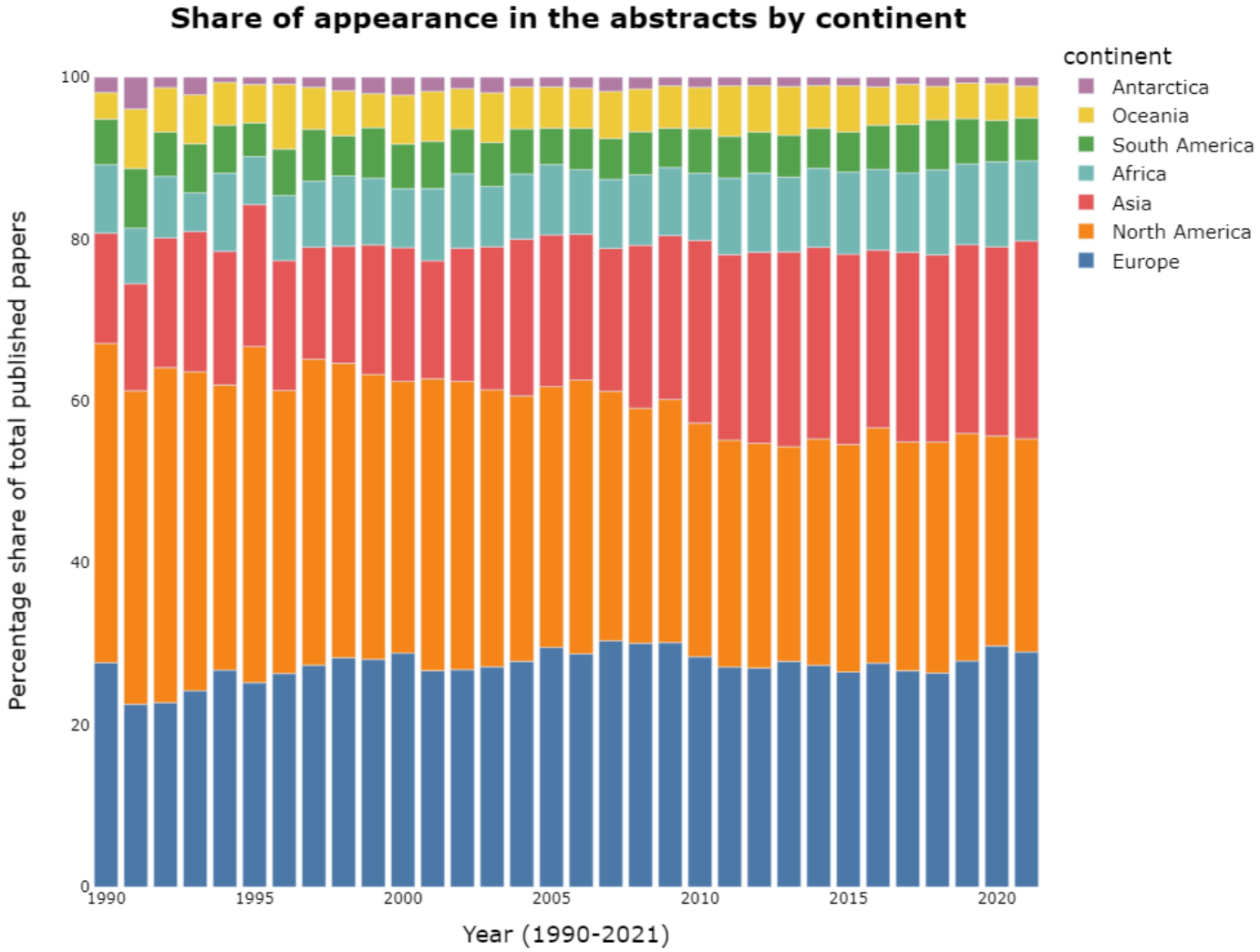


Figure 28: Representation of the relative percentage of continents presence in the abstracts from 1990 to 2021.

sitivity and capacity to adapt to the negative effects of climate change" while "Readiness measures a country's ability to leverage investments and convert them to adaptation actions." It is one of the few open-access indices that have comprehensive data regarding climate change on most countries. The index used in this thesis contains data from 1991 to 2021. Data from 2011 to 2021 were used because these are the years for which the network was built. "Readiness" and "Vulnerability" are independently constructed, both moving in the range between 0 and 1. "Readiness" is the better the closer to 1, while "Vulnerability" the opposite.

- World Bank Open Data [2]: the world bank offers lots of economic and social data, but not only that, at the country level. However, not all data are complete, so some potential climate change indicators on the platform were not included because they had temporally incomplete data or many missing values in various nations. A great plus of the platform is that it provides a usable API in Python, wbgapi [1], already integrated with the Pandas library, which allows

easy extraction and use of data. The following indicators were used:

1. CO_2 emissions (kt)
2. GDP (current US$)
3. Prevalence of severe food insecurity in the population (%)

Since, for reasons of both the quality of the results and computational limitations, only the years 2011 to 2021 were taken into account when constructing the network in the 5.1 section, since, as seen when exploring the data, in the section 3.3, after 2021, the number of papers on the Semantic Scholar platform and the metadata associated with them decreases, and since the clusters resulting from the network are crucial for the thematic analysis of papers, even in the case of the data downloaded from ND GAIN and World Bank, only those from 2011 to 2021 were considered. In addition, since with regard to the analysis it was necessary to have a single data representing these 10 years, in the case of both "Vulnerability" and "Readiness," as well as in the case of the World Bank indexes, the average of the indicator over these 10 years was taken for each country. Moreover, the same approach was also taken with regard to the number of papers in which a country is mentioned, taking the average number of papers over the 10 years from 2011 to 2021 for each country, so that a common baseline was defined for all the data used in the analysis. Taking into consideration the number of papers per country, it can be seen in figure 22 that there are few countries with many papers, and many countries with few papers. To make sure that the analysis is not distorted therefore by outliers, it has been represented, for the average number of papers per country between 2011 and 2021, their distribution, in figure 29. It can be seen that there is the absolute outlier, as expected, of the United States, with more than a thousand papers per year, on average, which can be seen in the top left graph of the figure 29. Looking at the top right graph, where the average number of papers has been trimmed to 200, it is possible to see that Africa, North America, and Oceania generally have a lower average number of papers published per year. In particular, looking at the violin plot at the bottom right, Africa has a much more squashed distribution on a low number of papers than all the other continents. There seems to be more equity within the continents in Asia, Europe, and South America. Finally, looking at the bottom left graph, what was initially said is confirmed, that almost 90 countries have an average number of papers citing them per year between 0 and 5, and that very few countries exceed 40 papers per year.

That most countries have a low number of papers per year is what also emerges from the analysis of the graph in figure 30 where the correlation between the average number of papers and and the average vulnerability was plotted. The scatterplots are configured in such a way that the first includes all countries, the second excludes the U.S., and the third excludes all nations with more than 100 papers on average, so as to gradually exclude the outliers identified in the figure 29. In all three graphs the point size is given by the GDP of the countries, but in relation to the other countries in the graph. Indeed, the countries in the last graph are the same as those in the first, excluding the outliers, it is just that the GDP is rescaled precisely because those countries are no longer present. A simple regression line is also shown in the graphs, representing the correlation between number of papers

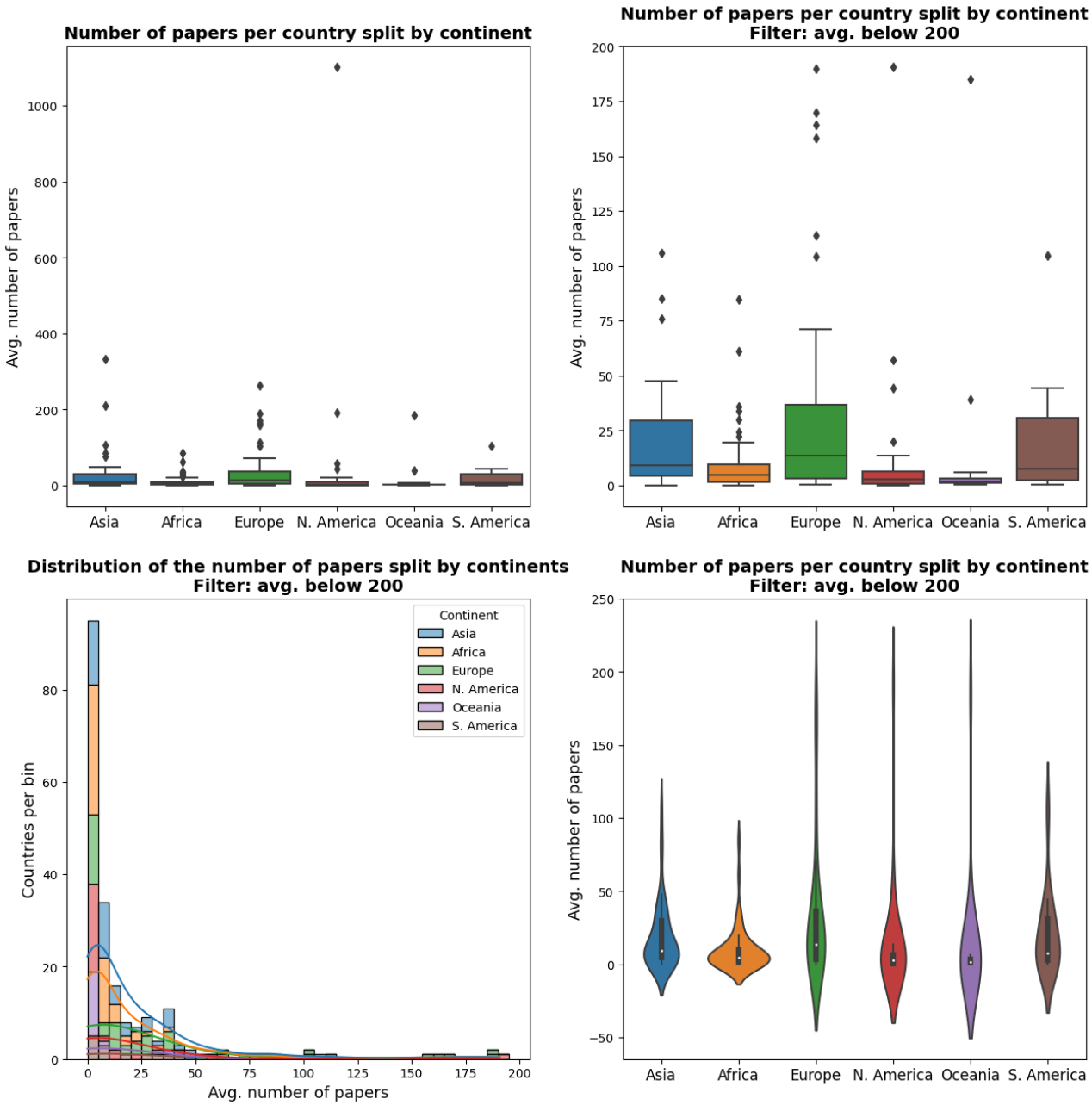


Figure 29: Representation of the distribution of the average number of papers per year in which a country is mentioned, between 2011 and 2021.

and vulnerability. As can be seen, again in figure 30, is that the U.S. tends to make the relationship flatter, because it is present in a huge number of papers compared to the others, which weighs heavily on the correlation. However, it can be seen that even when the outliers are removed, the correlation does not change that much, remaining negative, indicating that as the number of papers in which a country is mentioned increases, the country's vulnerability decreases. It also emerges from the figure, especially looking in the last graph, that a large proportion of the countries with high vulnerability are African or belonging to Oceania, followed by Asia. European countries, on the other hand, have lower levels of vulnerability on average. Within the continent groups, however, it can be seen that the

pattern repeats that as the number of papers increases, vulnerability decreases, with a few exceptions. It is important to establish that this does not imply a causal link between the two factors, not least because surely a great many other factors come into play, of which only GDP is tracked here. But even looking only at the size of bubbles, we see that on average, larger bubbles have lower levels of vulnerability, surely indicated that the economic factor has its weight in a nation's vulnerability. However, it is not the purpose of this thesis to analyze what macroeconomic factors lead to a country's climate vulnerability, but the results of these correlations are useful in interpreting the overall results.

In figure 31 what was done in 30 was replicated but with readiness instead of vulnerability. As might be expected, the results are virtually symmetrical, although it is important to note that the two measures are not the reciprocal of each other, but are simply calculated differently. Again, African countries are the least "ready" for change, and more economically developed countries are more so. The correlation with the average number of papers in this case is positive, but upon removal of outliers it tends to flatten out, unlike that of vulnerability. Again, European countries have higher readiness scores. These results confirm what Klingelhöfer et al. [20] concluded in the paper discussed in the 2.1 section.

To better interpret and provide additional inisghts on the graphs above, in figure 32 the average vulnerability of each country has been depicted, where as vulnerability increases, the area of the rectangle increases. As highlighted in the figures 30 and 31, the countries with the highest vulnerability reside in Africa and Oceania; where especially the smaller islands are very vulnerable, just think of the threat posed by rising sea levels. Asia also has high average levels of vulnerability, while Europe has the lowest.

Moving forward with the analysis, although readiness will be taken into account, efforts will be focused on vulnerability, as it better represents the fragility of a country, and sees the problem from the perspective of suffering, and not of the action taken toward change. In fact, there are geographic factors that do not depend on preparedness, on human action, that make, just based on its location, a country more or less vulnerable. It goes without saying that then how vulnerability is addressed is critical to being able to prevent the disastrous effects of climate change, but the aim of this thesis is to analyze whether academic research is symmetrical with precisely these vulnerabilities, that is, whether it succeeds in meeting the demand for how climate change can best be addressed, even for those countries that are more economically disadvantaged, that therefore have fewer opportunities to invest in research, and perhaps more immediate concerns to be addressed than climate-derived threats. In fact, as is already emerging, it is precisely the economically disadvantaged countries that are at greatest risk.

The next objective is to converge the two hitherto parallel analyses, i.e., geo-spatial and thematic, together with vulnerability and readiness data from the various countries. To do this, the information is compiled into a single figure representing the six communities identified through the construction of the network in the 5.2 section. To join the three different datasets, data on the cluster to which each paper belongs was added thanks to the corpus id, used as a key, to the main dataset constructed thanks to the NLP[5] in section 4, already containing information on keywords and geographic locations. In addition, thanks to the ISO 3166-1 alpha-3 classification [37] obtained and stored during the

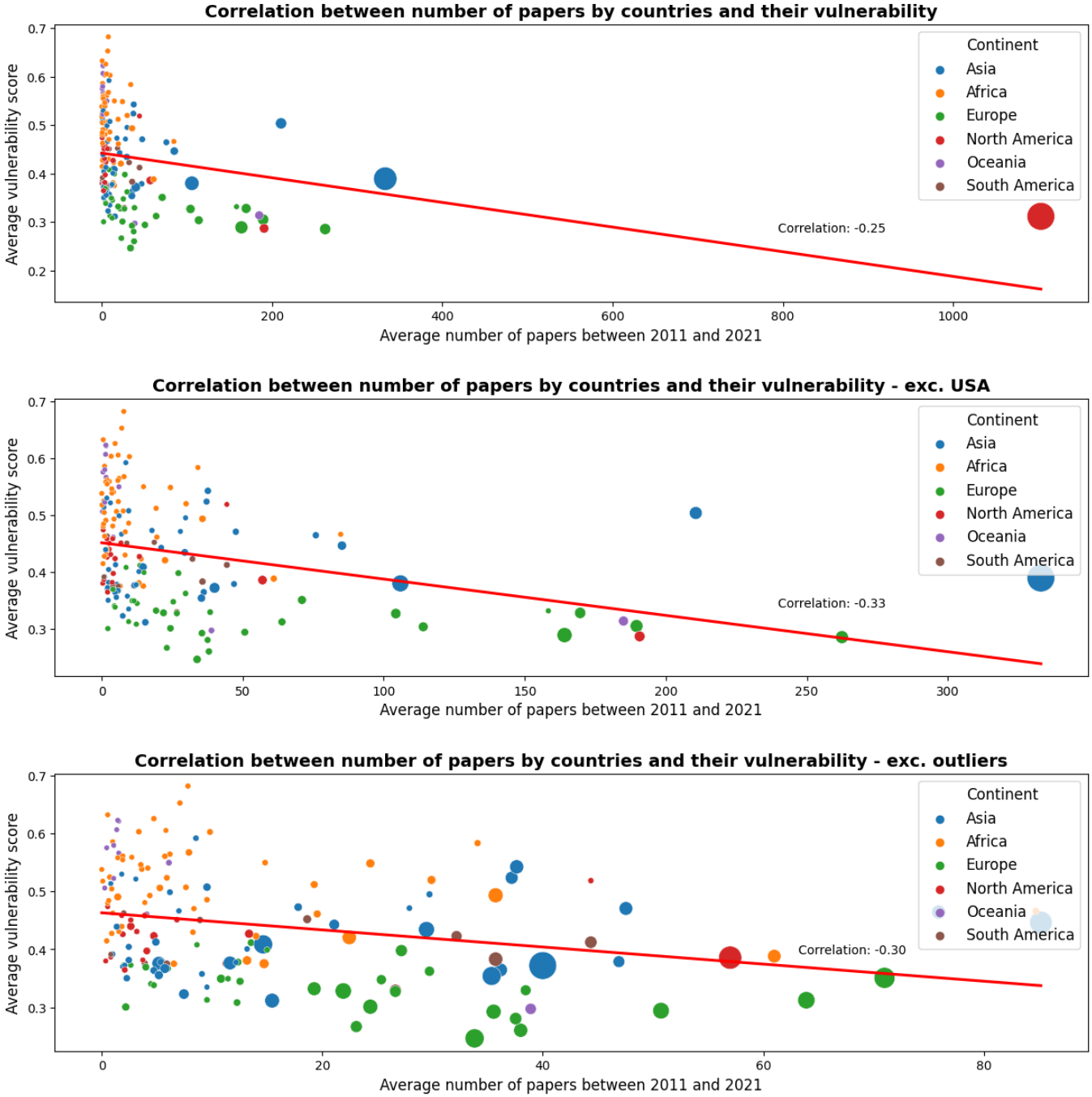


Figure 30: Correlation between average number of papers and average vulnerability by country in the time span between 2011 and 2021. The size of the bubbles represents the country's GDP weight relative to other countries in each graph. The first graph includes all nations, the second excludes the U.S. outlier, and the third excludes all countries with an average number of papers greater than or equal to 100.

extraction of geographic data, which is also present in the ND GAIN[10] [15] database of vulnerability and readiness, it was also possible to incorporate the latter information for each country for which it is available. The highest level of granularity available then, from here on, will be, at the geographic level, that of nations. Once the information has been combined into a single dataset, it is then possible to represent potential patterns existing in the vulnerability and readiness of particular communities.

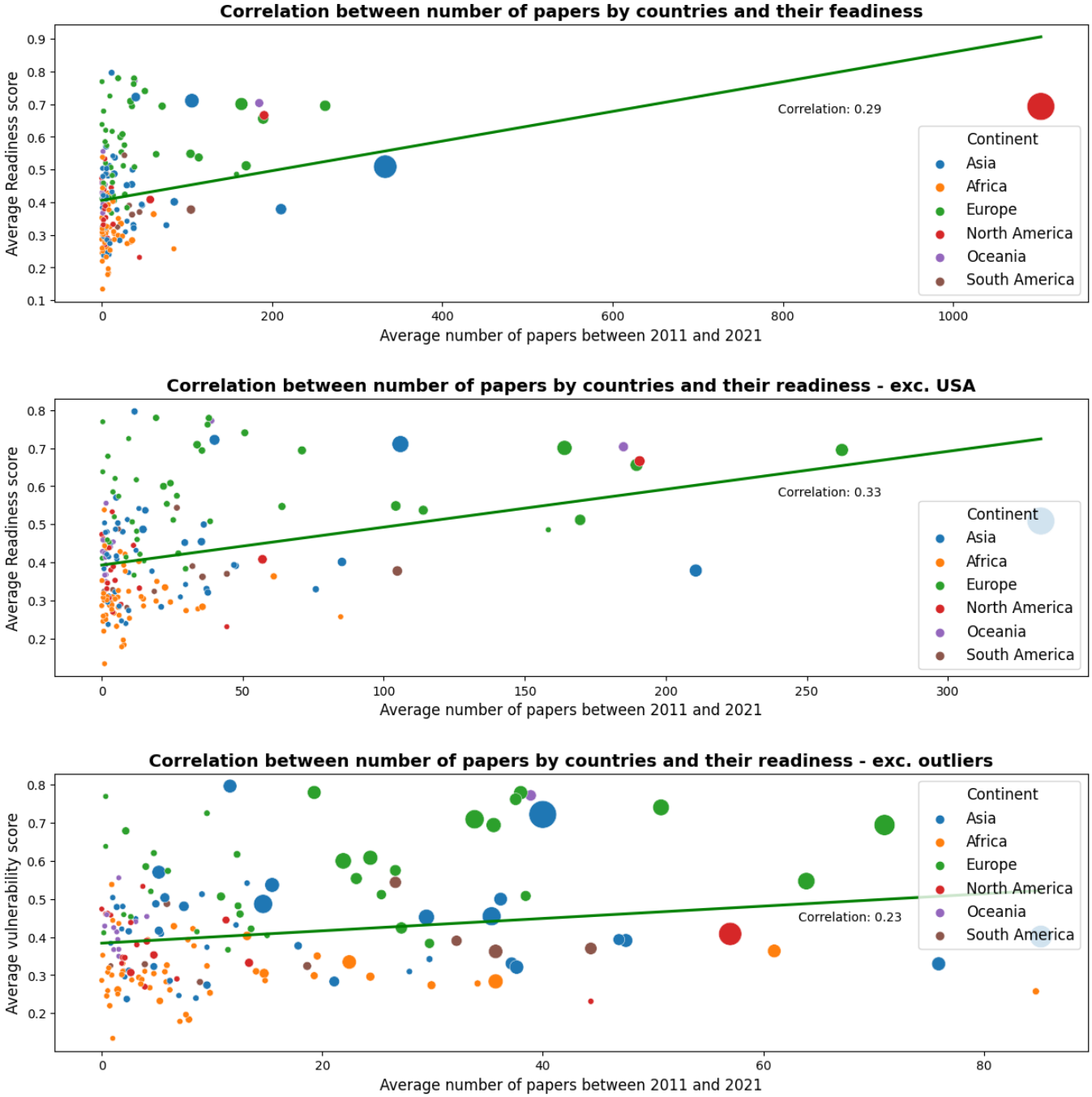


Figure 31: Correlation between average number of papers and average readiness by country in the time span between 2011 and 2021. The size of the bubbles represents the country's GDP weight relative to other countries in each graph. The first graph includes all nations, the second excludes the U.S. outlier, and the third excludes all countries with an average number of papers greater than or equal to 100.

Figure 25 has already represented the most frequent keywords for each cluster, based on which a name was found to be the common denominator of the overall cluster theme. Figure 33 again presents the same 6 clusters, with the names representing the previously identified thematic area, but, instead of the keywords, the nations most frequently mentioned in the abstracts of papers belonging to that specific community are represented in the barcharts. It should be recalled how all papers in the network

Figure 32: Treemap representing the vulnerability level, given by the size of the rectangle plane, of each country for which vulnerability data are present.

have geographic information because, with the aim of reducing the number of papers in the graph, for computational reasons, all papers without such information had been excluded, as explained in the 5.1 section. In addition, the United States was excluded from each barchart because it is the most recurring country, by a long margin, in each of the clusters, repeating the pattern seen in the figure 29, in the boxplot at the top left. Although their inclusion would not have created too many visualisation problems from a barchart point of view, as it would simply have made the bars of the other countries smaller, and thus in proportion to the difference with the USA would have given the impression that they were closer together, from a vulnerability and readiness point of view it would have flattened the results, leading all the clusters to look a little more similar. Indeed, the second piece of information hanging in the present graph, in two gauge plots, is about the weighted average of vulnerability and readiness for each cluster. Weighted average because the algorithm takes into account how many times a country is present in the cluster, so the vulnerability weight of the more present countries will be greater than the less present ones. Computationally, it is very simple because the dataset has been constructed in such a way as to have one row for each time a country appears in a community, because there is the detail of the corpus id, which is the primary key of the dataset. So there will be as many rows in the dataset as many times a country appears, and therefore its vulnerability and readiness will be repeated as well. It is therefore sufficient in this case to make a simple average by taking all the rows into account. Countries for which the value of vulnerability and readiness is missing are excluded from the average, but not from the barcharts, as they are marginal anyway, and do not stand out among the countries most present. Anyway, there are 164 countries present in the clusters in total, of which only 7 have no information on vulnerability and readiness, namely: Bermuda, Cook Island, Greenland, Guernsey, Palestine, South Sudan and Taiwan. The gauge plots for vulnerability and readiness are constructed to show the range between the maximum and minimum value reached in general by all clusters, i.e. between 0.35 and 0.43 for vulnerability, and between 0.44 and 0.55 for readiness. That the two measures do not move in the same range is not a problem, since it is not beyond the scope of this thesis to compare vulnerability and readiness values with each other, but rather it is to compare the scores for each cluster with each other, which then lie on the same scale
The information that can be extracted from the figure 33 therefore goes in two directions. On the one hand, it is possible to extract clues as to what the countries' greatest thematic urgencies are, and on the other hand, it enables a link to be made between the vulnerability and readiness of countries both with the theme and with the countries themselves, identifying clusters that share the same climate challenges. What immediately catches the attention is that two diametrically opposed clusters can be seen. Indeed, community 2, the one with agriculture as its main theme, shows the lowest levels of readiness and the highest levels of vulnerability, while community 5, which revolves around vegetation, shows the highest readiness and lowest vulnerability. Looking at the countries that make up these clusters the most:

- in cluster 2 there is a majority of African countries, which are recurrent in this cluster, but tend to be present in few papers considering the total figure. The incidence is therefore not only high, but it is also surprising that these countries emerge despite the fact that they are

scarcely mentioned in general, indicating their high concentration around food issues. In fact, in the figure 25 we see that the most recurrent word is agriculture, followed by "crop", and "farm/farmer", indicating that the concern is indeed agricultural, but centred on the issue of the impacts that change may have on crops, and therefore on food supply. The two most recurring countries, however, are India and China, which, however, compared to the African nations, participate more significantly in the number of papers citing them, as evidenced by the fact that they are consistently present in the majority of clusters. It is important to note that the country most present, India, has a pretty high vulnerability of 0.5 per se, which therefore certainly contributes, along with the others, to the classification of this community as the most vulnerable.

- In cluster 5 there is an almost total majority of European countries, and it is the only cluster not dominated by China or India, the latter not even present. Notable is the presence of Belarus, which is, however, one of the most mentioned countries in Europe, as can be seen from the Voronoi treemap in figure 22. Its vulnerability is 0.33, and this is a particularly low vulnerability, as it is generally low among European countries, as can be seen in figure 32. If we look at the words in this community, we notice, among others, in figure 25, the words "tree", "drought", "forest", "wildfire" and "ecosystem", which lead to draw a situation in which forest ecosystems are threatened, including by drought that leads to more and more frequent wildfires, as constantly seen by the negative records that each new summer brings. Europe is at the centre of this cluster probably for two reasons: on the one hand, the massive presence of forest ecosystems in its territory, and on the other hand, probably a particularly specialised research on the topic, or at least with the possibility of being oriented towards topics not directly related to the destructive effects that climate change can have, such as those related to flooding, or not strictly related to the potentially devastating consequences that burden the population, such as food insecurity.

As for the rest of the communities, the three clusters that make up the central core of the network in figure 26 differ for two reasons: the first is that the situation of vulnerability and readiness tends to worsen as moving towards the nodes belonging to climate polarisation and temperature warnings, and the second is because of the more frequent countries, where India is seen to be mentioned more in the cluster characterised by more violent words related to rainfall, such as "disaster" and "flood", i.e. cluster 3, and the one where the issue of drought is also central, i.e. cluster 4, than in cluster 2. Looking at the total number of papers per community, which can be seen in the first graph in figure 24, it emerges that the cluster with the lowest vulnerability and highest readiness is the one that has received the least attention from academic research while cluster 2 on agriculture is in third position. The first cluster, the most numerous, has an average level of vulnerability and readiness, in fact it is also the one that connects all the other clusters, as can be seen in figure 26. Taking the three similar clusters into account, the trend here is the reverse, i.e. the cluster with the best vulnerability is the one that received the most mentions. In general, it can be concluded that academic research has been more or less equally focused, with more attention being paid to those countries with higher vulnerability and less readiness, although it should be noted that the United States is excluded from this analysis.

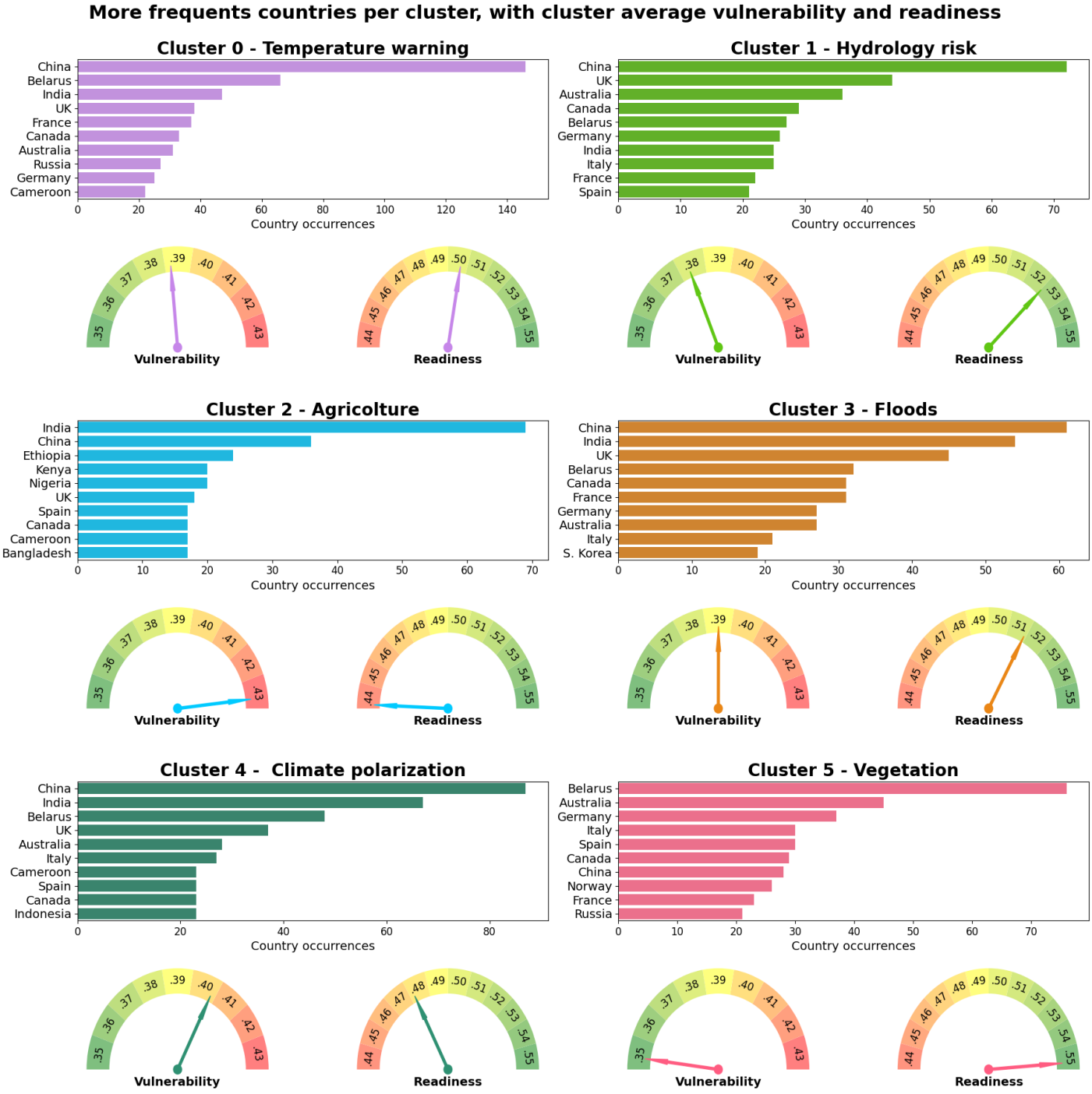


Figure 33: Representation of the communities identified in the 5.2 section with the countries most frequently mentioned in the papers and the weighted average vulnerability and readiness of each cluster. Excluded are the united states, which are the most frequent countries in each cluster, and thus would have simply flattened the results without providing valuable insights.

It thus emerges that the most disadvantaged countries from a climate change adaptation point of view have a particular focus on issues with potentially more direct harm to humans, such as those related to food security. As shown in figure 30 and 31, vulnerability is inversely proportional to the average number of published papers, while readiness the opposite. The fact that cluster 2 is the third largest in terms of number of papers, but that its vulnerability is very high, and that it is largely com-

posed of African and Asian countries, inidicates that the majority of the countries that compose it are those at the top left of the vulnerability scatter plot, i.e. countries that tend to be economically disadvantaged or developing countries, and thus characterised by a higher average share of the population dedicated to the primary sector. To highlight and better interpret the geographical distribution of the clusters, an atlas was constructed in which each country was coloured according to the cluster most frequently associated with the papers in which the country is mentioned. Looking at the figure 34, it is possible to confirm what has been said so far. Countries with agriculture and crops as their main cluster are also characterised by economic underdevelopment and a large proportion of the population employed in the agricultural sector. An exception from an economic point of view may be India, which is in a more advanced phase of transition, but still had 44% of its population employed in the primary sector in 2021 [2]. It can also be seen that the cluster on climate polarisation, i.e. highlighting either a focus on drought, or floods, or both, is also particularly present in countries with economies that are not fully developed, and in desert or very arid areas, such as the Saharan or Middle East. Most of the countries that have vegetation as their most recurring cluster are, as already mentioned, European, including Italy. Another fact that can be analysed is that the Indonesian, as well as the South American, and much of sub-Saharan Africa tropical zones revolve around the three clusters related to hydrogeological risk, flooding and climate polarisation, which are, as seen in the previous chapter, that are very interconnected. Furthermore, it seems that geographically neighbouring countries also tend to have a certain recurrence in the most frequent cluster in which they are mentioned, as can be seen in the blue of the south-eastern part of Africa, the pink in central Europe or the purple covering North America, China and Russia, as can be seen in more or less the whole atlas.

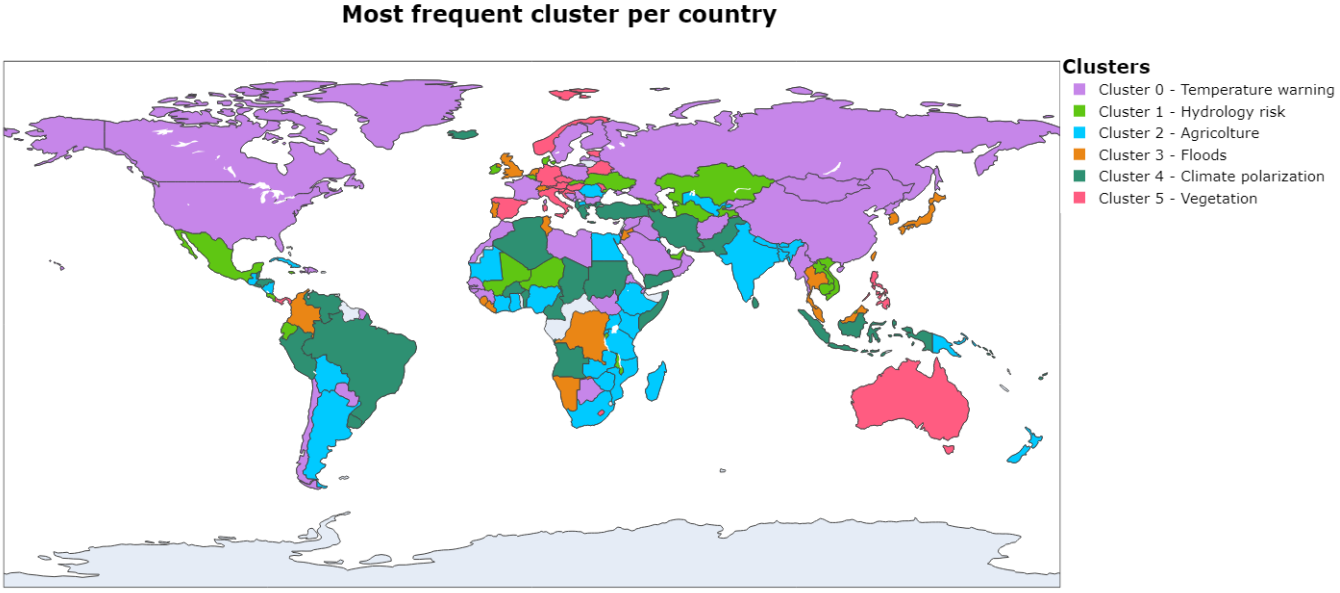


Figure 34: Representation of the most frequent community for each nation. In the atlas, each country is coloured in line with the cluster with which it appears most frequently associated

The most striking fact, however, is the massive presence of the theme of food and agriculture primarily in developing countries. Even if one looks at the results of the 4.3.1 section, in particu-

lar the 19 graph, one can see that "Agriculture" is the word that most recurs in abstracts related to the African continent. This aspect is worthy of further analysis. In particular, the figure 35 shows the correlation between the percentage of times that a nation is associated with a paper classified in cluster 2 out of the total number of appearances of that nation, and the percentage of the population living below the threshold of food security, according to the data of the World Bank [1]. The size of the bubbles, in this case, represents not GDP but the average number of papers in which the country is cited. From the scatter plot, it can be seen that there is a positive correlation, i.e. as the percentage of the population living in a food insecurity situation increases, so does the country's relative presence in the agriculture-related cluster. As might be expected from the results so far, most of the countries with high levels of food insecurity are African, and indeed their weight in community 2 is high. Furthermore, a further pattern can be seen, namely that countries with a high presence in this cluster, and with high food insecurity, are in general the least mentioned overall in the papers. This is obviously strongly linked to economic reasons, but it shows that research focuses less on those countries that are most economically disadvantaged, but that when it does, it goes to the needs of that country. In fact, food insecurity is certainly threatened by climate change, where without adequate adaptation measures the situation is bound to worsen as droughts and extreme weather events, destructive to crops, increase. Indeed, it can be seen that a large proportion of the citations in the abstracts are concentrated in the lower left-hand corner, where the most developed countries, European and North American, have very low levels of food insecurity but equally low levels of presence in the "agriculture" cluster. The disposition of countries is similar to what can be seen in figure 30, where a stratification of continents in terms of vulnerability can be seen. On average, African countries are more vulnerable, followed by Asian and North American countries, and, finally, Europeans. Indeed, looking at the scatter plot in figure 35, the same stratification can be found inidicating that, given the high vulnerability of the countries with the highest food insecurity, the need to take the necessary mitigation and adaptation actions, and thus the need for substantial academic input to be able to find targeted solutions, is crucial, especially in countries where change threatens the population in its most basic needs, often a matter of survival. Furthermore, as already addressed in the 2 section, not only, confirming these data and citing [40], is there inequality in the impact of climate change between countries, but also, citing [16], within the same countries, the most disadvantaged segments of the population absorb a large part of the damage related to climate change and extreme weather events. It is therefore immediate to think of the need for the international community not to leave behind those underprivileged countries, and their populations, with up to 40/50% of the population experiencing food insecurity. The role of academic research can be crucial in finding adaptation solutions tailored to the specific needs of these countries.

The last analysis proposed in this section analyses precisely the differences between countries according to their vulnerability. From a methodological point of view, countries were divided into roughly equal classes according to their average levels of vulnerability between 2011 and 2021. In particular:

1. first class with vulnerability between 0 and 0.36, in which 41 countries fall.

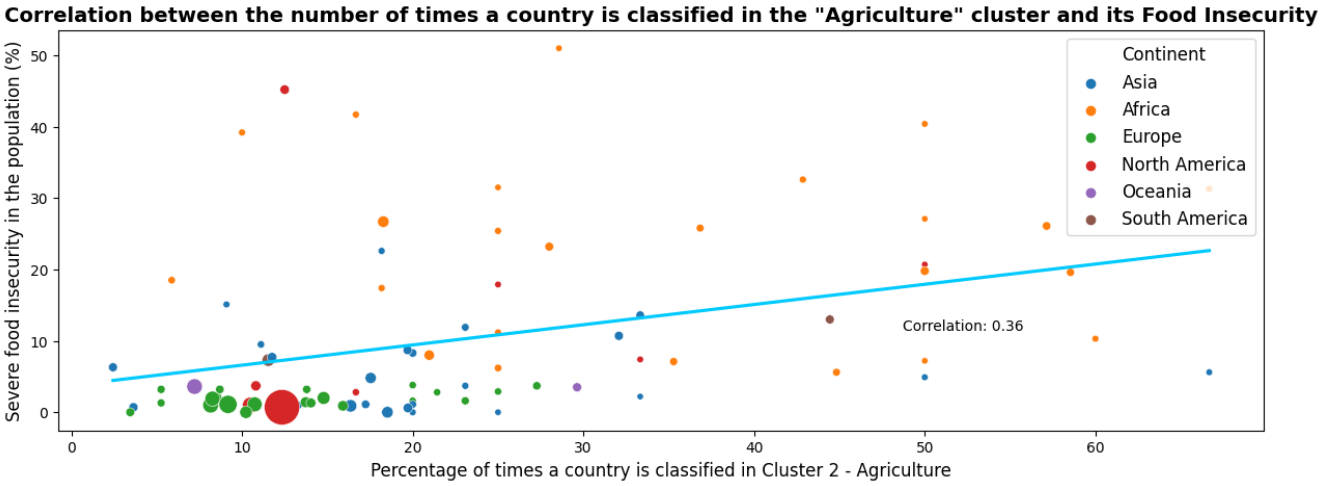


Figure 35: Correlation between the percentage of times that a country is associated with a paper belonging to Cluster 2, out of the total number of times it is associated with a cluster, and the percentage of the population below the threshold of severe Food Insecurities [1]. The size of the bubbles represents the average number of papers in which that country is cited.

2. second class with vulnerability between 0.36 and 0.42, in which 48 countries fall.

3. third class with vulnerability between 0.42 and 0.51, in which 52 countries fall.

4. fourth class with vulnerability between 0.51 and 1, in which 44 countries fall.

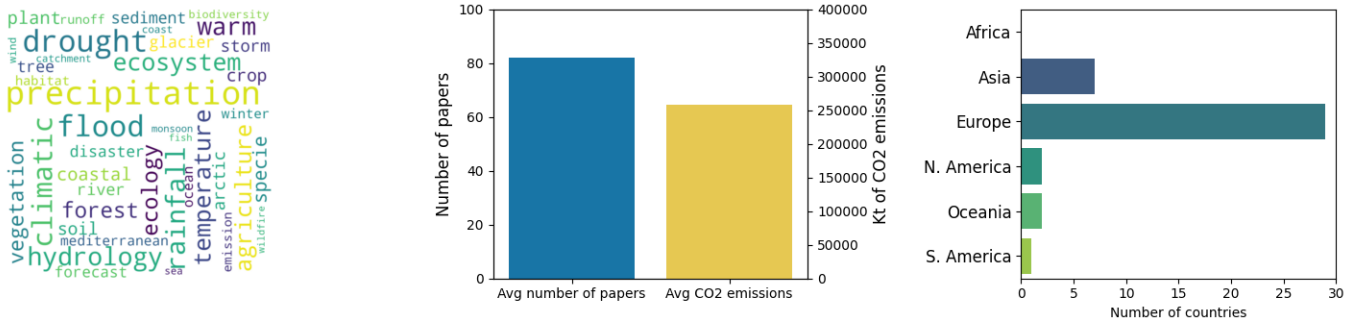
In each class, its upper limit is included, and its lower limit is excluded. On average, there are 46 countries per class; it was decided to leave the first and last classes slightly smaller in order to better capture characteristics of the more extreme countries on this scale.

Three graphs are plotted for each class, as it is possible to see in figure 36. The first is a wordcloud containing the 40 most frequent words, whose size varies depending on the number of times they appear in the class. The second graph is a barplot providing information on two different aspects: on the left is the number of citing papers in a year on average by the countries belonging to the class, the bar refers to the left axis, while on the right is the kt of CO_2 emitted in a year on average by the countries, in this case the bar refers to the right axis. Finally, there is a bar chart that collects the number of countries per continent that belong to the class. The first thing that clearly emerges is that the countries in the most advantaged class from the point of view of vulnerability are also those for which, on average, more academic research is concentrated, and not by a little. In the first class, in fact, there are mainly European countries, with around 80 papers on average among the countries citing them per year, while already in the second class it drops to just over 20, and then gradually decreasing. The other thing to note is that no African countries are present in the countries with lower vulnerability. There are about 30 European countries, followed by seven from Asia, and very few from other continents. The other pattern observed is also the decrease in emissions on average as vulnerability increases. In this case, however, the second class, probably due to the presence of many Asian countries, emits more tonnes of CO^2 into the atmosphere on average. It should be noted that this class of countries emits the most CO^2 of all, but is very little in the focus of research, compared

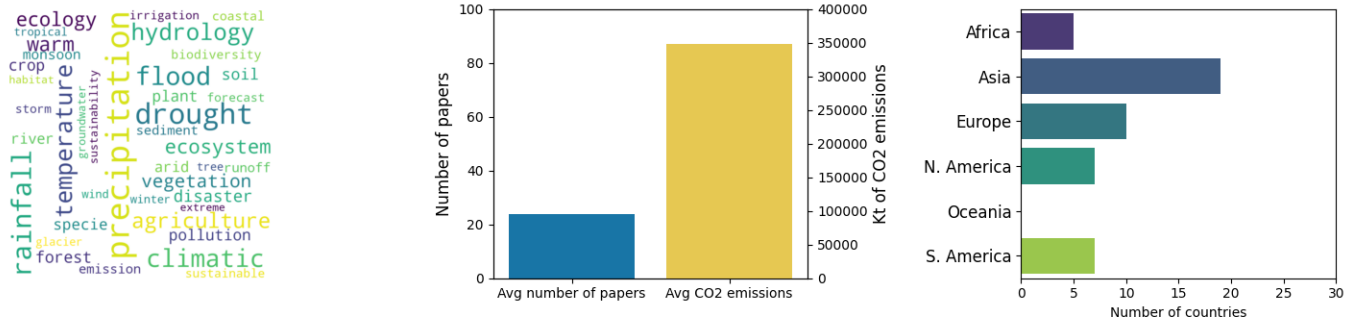
to the countries in the first class. The other thing one notices is that countries in the last two classes, which account for half of the total countries considered in this research, are mentioned very little in the abstracts, that they contribute equally little to CO^2 emissions into the atmosphere, but that they are the ones that will suffer most from climate change. In practice, they have contributed little to the causes of change, are not particularly helped in mitigation and adaptation activities by the academic community, and will be those most exposed to the disastrous effects of climate change. In addition, none of the European countries are included in the most vulnerable classes, on the other hand, most of the countries in the most vulnerable class are African, and most of the Oceania/Indonesia countries are concentrated in the last class. North America is also highly represented in the penultimate class, due to the many small states near Mexico, many of them islands threatened by rising seas and violent climatic events. What clearly emerges from this picture, and which confirms what we have seen above, is that the countries with the lowest risk of vulnerability are the ones that are most discussed in academic research, and are the ones that are most responsible for their emissions, while the most vulnerable countries seem to be left on the margins. This mirrors any other macroeconomic indicator in its distribution around the globe, from a research point of view, however, it can only be positive if research focuses on mitigating the effects of climate change, i.e. precisely on reducing emissions of environmentally harmful gases. On the other hand, however, academic research, along with certainly other social and economic factors, has contributed to making these countries less vulnerable, so it is to be hoped that the same attention will be paid to the most disadvantaged countries in the future, along with the necessary investments to put together the necessary adaptation measures. In fact, as pointed out in 2, climate change is already happening, and we are now suffering the effects of the gases emitted into the atmosphere 30 years ago, so from a mitigation point of view, it is already too late to help highly vulnerable countries, the only alternative is to help them prepare for these changes.

The advantage of the analysis so far is that it also provides insight into the thematic focus of the research, so that it is not merely a quantitative evaluation. On the left column of the figure 36, it is possible to compare the most common words of each class. The first thing notable is that, moving towards classes with higher vulnerability, there is a shift from a situation where there is a certain degree of equity in the number of times the most common words are repeated, with the word "precipitation" being the most numerous, but remember that it is the most numerous in general taking all the papers together, to a situation where gradually there are fewer words that prevail over the rest in terms of recurrences. This in part surely stems from the fact that the number of papers in the classes with lower vulnerability is certainly higher, so it also contributes to an increase in the diversity of topics addressed. Another thing that can be noticed is the gradual increase, as expected, of the word "agriculture" moving towards the more vulnerable class, but interestingly, the word "drought" also becomes relatively more present in the more vulnerable countries. Another thing worth noting is the presence of the word poverty in the wordcloud of the last class, confirming how the concern about agriculture is linked not only to a natural and environmental fact, but also to the difficulty in food procurement that large groups of the populations of vulnerable countries face, and are likely to face more and more because of climate change.

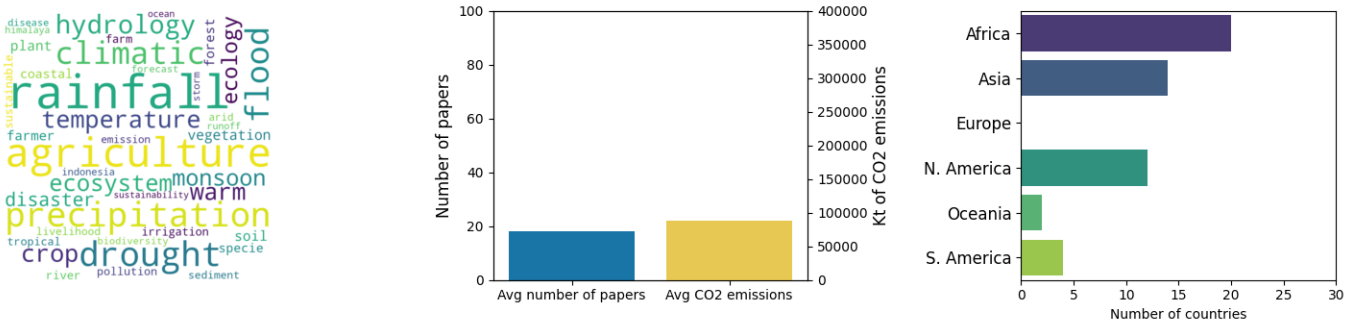
Class of countries with a Vulnerability Score of less than or equal to 0.36



Class of countries with a Vulnerability Score between 0.36 and 0.42 (included)



Class of countries with a Vulnerability Score between 0.42 and 0.51 (included)



Class of countries with a Vulnerability Score greater than 0.51

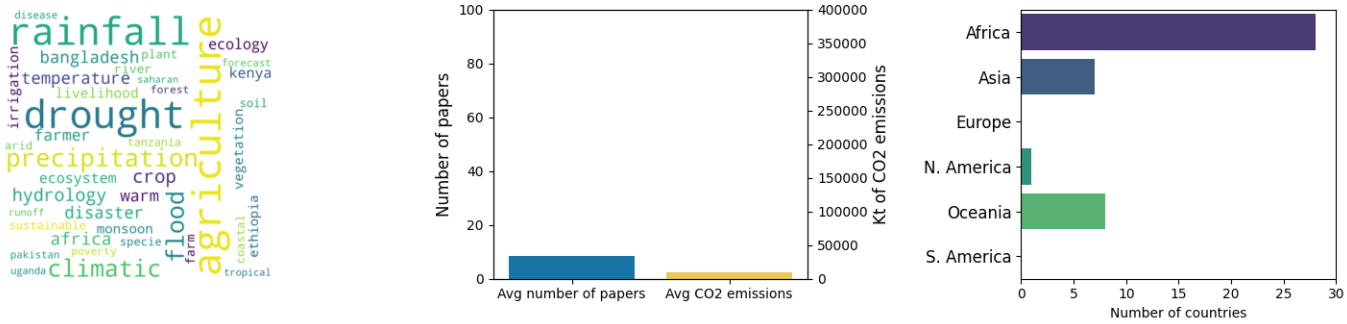


Figure 36: For each vulnerability class, the 40 most frequent words, the average number of papers and CO_2 emissions produced in the 10 years from 2011 to 2021 among the countries, and the countries most present per continent in the class are shown.

7 Open Points

There are numerous open points, or rather, further questions, arising from the analysis of this thesis, and numerous possibilities for improvement. Firstly, as already pointed out in the 3.4 section, the data on the papers come from a single source, thus making the entire analysis a child of the logics used for the construction of the Semantic Scholar logics. Expanding the source of the papers could lead to a more comprehensive view of academic research efforts globally, although it would create quite a few problems in harmonising data from different sources. Secondly, from the point of view of data processing and organisation, the limitation that can be circumvented is that related to the limitations of memory and processor power of a laptop, creating openings to be able to compare research on climate change and extreme climate events with research in general, something that was only possible with regard to the number of papers published over the years, but would have been impossible to do with regard to the processing of abstracts or taking other metadata into account. Computational limitations also created another barrier that could not be overcome, namely harnessing the power of LLM[6]. to analyse abstracts and extract information other than just keywords, perhaps trying to synthesise topics to make a direct classification of the paper, broader than the categories into which papers are directly divided by Semantic Scholar, and more focused on climate change. For example, papers falling under the category "Environmental Science" could have been broken up into sub-categories, or how papers categorised as "Medicine" were climate related. The keyword extraction method was also a compromise between quality and efficiency, and can certainly be improved and make the whole analysis more accurate. Another aspect that is successful as a method and in terms of the quality of results, but which would certainly have benefited from greater processing capacity, is that of network construction and the division of its nodes into communities. Indeed, the papers analysed in relation to the network, i.e. each time the identified communities were involved, numbered only 3 216, which is a good number but very slim compared to the total number of papers available. Therefore, it would have been possible not only to involve more years in the analysis, but also to carry out an analysis of the evolution of the network over the years. In general, the aspect of the thematic evolution of the papers over the years has not been particularly deepened in the course of this thesis, because the focus has been more on taking a snapshot of the situation of academic research to date, but the temporal aspect is certainly something that would be interesting to explore in more depth, in order to be able to grasp further nuances.

Another major open point that would lead to a highly interesting comparison, would be to be able to obtain information on the place of publication of the papers, even if this would mean cross-referencing the data with databases of different origins, in order to be able to compare them with the geographical data extracted from the abstracts, which would lead to a whole series of possible analyses to be made. First of all, the impact of academic cooperation between different countries could be assessed, and how much the country being researched is also the one where the paper is published, or not. Analyses could then be carried out on how macroeconomic indicators influence a nation's academic productivity, and whether countries that are advantaged in terms of vulnerability share knowledge with those that are either more vulnerable or have a weaker academic infrastructure. Moreover, there are data

in the dataset of this thesis that have not been used, but one could also direct the search by considering citation numbers and author information, perhaps with the construction of additional networks, to study the relationship between papers from a more technically academic point of view, but which could provide insights into international cooperation. Another interesting aspect that could be taken into account concerns the publication journal of papers, in order to identify virtuosities related to knowledge sharing on climate change mitigation and adaptation measures.

Moreover, correlation is often mentioned throughout this thesis, but from a strictly statistical point of view, correlation does not imply causation. In fact, surely the correlations that emerge between the number of papers and vulnerability, in the recurrence of cluster 2, and so on, underlie social and economic factors that are only scratched from afar. It is certain that countries with more economic means can make greater investments in research, and thus are inevitably more productive from an academic point of view, and that a country's vulnerability is immensely influenced by its political and economic history. It is not the aim of this thesis to give conclusions on causal links between the number of published papers and a country's vulnerability, but rather to give an overall view of the situation, providing an interpretation, albeit not a strictly statistical one. Indeed, multiple analyses can be carried out to assess the impact of one economic factor rather than another, such as how much of a country's GDP is spent on climate change research. However, for an analysis of this type, this thesis can be an effective starting point, thanks to the dataset, the extracted features, and the evidence already brought to the surface.

It is often mentioned that one wants to identify whether research meets the demand of countries from a climate change perspective, but this is done from a high-level point of view. A whole range of research on the current state of the art can therefore be carried out in order to understand one thing that does not emerge from this analysis: that is, what the countries' needs actually are, and on which they are facing the greatest climate difficulties. This is beyond the scope of this thesis, but, again, it could be a starting point.

Finally, related to what has been written above, an accurate representation of the distribution of keywords at country level has not been made, but the analysis has always been kept more aggregated at continent level, in order to make it more readable. Doing such an analysis, however, together with an identification of the actual needs and sufferings of each country, is among the possible future developments.

8 Conclusion

Attention to climate change concerns is increasing across all disciplines. There is more discussion in the media, there is increasing social attention, with protests spreading across the globe. Nonetheless, scientific research is increasingly focusing on the topic of climate change, its causes and effects, and impacts spanning many different areas. The growth in the number of papers on climate change, when compared to the growth in the number of papers published in general, has been greater, especially in recent years. This thesis follows the path traced by the literature, but also undertakes new strands that provide important results. A global focus on precipitation issues, and on runoff in general, is shown. Hydrogeological risk is a constant that is highlighted and seems to be the main theme addressed by the literature with regard to risks from extreme climate events. There seem to be two kinds of constants among climate effects, the first being that in the northern hemisphere of the planet there are certain recurring patterns in the themes, but these places also receive more attention from the scientific community, while the southern hemisphere has more varied themes, and yet is often less mentioned in abstracts. The second is related to macroeconomic indicators, i.e. in developing countries the levels of vulnerability are extremely higher, but lower is the recurrence with which these places are mentioned in the papers. This research shows how the most disadvantaged countries are less covered by academic research, even though the research itself seems to be focused on topics relevant to these nations, particularly related to agriculture and food. It was pointed out that there is not only a disparity between countries, but also one between the population within countries, which makes the need for action extremely urgent. A link of climate vulnerability to poverty is also shown, underlining the fact that the poorest people in the poorest countries will suffer the most direct impacts of climate change. The issue of drought, on the other hand, is very present but seems to have less global relevance than the issue of rainfall, and is more present in developing countries, especially in Africa, often linked to crop and irrigation issues. However, it seems to have more regional patterns, with countries where the issue is virtually absent, and in others where it is dominant. In general, it can be argued that vulnerable countries are also those that have greater food insecurity and are more economically disadvantaged. The impact of climate change in these countries creates vicious circles because they are more focused on recovery after disasters than on adaptation, having limited resources at their disposal. Disasters resulting from climate change also require the allocation of economic resources, which are in themselves scarce and are diverted away from country development. Most countries where agriculture is most relevant also have the majority of their population employed in the primary sector. The link therefore with agriculture and crops is not only related to food supply, which puts an immeasurable number of people at danger, but also to the economic livelihood of these countries. Moreover, often the same countries that are less prepared to deal with climate change are also those that have not caused it, with extremely lower CO^2 emissions than developed countries. The strength of academic research is therefore assessed on two different parameters. On the one hand, it seems to be varied and to address the issues and urgencies of each geographical location, thematically, but on the other hand, the focus seems to be more on the less vulnerable countries, following economic patterns rather than actual needs. Since there appears to be a link between academic research and the

reduction of a country's vulnerability, albeit all very much related to the economic level of the country itself, it is to be hoped that in a not too far future there will be international academic cooperation to share the knowledge required for mitigation and adaptation actions against extreme climate events. A glimmer of hope lies in the fact that countries with relatively recent economic development, i.e. Asian countries, are already increasing their weight in global academic research enormously, so the hope is that even the most disadvantaged countries can, in parallel with their economic development, bridge the gap in terms of both their research capacity and their preparedness for climate change, which is already underway, but whose violence will inevitably increase.

As therefore highlighted throughout the analysis, the challenges posed by climate change cut across any economic logic, and do not depend on the borders of a country. It is therefore crucial that, for an effective response, there is a collective effort involving all the nations of the planet, because as never before, the objective is too far away to be won by the efforts of individuals.

References

[1] World Bank. *World Bank data API*. URL: https://blogs.worldbank.org/opendata/introducing-wbgapi-new-python-package-accessing-world-bank-data.

[2] World Bank. *World Bank Open Data*. URL: https://data.worldbank.org/.

[3] Dimitrije Curcic. *Number of Academic Papers Published Per Year*. Wordsrated. 2023. URL: https://wordsrated.com/number-of-academic-papers-published-per-year/#:~:text=As%20of%202022%2C%20over%205.14,5.03%20million%20papers%20were%20published..

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (2019). arXiv: 1810.04805 [cs.CL].

[5] Ole Ellegaard and Johan A Wallin. "The bibliometric analysis of scholarly production: How great is the impact?" In: *Scientometrics* 105 (2015), pp. 1809–1831. DOI: https://doi.org/10.1007/s11192-015-1645-z.

[6] Leonhard Euler. "Solutio problematis ad geometriam situs pertinentis". In: *Commentarii academiae scientiarum Petropolitanae* (1741), pp. 128–140.

[7] Alex Evans and David Steven. "Climate change: the state of the debate". In: *Center for International Cooperation, New York University* (2007).

[8] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3 (2010), pp. 75–174. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2009.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0370157309002841.

[9] *GeoPy's documentation*. URL: https://geopy.readthedocs.io/en/stable/.

[10] *Gephi*. URL: https://gephi.org/.

[11] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[12] Maarten Grootendorst. *KeyBERT*. URL: https://maartengr.github.io/KeyBERT/index.html.

[13] *H2-keywordextractor*. URL: https://huggingface.co/transformer3/H2-keywordextractor.

[14] *Hugging Face*. URL: https://huggingface.co.

[15] Notre Dame Global Adaptation Initiative. *Country Index*. URL: https://gain.nd.edu/our-work/country-index/.

[16] Nazrul Islam and John Winkel. "Climate change and social inequality". In: (2017).

[17] Kcnarf. *Voronoi treemap, sample code*. URL: https://gist.github.com/Kcnarf/fa95aa7b076f537c0

[18] Kcnarf and KathyReid. *Voronoi treemap*. URL: https://github.com/Kcnarf/d3-voronoi-treemap.

[19] Rodney Michael Kinney et al. “The Semantic Scholar Open Data Platform”. In: *ArXiv* abs/2301.10140 (2023). URL: https://api.semanticscholar.org/CorpusID:256194545.

[20] Doris Klingelhöfer et al. “Climate change: Does international research fulfill global demands and necessities?” In: *Environmental Sciences Europe* 32 (2020), pp. 1–21. DOI: https://doi.org/10.1186/s12302-020-00419-1.

[21] Nada Maamoun. “The Kyoto protocol: Empirical evidence of a hidden success”. In: *Journal of Environmental Economics and Management* 95 (2019), pp. 227–256. ISSN: 0095-0696. DOI: https://doi.org/10.1016/j.jeem.2019.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0095069618300391.

[22] Gerald A Meehl et al. “Trends in extreme weather and climate events: issues related to modeling extremes in projections of future climate change”. In: *Bulletin of the American Meteorological Society* 81.3 (2000), pp. 427–436.

[23] M Monirul Qader Mirza. “Climate change and extreme weather events: can developing countries adapt?” In: *Climate policy* 3.3 (2003), pp. 233–248.

[24] *NetworkX*. URL: https://networkx.org/.

[25] Mark EJ Newman. “Fast algorithm for detecting community structure in networks”. In: *Physical review E* 69.6 (2004), p. 066133.

[26] Suphakit Niwattanakul et al. “Using of Jaccard coefficient for keywords similarity”. In: 1.6 (2013), pp. 380–384.

[27] *NLTK, Natural Language Toolkit*. URL: https://www.nltk.org.

[28] University of Notre Dame. *Readiness and vulnerability index*. URL: https://gain.nd.edu/our-work/country-index/.

[29] *OpenStreetMap*. URL: https://www.openstreetmap.org/about/.

[30] A. Ozpinar. “A Hyper-Integrated Mobility as a Service (MaaS) to Gamification and Carbon Market Enterprise Architecture Framework for Sustainable Environment”. In: *Energies* (2023). URL: https://api.semanticscholar.org/CorpusID:257386701.

[31] Sabine Pahl et al. “Perceptions of time in relation to climate change”. In: *WIREs Climate Change* 5.3 (2014), pp. 375–388. DOI: https://doi.org/10.1002/wcc.272. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.272. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.272.

[32] Serge Planton et al. “Expected impacts of climate change on extreme climate events”. In: *Comptes Rendus Geoscience* 340.9-10 (2008), pp. 564–574.

[33] *Pycountry*. URL: https://pypi.org/project/pycountry/.

[34] *Semantic Scholar API*. URL: https://www.semanticscholar.org/product/api.

[35] Monica Singhania and Gurmani Chadha. “Thirty years of sustainability reporting research: a scientometric analysis”. In: *Environmental Science and Pollution Research* 30.46 (2023), pp. 102047–102082. DOI: https://doi.org/10.1007/s11356-023-29452-2.

[36] *SpaCy - Linguistic Features*. URL: https://spacy.io/usage/linguistic-features.

[37] International Organization for Standardization. *ISO 3166-1*. URL: https://www.iso.org/iso-3166-country-codes.html.

[38] Giorgio Strano. *La biologia molecolare*. Treccani Encyclopaedia. URL: https://www.treccani.it/enciclopedia/la-biologia-molecolare_%28Storia-della-civilt%C3%A0-europea-a-cura-di-Umberto-Eco%29/.

[39] Chi Sun et al. “How to Fine-Tune BERT for Text Classification?” In: (2020). arXiv: 1905.05583 [cs.CL].

[40] Nicolas Taconet, Aurélie Méjean, and Céline Guivarch. “Influence of climate change impacts and mitigation costs on inequality between countries”. In: *Climatic Change* 160 (2020), pp. 15–34.

[41] Anastasia Telesetsky. “The Kyoto Protocol”. In: *Ecology Law Quarterly* 26.4 (1999), pp. 797–813. ISSN: 00461121. URL: http://www.jstor.org/stable/24113942.

[42] *The Global Climate Risk Index*. URL: https://www.germanwatch.org/en/cri.

[43] Patrick J Tierney. “A qualitative analysis framework using natural language processing and graph theory”. In: *International Review of Research in Open and Distributed Learning* 13.5 (2012), pp. 173–189. DOI: https://doi.org/10.19173/irrodl.v13i5.1240.

[44] Marco Tobio. “The Effects Atmospheric Changes Have on Runoff”. In: *Inquiry@Queen’s Undergraduate Research Conference Proceedings* (2023). URL: https://api.semanticscholar.org/CorpusID:257882443.

[45] Unesco. *UIS.Stat (2019) Data 2017*. URL: https://data.uis.unesco.org/Index.aspx.

[46] European Union. *EU (2020) 2030 climate & energy framework*. URL: https://ec.europa.eu/clima/%20policies/strategies/2030_en.

[47] Ashish Vaswani et al. “Attention Is All You Need”. In: (2023). arXiv: 1706.03762 [cs.CL].