SciencesPo
SCHOOL OF PUBLIC AFFAIRS

PUBLIC POLICY MASTER THESIS

April 2021

Can scientific evidence contribute to the improvement of environmental policies?

A semantic analysis of three supranational institutions challenges and biases in the process of aggregating science to improve policies

Yann David
Master's Thesis supervised by Philipp Brandt
Second member of the Jury: Jean-Philippe Cointet

Abstract

In this Master's thesis, I explore biases in the relation between environmental policy-making and science, introduced both from the supply and demand sides. This research specifically questions the lack of popularity of impact evaluations in the environmental field. It uses a methodology of semantic and network analysis to illustrate biases in the use of science in reports from three supra-national institutions: the OECD, the European Commission, and the World Bank. The study allows to observe the political orientations of the entities as well as clear evidence of *selectivity in evidence* and *altered aggregation bias*. It ultimately provides four directions along which substantial improvements could be made to mitigate supply and demand issues identified.

Key words

Evidence-based policies, science network, counterfactual, impact evaluation, policy-making, bias, systematic, meta-analysis

Contents

Acknowledgments

I would like to thank my supervisor Philipp Brandt who, on top of being both a great professor and as interested as I am in nerdy discussions about programming, has been of real help in the elaboration of this Master's Thesis. I am very grateful for his precious advices, and especially when he brought me back on track at this moment I was about to scrap everything I could on the web but had forgotten that the goal of this exercise was to come up with a strong methodology and interesting results. I hope it will be the case.

I am also eternally grateful for the luck, and substantial supports from my wonderful parents, that allowed me to spend six incredible study years that taught me so much. I stepped into this university with the will to make things outside a bit better. This final work shows that this objective has evolved in conjunction with memorable friendships and discussions as well as with inspiring courses and readings, but has surely not faded away.

Finally, I cannot avoid mentioning my profound gratitude towards my mother, father, brother, sister, and friends: their support has been invaluable throughout this structuring period in one's life, mine included.

Why care about evidence-based environmental policies and read this research?

With the development of a wide range of statistical techniques ensuring causal identification, a scientific moment entitled the "credibility revolution" by Angrist & Pischke (2010), development policies have been increasingly evaluated in the academic literature over the past twenty years. These results have been more and more used by public and private institutions to increase their legitimacy and justify their decisions regarding the future of these evaluated programs, hence renaming their decisions *evidence-based policies*.

While this approach to policy-making has become a golden standard methodology in development, education or health fields, environmental policies have remained quite indifferent to impact evaluations. This situation is very paradoxical for a policy area that has benefited from the most important scientific effort ever realised to characterise the stakes and potential consequences of climate change and biodiversity decline with the creation of the IPCC and the IPBES platforms of scientists. Moreover, scientific consensus has been reached since at least two decades in these research fields: global, rapid and efficient actions need to be implemented to limit humans footprints on earth and preserve its livability. So why do environmental policy-makers resist to impact evaluations that would help them fine-tune their programs?

In this research I first synthesise, in the interdisciplinary literature review, reasons explaining the lack of policy evaluation to inform environmental decisions. I argue that the problem comes both from important supply-side methodological barriers that have not yet been overcome, as well as from demand-side lack of training and resources.

This Master's Thesis explores demand-side problems: policy-makers use and misuse of science in the specific context of environmental policies. The focus is put on the observation of biases introduced by practitioners in their use of academic evidence. More specifically, I use a corpus of 1,505 institutional reports from the Organisation for Economic Cooperation and Development, the European Commission and the World Bank and analyse their semantics and references networks to illustrate both issues of *references picking* and *evidence oriented aggregation* biases.

The main contribution of this work is to isolate and assess the magnitude of bias introduced in each of the two steps mentioned just above. These findings open the room for improved use of science in the design of environmental policies. Hence, the final section of this research provides four major work tracks along which demand and supply side hurdles to informed and efficient policies can be importantly mitigated.

1 Introduction

With the rapid development of a quantitative approach to measuring the efficiency of public policies, a powerful paradigm has emerged and progressively imposed itself in the world of politics over the past decades. This approach to building political agendas is often called *evidence-based policy making*. It has become the new golden standard for national and supra-national institutions to legitimate the rigour and unbiasedness of their respective orientations. Evidence-based policy making consists in the extensive use of academic knowledge to asses and prioritise stakes as well as to measure the ex-ante and ex-post efficiency of policy strategies. Evaluation of public policies is therefore at the core of the process. Moreover, under the influence of empirical economics, *counterfactual* thinking is nowadays extensively applied for the measurement of programs impacts on ranges of indicators. This approach gathers a range of statistical modeling methods that aim at isolating and identifying the effect of a policy on different outcomes. As defined by Ferraro (2009):

> *The essence of counterfactual thinking is elimination of plausible rival interpretation of observed outcomes*

Everything else is theoretically held constant (*ceteris paribus*) such that if a positive impact is measured, it implies that the program has a significant impact and should be further developed. The theoretical intuition behind this approach is very straightforward and has therefore convinced much policy making institutions to use these results as reliable and powerful guides to their decisions.

While counterfactual evaluation of programs has become dominant in the evaluation of development, education and health policies - this triumph being crowned by Economics' Nobel Prizes in 2019 - an important policy area seems to remain out of the scope of this paradigm: environment. Indeed, very few impact evaluations of programs have been proposed in this field. Newig & Rose (2020) explain that this specificity of environment policy making may be attributable to the high variety of profiles and backgrounds both researchers and policy-makers working in this field have. Nonetheless, the policy challenge is at least as crucial to the future of societies as development programs. Facing the emergency to reduce humans pressures exerted on climate and ecosystems, informed and efficient policy choices should be made. As Maki et al. (2018) argue:

> *To more effectively influence these environmental behaviors, we need policies informed by sound social science that help people engage in behaviors that benefit the environment, and at the same time are not too costly or onerous to the individuals being asked to change their behavior*

So why hasn't the counterfactual paradigm become a standard in the process of environmental policy making? The idea that such approach is required to improve programs' design is widely shared across the literature, but a central challenge contravene this will: complexity. In contrast with most development policies evaluations, environmental policies are often implemented within interconnected ecosystems, making it a real empirical challenge to quantify multifaceted effects of a program. For this reason, many environmental policies are monitored

with the use of descriptive data and are forecasted using modeling approaches. From a counterfactual thinking view, these ex-ante and ex-post approaches are a real concern because it implies no one knows how effective policies are, and if the money is thus spent efficiently to meet targets. Even though the counterfactual paradigm is challenging to implement in some contexts, the need for *evidence-based policies* is strongly agreed upon in the literature.

In this Master's Thesis, I explore how environmental policy-makers, broadly defined as public actors enrolled in the design or evaluation of public policies, interact with scientific knowledge and counterfactual-based studies to extrapolate these *evidences* into informed policies. My fundamental interest is to understand how science can be better used to improve environmental policies outcomes. Three stages of the problem can be identified. The first one is the process of science creation: do publications answer to the specific questions of policy-makers? The second source of distortion may simply be barriers to accessing knowledge: academic publications are (wildly) published on multiple platforms, some of them being accessible only after payments. Finding relevant articles is thus a very time and cost expensive process. Hence, policy makers in charge of screening the literature may introduce a *selection bias* to the analysis - implying that sources picked are not representative of the true state of knowledge on an issue. The third source of deformation may come from deliberate or unconscious information alteration during the process of summarising retrieved publications. The consequence being a distorted representation of actual scientific knowledge.

This research investigates how policymakers are currently managing the two latter stages. More specifically, I focus on the role and behaviours of the environmental departments of three supranational entities, namely the Organisation for Cooperation and Economic Development (OECD), the European Commission (EC) and the World Bank (WB). I selected these three entities because they position themselves as technico-scientific neutral actors of public debates. Their distance to country-specific politics make them more likely to use science as a source of legitimisation for their policy orientations. My interest is to understand how they build on different sources of knowledge to construct their policy proposals.

I attempt to answer to two questions. Firstly, I want to understand how the *evidence-based policy* paradigm is shaping these three institutions' environmental policies. Secondly, I try to measure how unbiased their use of science is. According to the description provided in the previous paragraph, bias may stem from unrepresentative selection of articles and oriented synthesis of articles.

To meet these ambitious research objectives, I rely on the methods of semantic and network analysis. My study is thus divided into two main components. I start with the examination of a corpus of more than 1,500 environmental reports from the three institutions. Reports are all published between 2008 and 2021. The study is focused on the comparison of narratives evolution across time and institutions, as well as the examination of how the *evidence-based policy* paradigm is impregnating their publications. The second component of the study is a case-study of three reports discussing *nature-based solutions to water-related risks caused by climate change*, each one of them published by one of the three entities between 2018 and 2020. This second moment in the study is used to adopt a sort of counterfactual approach. The goal is

indeed to compare how the entities address a very similar research question in terms of articles selection and retrieved literature synthesis. I use semantic and network analysis in this second part.

The following chapters are organised as follows. In the first chapter, I realise an interdisciplinary state of knowledge on the identified limitations to the development of evidence-based environmental policies. I then present in details the protocol used to construct the two databases, the methodologies to analyse them, and I finally present the retrieved data characteristics. In the fourth chapter, I turn to the examination of the two data sets of OECD, European Commission, and World Bank publications on environmental topics. After conclusive remarks in chapter five, bridging the literature review and the study of the data set, I develop some policy recommendations in the sixth and final chapter.

2 Interdisciplinary state of knowledge: What are the limitations to evidence-based environmental policies?

In this chapter, I propose an overview of the stakes discussed in the academic literature and related to evidence-based policy making. Because the implicit assumption throughout the next analysis chapter is that designing policies based on academic knowledge necessarily improve outcomes, the present chapter focuses on limitations to this idealised vision. The first section discusses general issues in the literature that may prevent one from considering academic material as a perfectly relevant and trustworthy source for policy making. In the second section, I deep further into the specificities of environmental policies, and what makes them so hard to be analysed through the counterfactual lens evoked in the introduction. Finally, because the present thesis primarily focuses on policy-makers biases, the third section of this chapter covers literature inputs on that matter.

2.1 General limitations to scientific knowledge accumulation for policy-making

Should we trust science? This question, as provocative as it may look like within a *wanna-be-academic* document, is a legitimate inception to the examination of how scientific results should be used to improve policy outcomes. Of course, the scientific method - with the decisive role of the peer-review process as a quality ensuring institution - is the more robust approach to evidence building. As such, academic publications do not provide *truth* about things, but rather consensus about the best possible knowledge on things, given the best methodological techniques known so far and commonly approved by a community of researchers. This characteristic of science implies that its results are inherently associated to varying degrees of uncertainty. Jasanoff (2007) calls this feature *the asymptoticity of perfect knowledge*, a metaphoric expression that symbolises the never-ending iterative and cumulative nature of science. Furthermore, as Shwed & Bearman (2010) synthesise in their paper, sociology of science has shown how scientific consensus building is influenced by politics, culture, fundings and credibility. An interview realised for the present thesis with a Commissioner from the European Union stressed the important role of the institution as as research grant-maker. The interaction between science and politics is therefore not unilateral but is structured around feedbacks that progressively shape orientations of all stakeholders. As Fujimura (1996) emphasises, academic consensus is also built on the fortification of bandwagon practices. There exists path-dependency in knowledge accumulation, such that some topics may not yet generate consensus simply because they have not been explored.

However, even with the most rigorous methods to ensure publications' quality, researches have shown that the academic literature is not *bias-proof*. Research institutions themselves may indeed generate incentives altering the quality of publications. The most famous *autoimmune illness* in academia is the *publication bias*. This expression designates the impetus that pushes reviews to favour articles showing positive results, as explained by Peplow (2014). This constitutes a serious threat to the reliability of science because it lowers chances of articles presenting a null result to be published, even though they may accurately represent reality of the studied phenomenon. In empirical economics, this problem is entitled *p-hacking*. This expression comes

from the rule-of-thumb to consider results to be statistically significant when their p-value is higher than 95%[1]. This 95% confidence level was for long arbitrarily considered as a cut-off to reject the "null hypothesis". As shown by Brodeur et al. (2020), this *tradition* has had detrimental effects on the entire causal economics literature, where researchers have started to adjust their models and approaches in order to *hack* and reach this p-value level. This research shows that the distribution of published articles p-values, which should theoretically follow a t-distribution, features a bump just after the 95% significance level - indicating an important publication bias in favour of positive results. Moreover, the article shows that the phenomenon is observed in all journals indifferently from their reputation and across all empirical methods. On the positive side, authors also find a decrease in the magnitude of the problem over time. In their article, Andrews & Kasy (2019) propose an approach to re-estimate published results and correct for the p-hacking problem. More generally, the problem is now acknowledged by the entire discipline and actively debated. Methods ranging from study protocols mimicking medical trials to most sophisticated machine learning techniques are being tested to reduce this publication bias - which contravene to the *empirical credibility revolution* optimistically narrated by Angrist & Pischke (2010). In the perspective of extrapolating evidence to improve public policies, publication bias is a serious concern. Indeed, the existence of these *fake positives* can lead to misleading interpretations and in turn cause wrong policy decisions.

Finally, raw results from policies impact evaluations papers may be inappropriate to policy-makers' needs. Indeed, the discipline suffers from a blinded quest for causality in the context of the study, also called *internal validity*, which completely sets aside issues of results' generalisability. Furthermore, some methods employed to identify an effect are only *local*: they measure a causal impact only for individuals who respond positively to an incentive (traditionally called an instrument in this literature) - a behaviour that is very likely to be context-dependent. Hence, results are often not reproducible as demonstrated by Chang & Li (2015). To address these issues, reflections about *external validity* has only recently emerged in the policy evaluation literature, with two seminal papers by Meager (2016) and Vivalt (2015). From a policy-maker's perspective, it is therefore not self-evident that empirical estimates from studies could be used to design policy programs in different contexts. Another institutional incentive, briefly mentioned by Vivalt (2015), exacerbates the problem caused by the internal validity centered approach. Indeed, researchers are incited to be first-mover on a topic such that they can be attributed the parenting of a concept and be very much cited. This implies scarce evidence on numerous topics, which reduces the possibility to gauge external validity of results.

Evidence-based policies should therefore not consist in a naive use of science to justify political orientations. Rather, policy makers should account for the potential biases exposed above and be extremely cautious in hasty extrapolation of available evidence. However, science availability on a topic is not ex ante guaranteed. Hence, consensus building about environmental policies particularly suffers from these evidence gaps.

[1] In standard language, this means that if the true population effect was null and if we were indefinitely re-sampling from the population and re-measuring the effect, the probability to observe an effect at least as extreme as the one measured in the observed sample would be equal to 5%

2.2 Specific challenges for evaluating environmental policies

In this paper's introduction, I briefly mentioned some specificities of environmental policies increasing the difficulty to assess their impacts. This section details these discussions from the literature.

According to Newig & Rose (2020), the first hurdle to consensus building in the study of environmental policies is that the community of researchers it groups together come from "*very different disciplinary backgrounds [..] loosely held together by a common research topic*". The direct consequence is a multiplication of concepts proposed to define phenomenon that are often very related. In this context, Fujimura (1996)'s *fortification of bandwagon practises* does not occur. Furthermore, because researchers still consider their disciplinary belonging as central for their legitimacy, they value publishing in their own discipline's journals more than interdisciplinary reviews. Aggregating evidence in the perspective of policy-making is thus complicated by the dispersion of resources across journals and disciplines.

A second challenge in the creation of a structured knowledge on environmental policies, is the multiplicity of impacts they can have. Indeed, political, social, economic, biological, chemical and ecosystemic indicators may be considered. These evaluations are thus by nature interdisciplinary - which creates two challenges. First of all, researchers' lack of interdisciplinary skills may prevent them from assessing the effect on all indicators. Assessments are thus partial and may not address all policy-makers' concerns. Secondly, interdisciplinarity implies varying and sometimes clashing approaches to answer a similar question. It makes this pool of evidence methodologically very heterogeneous and in turn challenging for policy-makers to be managed and exploited.

Finally, Ferraro (2009) proposes a discussion about challenges of adapting counterfactual thinking to environmental policies impact evaluation. He argues that most of what is nowadays called evaluation of environmental program should in fact be called monitoring of indicators: no causality is identified. These measures are thus incapable to isolate the true effect of implemented policies. Many confounding factors, other than the intended program, may influence the observed indicators in one direction or another. Using dashboards of raw time series to analyse the impact of a policy is not informative and can lead to false interpretation and wrong decisions. For this reason, Ferraro (2009) argues for environmental policies to be evaluated under the counterfactual paradigm.

Nonetheless, very few examples of experimental or quasi-experimental evaluations exist up to now in this field. Reasons behind this lack of academic interest are mainly methodological. Empirical challenges are indeed very important in the identification of causal effects. Amongst others, some specificities of environmental problems include nonlinear response outcomes such as threshold, high rate of outcome variability, infrequent data sampling, long time lag between intervention and response, spillover effects, large spatial effects. Some of these challenges are also found in other social policy fields, but they are particularly pervasive in the context off environmental policies. Ferraro (2009) argues that second best approaches can be adopted to approximate answers to the big questions. For instance, he proposes that instead of trying to assess the impact of a policy on long-term environmental indicators, one could start by looking

at behavioural changes that were triggered. If positive changes are observed, it can be assumed that they will have an impact on the environmental factors of interest. The evidence puzzle may hence be simplified, but challenges to extrapolate its pieces into policy decisions remain important.

2.3 Translation of science into policy-making or the room for additional biases

In this final section of the literature review, I will assume that research limits presented above are acknowledged and controlled for by policy-makers. Assuming a body of literature exists and is exploitable for policy decisions, it is now in the hand of policy-makers to do so. Unfortunately, this final process of academic evidence extrapolation into policy decisions may still add layers of bias to the foundations of *evidence-based policies*.

Experts and policymakers may for instance select a subset of articles best aligned to their political preferences as explained by Ingold & Gschwend (2014), leading to the construction of an unrepresentative set of publications, which ultimately yields an eroded aggregation of scientific evidence - whatever methodology be employed for aggregation. The problem is that there exists no simple solution to build an exhaustive sample of articles - and no one can even tell what a relevant set of articles is for any given topic. In the context of environmental policies, the dispersion of concepts across an important number of journals from different disciplines, as explained by Newig & Rose (2020), makes the screening process even more likely to be skewed towards an unrepresentative subset of the literature. This is not to mention politically oriented articles' picking.

With article selection as an unavoidable, but mitigatable, source of bias, policy-makers are unfortunately more prone to make *evidence-biased* than *evidence-based* policies. But problems do not stop here, and the challenge of studies' *external validity* constitutes another hurdle for practitioners. Indeed, extrapolation of impact evaluations to another policy context is a very tricky step. To complicate this task, Ferraro (2009) notes the lack of practitioners' skills in understanding the different policy evaluation methods, and researchers often avoid discussing the real extent of generalisability of their results. In an ideal world, policy-makers could use meta-analyses methods to estimate credibility intervals of the likelihood of external validity of a set of papers analysing a similar policy impact in different contexts. However, these sophisticated statistical methods are very new in the literature (Meager (2016), Meager (2019), Vivalt (2015)), have never been applied to environmental policies impact evaluations, may not be properly implemented nor interpreted, and rely on the quality of source studies to yield informative results.

Finally Vivalt & Coville (2017) show how policy-makers *ex-ante* overestimate the likelihood of a program positive effect compared to researchers. She also finds that practitioners do not update symmetrically to evidence, meaning that they do not easily accept and use academic evidence when it is not aligned with their ex ante beliefs.

In the following chapters, I will focus on the issues of *selectivity* and *aggregation* bias explored in this third part of the literature review. I will investigate how three supra-national institutions,

the OECD, the World Bank, and the European Commission, compare in terms of their use of evidence to construct environmental policy reports.

3 Methodology

3.1 Approach

The analysis of this Master's Thesis is dedicated to the understanding of processes through which environmental policymakers currently aggregate and extrapolate knowledge from academic publications - the latter being here referred to as *evidence*. To achieve this goal, I analyse reports' writing practices developed by three supra-national institutions which were chosen because they partly base the legitimacy of their policy proposals on their supposed scientific foundations. From a theoretical perspective, if the three institutions were to translate the current state of scientific knowledge into their reports, we would observe a convergence in the methods and topics tackled. In practice, however, one can expect *frictions* between academic discourses and the policy-world. This report investigates the magnitude of these institutional *touches of salt* and orientations. The analysis is divided in two parts.

In the first stage of the study, I perform a semantic analysis of 1,505 reports covering environmental policies and posted online by these three institutions. The objective is to observe the degree by which their orientations on the topic diverge. If the three entities were simply and objectively aggregating results from science, one would observe a degree of divergence equal to zero. To perform my analysis, I first try to understand how closely these three institutions approach environmental policy issues as a whole. To do so, I examine trends in words frequencies across reports and years. The goal is to identify the key ideas put forward in the entities' narratives about environmental policies. Secondly, I extract key words from an academic corpus discussing *evidence-based environmental policies* and examine the prevalence of these words, through time, in the institutions reports. This second step wishes to identify how much the *evidence-based policies* paradigm is used in the reports.

The first stage of the analysis described just above provides an overview of the three institutions narratives about environmental policies. Nonetheless, it fails at capturing how much these writings differ from the underlying evidence they cite. Furthermore, the three institutions may be interested in different topics such that the first stage would be comparing reports with non-overlapping contents. In other words, it may be biased because it lacks a *counterfactual*. In the second stage of the analysis, I try to correct for this issue. I restrict my focus on three reports, each written by one of the three institutions, discussing the same issue of *nature-based solutions to water-related risks*. I explore potential *aggregation bias* and *selectivity* in the articles picked to write reports. To examine potential aggregation bias, I try to build a counterfactual corpus from reports' references abstracts. Additionally, I make a cross-comparison of the level of penetration of the counterfactual paradigm in these reports compared to the importance of this narrative in cited references. Finally, to talk about selectivity of evidence, I study how and what references are used in the three reports. An ideal approach would have been to compare these networks of citations to the comprehensive network of scientific knowledge on this topic, but that would require a systematic approach to retrieving articles and delimiting boundaries of the topic. Time constraints impeded me to realise this project. I therefore turn to a second best approach where I get a sense of sources' representativity by comparing each report's network to the two remaining ones.

3.2 Data and Sources

Reports retrieval
The first pillar of the analysis is a systematic semantic analysis of publications from the OECD, the World Bank and the European Commission related to environmental policies. My main concern during the data set construction was the selection of irrelevant reports that would in turn alter the quality of any analysis. For this reason I defined, prior to any download, all the inclusion criteria described below. I then downloaded all reports from the three institutions websites meeting these criteria.

On the OECD website, I downloaded all the "Policy Papers" and "Environment Working Papers", which amounts to a total of 184 documents. I automatised the download of the PDFs and the extraction of dates and titles on the webpage, before turning these documents into a regular dataframe. Similar webscraping approach is adapted to the European Commission website. In the search bar, I request "general studies" about "environment" published either by the Directorate-General for Research and Innovation or by the Directorate-General for Environment. 583 publications' texts are extracted in this way. Finally, the World Bank makes our lives easier as it proposes an *application programming interface* (API) allowing download of metadata and texts in an easily manipulable format. I restrict my request to "reports" publications focused on "environment". The *API* returns 738 results.

Once the three databases are constructed, I clean the data for semantic analysis. The first step is reports' tokenization: a new dataframe is created with a row generated for each word in each report. The second step is to remove "stop words", that is to say words that do not provide any significant meaning, such as "and', "the", "if"... I use the list of *stopwords* from the *tm* package in *R* to accelerate this step. Finally, I apply a stemming algorithm to remaining words, in order to shrink them to their roots and gather all unique meanings behind a similar token. The algorithm works as in the following example ; it transform, all words like "technics", "technical" into a similar "techni" token. The resulting data set is cleaned and ready for analysis.

Citations data extraction
For the second part of the analysis, I restrict my attention to three reports focusing on the topic of "nature-based solutions to water-related risks", as presented in Table 1. As previously mentioned, I am interested in performing a counterfactual-like analysis based on the semantics and networks of citations. To build this dataset of references, data extraction is performed in two steps.

The first step of the data extraction is to retrieve all sources from the three reports' PDF. Seeking results' replicability, the procedure is coded in *R*. The extraction is realised in several substeps. I start by extracting the references' pages and create a vector of *raw* references formatted in the style chosen by the report. I then isolate the title, the vector of authors, the DOI if available, and the publication date for each source using regular expressions. The final output is a list of references for each original report, with a sub-list for each reference containing its title, date and vector of authors.

Table 1: Description of the three reports compared

Title	Institution	Year	Number of references
Nature-based solutions for disaster risk management	World Bank	2018	57
Nature-based solutions for flood mitigation and coastal resilience	European Commission	2020	91
Nature-based solutions for adapting to water-related climate risks	OECD	2020	93

The second step is to seek for metadata on each of these references and for their inner references. To perform this research, I was granted access to the *Web of Science* database. However, given the slowness of looking at all references manually, I propose an alternative automatised approach. I have coded a class of functions in Python using the *Cross-Ref API* from Ynnig (2020) and the *Selenium* library from Muthukadan (2011) that allow me to:

1. Use CrossRef API to retrieve meta-data for references that have a *DOI* reported in the reports. I also search for the DOI with the Cross-ref API if it is not reported. The meta-data returned by cross-ref is not consistent, because it depends on what the publisher has decided to report. However, it always gives the title, the type of document, the authors, their affiliations, and sometimes the references included in the document. For references that did not yield results on this database, I use the Web of Science website.

2. I have also automatised search and download of metadata on Web Of Science using the selenium library for python. Selenium allows me to simulate navigation in my browser. The bot speed can be really fast, as it is only limited by my internet connection - but I have decided to send less than 600 requests per hour, with random breaks in the script - so that the website does not suspect I am using a bot and does not block me. In this way I can download information about 2 to 3 references per minute. It is not record breaking, but I don't have to do it "by hand". The idea here is just to look for the references that had no DOI, and also to look for the references found on cross-ref but that did not report their inner references.

The clear advantage of this approach is replicability, and relative rapidity. Furthermore, it allows for rapid scale increase of the network of inner references retrieved. However, this approach does not prevent from a final manual data cleaning step to ensure the consistency of article titles required for the network analysis. Furthermore, it failed at finding meta-data for all references.

3.3 Data Overview

I now briefly present the data retrieved. After a focus on the reports data set I describe the citations data set.

Reports Retrieved

The automation of reports download and text cleaning allow me to create a corpus from 1,505 documents published after 1990. As shown in Figure 1, the density of publication is not constant over time and do not feature a similar shape between institutions. Indeed, the OECD publication intensity is almost constant as of 2009, whereas the European Commission primarily starts publishing after 2013 but with a decline in number over time. The World Bank trend is in contrast exponential after 2015. These changes in published reports may be an artefact: the publication date extracted from websites may only indicate the moment they were put online and not the actual year they were written - even though this is the information I requested to the World Bank API. If this is the byproduct of a new open-data policy, it could explain the European Commission peak in 2013 and the World Bank exponential growth in publication. The trend may also partly capture diverging interests in environmental topics.

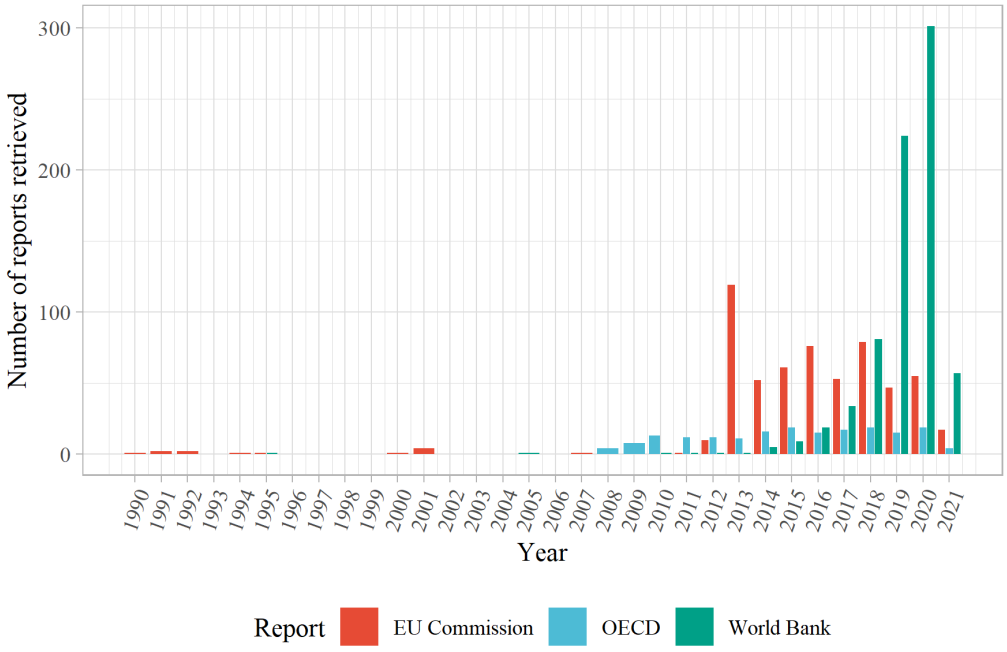


Figure 1: Number of Reports Retrieved by Year by Institution

Citations

In the two following tables, I present the results from the procedure presented above and applied to download meta-data about references. Table 2 presents summary statistics for the meta-data that could be retrieved about references from the three original reports. The success rate is not very high for the World Bank and OECD reports and is higher than 50% only for the European Commission. The third column shows the number of references for which inner references were listed. This allowed to seek for them as presented in Table 3.

Indeed, Table 3 reports the number of second degree references identified. The OECD is the most populated network, whereas the World Bank one is the smallest. Nonetheless, the gap in data availability surprisingly shrinks in the last column as the World Bank references show the

Table 2: Summary Statistics for Meta-Data Retrieval of First-Degree References

Report	Nb References	Found Meta-data	With Inner Ref
OECD	93	38	29
World Bank	57	22	14
European Commission	91	61	22

most frequent abstract finding rate.

Table 3: Summary Statistics for the Second-Degree References

Report	Number of Second degree refs	Second degree ref with abstract
OECD	1061	219
World Bank	291	130
European Commission	799	120

Overall, the data is clearly unbalanced and imperfect. The representativity of the networks and of the abstracts' corpus can clearly be questioned and criticised. Substantial improvements in the data would have required either a lot of coding time, to automatise references extraction in PDFs of references for which no meta-data was available, or patience in doing it entirely manually. I leave the improvement of the dataset creation functions for further research. In the following sections and chapters I propose to start by assuming that the data is representative, and then to relax this strong assumption and see how it can also help to explain some patterns.

3.4 Analytical Methods

Now that the data set creation and structure are explicit, I turn to the presentation of my analysis methodology. The study is decomposed in two parts: an analysis of the whole corpus of reports and a comparison of the three publications on nature-based solutions to water-related risk. I present my methodology following this structure.

3.4.1 Reports Analysis

The first part is hence the study of the entire corpus of reports. My objectives are twofolds. I want to understand the topical trends across institutions and time as well as the degree of impregnation of the *evidence-based policies* paradigm. The content at hand is a data frame containing rows of words classified by year, institution, report.

To reach the first goal, I choose to stick to the simplest possible form of data examination: I look at the evolution of words frequencies across the two dimensions of interest. A word's frequency is defined as its number of occurrence per institution per year over the total number of words in this institution-year corpus. For this statistics to make sense, I assume that words with the highest frequencies in each corpus are key-words illustrating political priorities and orientations

about environmental issues.

To study the impregnation of an *evidence-based policy* paradigm, the first challenge is to define concretely what it means and how to detect it. Once again, I decide to adopt a transparent and straightforward approach to this issue. I create a corpus of eight academic publications discussing the issue of environmental policies evaluation and extract the most frequent words in this corpus. Table 4 presents these articles. I assume that tokens present more than 50 times in the corpus capture the *evidence-based policies* paradigm. Screening the evolution of their usage in the corpus of institutions' reports is therefore a proxy for the popularity of this narrative in the policy process. As for the topical part discussed above, I stick to the examination of words' frequency across the two dimensions of interest to understand the impregnation of an *evidence-based policy* paradigm.

Table 4: Articles used in the *evidence-based policy* corpus

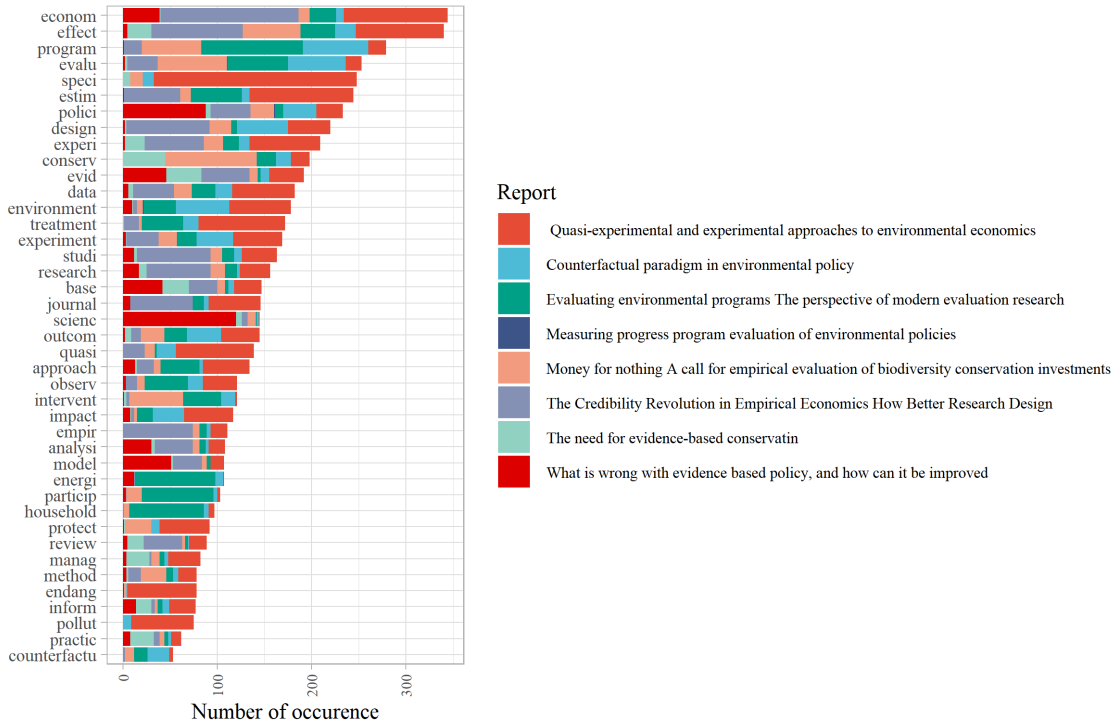
Title	Authors (Year)
What is wrong with evidence based policy, and how can it be improved?	Saltelli & Giampietro (2017)
The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics	Angrist & Pischke (2010)
Counterfactual Thinking and Impact Evaluation in Environmental Policy	Ferraro (2009)
Quasi-experimental and experimental approaches to environmental economics	Greenstone & Gayer (2009)
Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments	Ferraro & Pattanayak (2006)
Measuring progress: program evaluation of environmental policies	Bennear & Coglianese (2005)
Evaluating environmental programs: The perspective of modern evaluation research	Frondel & Schmidt (2005)
The need for evidence-based conservation	Sutherland et al. (2004)

In Figure 2, I plot the number of occurrence of the most frequent words throughout the eight articles. Before turning to the semantic examination, I remove words that are describing environmental topics rather than the *evidenced-based policies* paradigm, such as "speci", "conserv", etc.

3.4.2 Comparison of the three reports on nature-based solutions to water-related risks

In the second time of the study, I focus on three reports tackling the similar issue of nature-based solutions to water-related risks. By narrowing the research to only three documents, the objective is to compare reports that are very close to one another in terms of topic. I thus adopt an

Figure 2: Top words in the academic corpus about evidence-based environmental policies



approach that attempts at getting as close as possible to counterfactual thinking. Moreover, this focus allows to dig further into the details of each publication. I can for instance compare their references and therefore discuss the issue of *evidence selectivity*.

Before turning to a network analysis of references, I first try to answer the same questions as for the whole corpus. I therefore look at words' frequency to understand how differently the three reports tackle a similar policy issue. I then focus on potential *aggregation bias*, where I compare the most frequent words in the references abstracts to the most frequent words in the reports. If reports' authors were purely translating evidence into policy proposals, we should observe similar patterns of words frequency. Unfortunately, my counterfactual pool of abstracts is not representative of the real domain of abstracts as I could not retrieve all of them on *Cross-Ref* or on *Web of Knowledge*. As there exists no second best solution to this issue, I propose an indicative analysis with the data at hand. I also use this counterfactual-inspired approach to analyse the impregnation of *evidence-based policy* semantics in reports in comparison to their references' abstracts.

Finally, the ultimate analysis section uses network representations to compare the sets of references within each report. Once again, the proposed analysis only holds if abstracting from the data incompleteness detailed before. The idea is to get a sense of the *evidence selection bias*. I propose to approach this issue by focusing on two network features. First of all, applying the law of large numbers to our problem, the more populated is the network, the more one may expect that selected references are representative of the underlying academic controversies and trends. Secondly, the more connected is the network, the more sources retrieved are citing each other -

which in turn implies that a consensus is likely to have been reached within the network, and that the more cited articles are recognized by their peers. In an ideal world, I would be comparing each report's network to the true network of evidence on the topic of nature-based solutions to water-related risks. However, time constraints and methodological barriers made this project unfeasible, for now. I therefore turn to a second best approach which consists in comparing the reports characteristics between one another.

4 Analysis - Policies and evidences in supra-national organisations, the case of the World Bank, the European Commission and the OECD environmental reports

In the following chapter, I now move to the analysis of the corpus of reports and references retrieved. Data examination answers to three objectives. Firstly, I want to understand how similarly the three institutions have approached environmental issues, thanks to careful study of their vocabulary choices. Secondly, I analyse how much the "evidence-based policies" paradigm has impregnated reports' semantics over time. Finally, I will focus on three reports discussing nature-based solutions to water-related risks, each of them published by one of the three institutions, to discuss the issue of references selection and representativity.

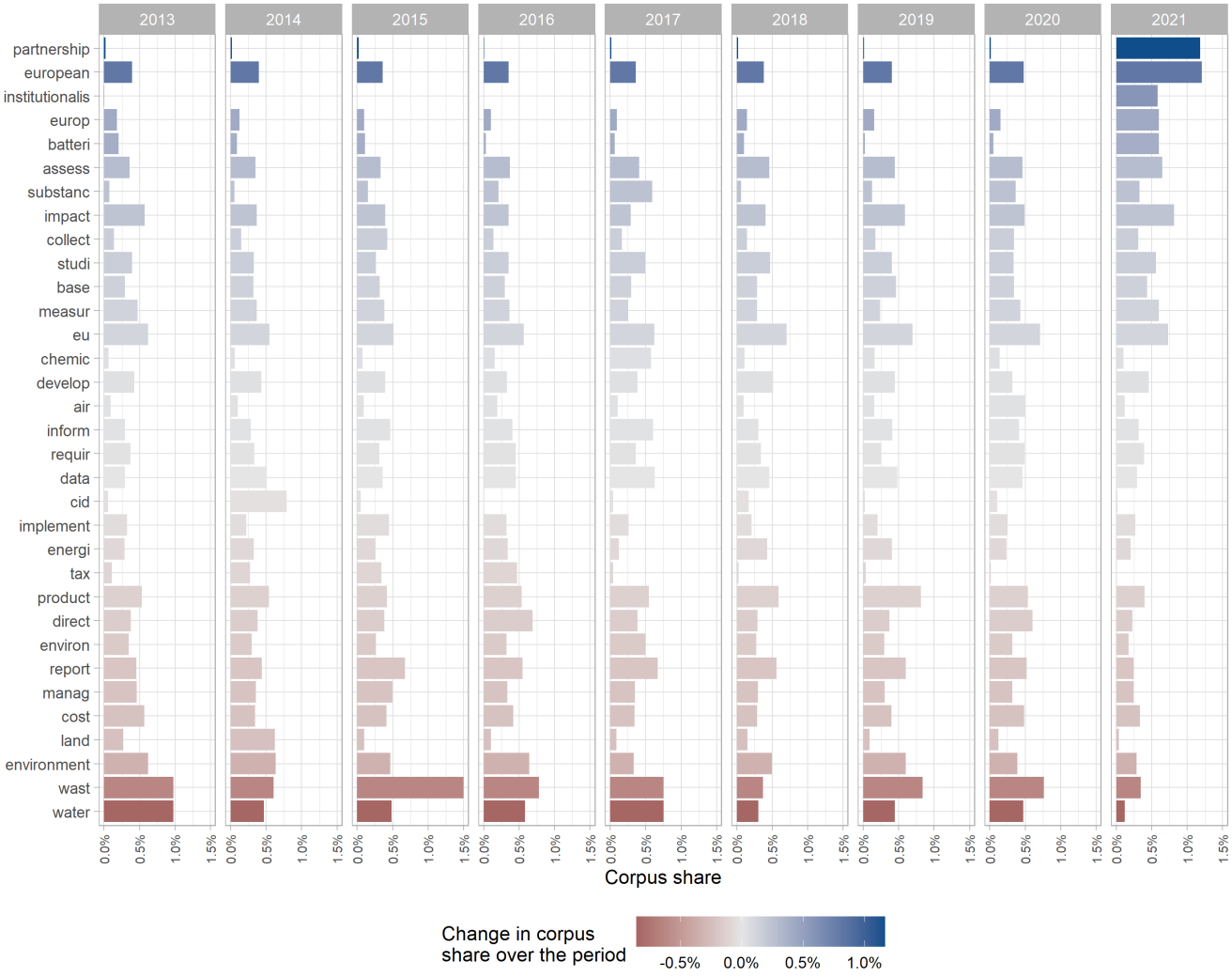
4.1 Trends in the institutions' semantics about environmental topics

The first pillar of the analysis is semantic and consists in counting the most frequent words in each institution corpus over time, once stop words are removed. In provided graphs, I have ranked words by descending order of change in frequency over time, from the older period until 2021.

The first graph, Figure 3, shows the words representing the largest share in the European Commission corpus of 583 reports on environmental policies. Some of these words are shrunk because of the stemming step during data cleaning. Each word therefore represents a meaning, such that the token "batteri" for instance captures the words "battery" and "batteries". Turning to the graph examination, one can observe an interesting change in the most popular words between 2013 and 2021. Indeed, most frequent tokens employed at the beginning of the period - such as "wast", "water", "environment", "land", "cost" - are generic and descriptive of environmental stakes. At the end of the period, the most frequent tokens have become "partnership", "europe", "institutionalis", "assess", "impact", "batteri", "substanc", which suggest a shift towards a more action-oriented approach. Of course, one should remain cautious about extrapolation of the 2021 semantics, given the sample of reports is not yet comprehensive of future publications for this year and is therefore biased towards the topics tackled from January to March. Nonetheless, this evolution in the narrative seems to be correlated with an increase in the frequency of words related to *evidence-based policies*, such as "assess", "impact", "collect", "studi", "measure" and "inform" which are all descriptive of a process of policy empirical evaluation. To be more precise on this important aspect of the analysis, I will observe the evolution of *evidence-based policies* semantics in the next subsection. One can finally notice that the word "tax" was increasingly employed until 2017, when it suddenly disappeared from reports. This is a surprising feature of this corpus given the popularity of carbon taxation within the very influential world of economics. This sudden "evaporation" cannot be linked to a specific political event and constitutes an anomaly in a context where some European policy-makers have struggled, faced social movements[2], and ultimately failed to implement such policy tool. Moreover, the "energy" topic, to which carbon taxation is very often associated, has remained almost

[2]The *Gilets Jaunes* movement is particularly illustrative of the prevalence of carbon taxation in France's political debates in 2018.

Figure 3: Evolution of Top Words in the European Commission Corpus (2013-2021)



Note: This graph plots the share that each top word represents in the corpus of tokens for the European Commission reports, by year. These tokens are selected because they are amongst the top 10 most used words during at least one time period.

Reading: The token "partnership" shows a peak in use in 2021 as it represents more than 1% of the corpus, stop words removed. Furthermore, it features the highest increase in usage between 2013 and 2021.

steadily evoked in the European Commission reports. This suggests that the Commission policy perspectives on energy has shifted over this period.

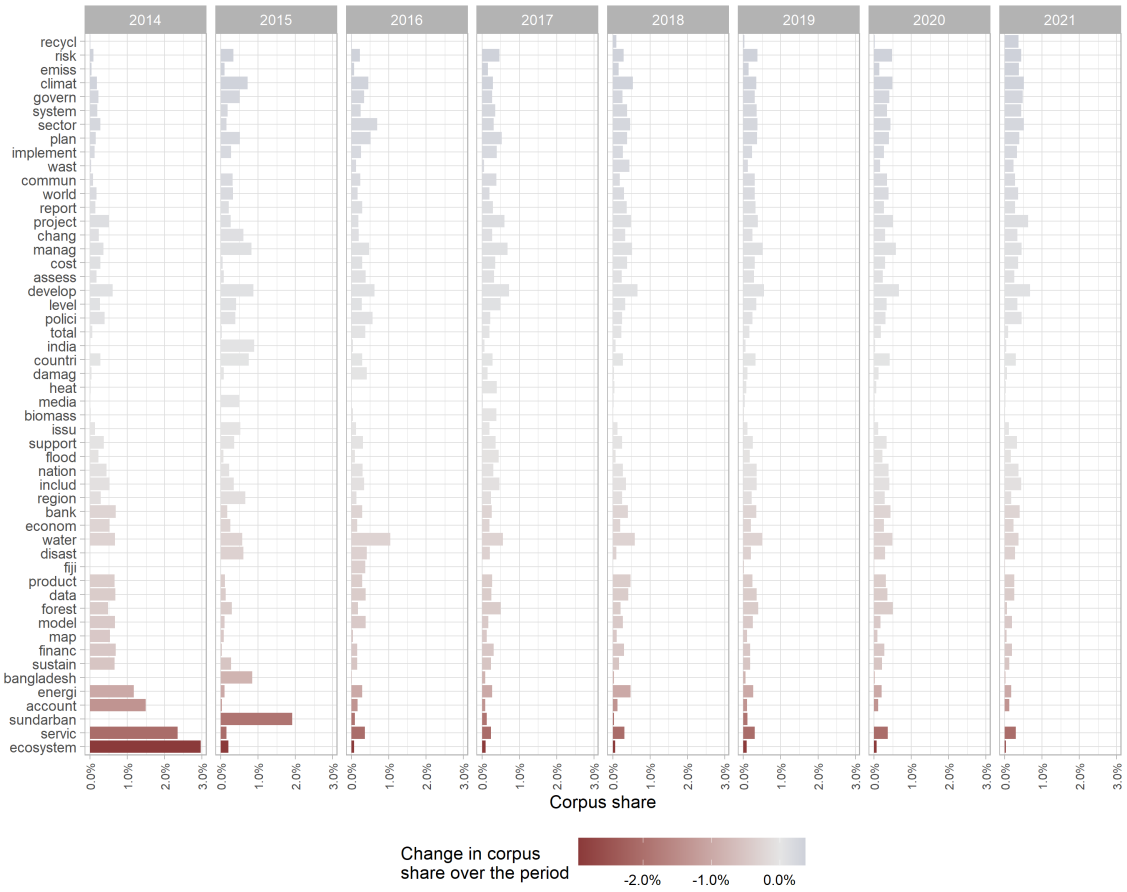
The second graph, Figure 4, presents the evolution of semantics for the World Bank 783 reports published between 2014 and 2021. One can notice that in contrast to the European Commission, the evolution of words frequency is only marked negatively. Five tokens that were very representative of the policy paradigm at the beginning of the corpus, namely "ecosystem", "service", "sunbardan", "acount" and "energi", (almost) disappeared from the words used in 2021. Other words such as "sustain", "financ", "map", "model", "forest", "data", "product", "disast", "water", "econom" and "bank" are also found to be decreasingly used over the period. Given the diversity of meanings to which these words can be linked, and their almost steady use after 2014, a cautious examination would suggest that the observed pattern only results from new topics covered since 2015.

However, the study of words which remained constantly used provides an interesting illustration of the policy paradigm that the World Bank has continuously applied to environmental topics over the covered period. A group of words reminds of the development bank nature of the institution: "sector", "plan", "implement", "project", "manag", "cost", "develop", "support". Words directly describing environmental issues are "recycl", "risk", "emiss", "climat", "wast", "damag", "heat", "biomass", "flood", "water". These words suggest that the World Bank is primarily concerned by two environmental topics: waste management, that involves infrastructure development projects, and climate change, for which they have a very economic approach based on risk and cost. Three words also indicate that the World Bank is sensitive to the communication around its actions: "commun", "media" and "support". Overall, this semantic analysis is consistent with the mission of the World Bank which is to provide financial and technical support to governments in the development of their infrastructures.

The evolution of the OECD reports semantics is finally graphed in Figure 5, page 25. From first sight, the depth of colours suggests important changes in the topics and meanings covered through times in this corpus. Generic descriptive tokens such as "adapt", "level", "chang", "sea", "loss", "impact", "econom", and "climat" have been progressively removed from reports. In contrast, as in the European Commission publications, new words that are action-oriented have been increasingly used. This is the case of "invest", "build', "financ", "price", "energi", "sector", "tax", "innov", "project" and "technologi". The expansion of these meanings in the OECD semantics could mean that the institution has positioned itself in favour of *market-based* policies to tackle environmental issues. One can also notice that words related to the evaluation of policies, such as "impact", "assess", "polici", and "model" represent a lesser share of the corpus at the end of the period than at its start. The evolution of top words used in the OECD reports illustrates the prominence of an economy-centered narrative developed by the Parisian office when it comes to proposing solutions to environmental issues.

The bag of words used by the three institutions has evolved in different directions over time. The European Commission and the OECD both have abandoned the use of very generic words which were positioning them in a role of outsider or observer of environmental issues. They have nonetheless located in different clusters of the policy debates. The European Commission

Figure 4: Evolution of Top Words in the World Bank Corpus (2014-2021)



Note: This graph plots the share that each top word represents in the corpus of tokens for the World Bank reports, by year. These tokens are selected because they are amongst the top 10 most used words during at least one time period.

Reading: The token "ecosystem" shows a peak in use in 2014 as it represents approximately 3% of the corpus, stop words removed. Nonetheless, it features the greatest drop in popularity as it practically disappears from the corpus in the following years.

words that have emerged throughout this period suggest that they are advocating for thorough monitoring of policies and practices, related to "chemic" and "substanc", to update and improve European policies. Indeed, one of the major policy success of the European Union in the past decade is the development of the REACH regulation which delivers market authorisations to chemical substances. On the other hand, the OECD has increasingly published in favour of "invest"(ments) and "financ", as well as for the regulation through "price" mechanisms. The positioning of the two institutions is very likely the pure extension of their political powers. It is very hard for the European Commission to implement a carbon tax system, because of the sovereignty of its state members, whereas it can increase its role of chemical substances regulator with the *REACH* regulation. The OECD has no political power and only acts as advisor to countries and as a policy narratives shaper. One way to weight in the policy balance may therefore be to influence businesses and governments to work hand in hand in a similar direction. Hence, the institution may be trying to preserve an apparent political neutrality. Indeed, making concrete policy recommendations to government may be politically cleaving and not tolerated by member states, as it can question the efficiency of their own actions. This could explain the inclination of the institution in favour of market-based solutions - which in a neoliberal momentum of international politics can be viewed as politically neutral. Finally, the World Bank had remained consistent to its positioning as a development bank.

In this subsection, I have observed that the three supra-national entities are addressing environmental policies under distinct narratives, in perfect consistency with their political roles and powers. However, these three institutions claim for an unbiased use of science to inform and improve their political decisions. But to what extent do they actually rely on *evidence* to support their proposals? I address this question in the next section.

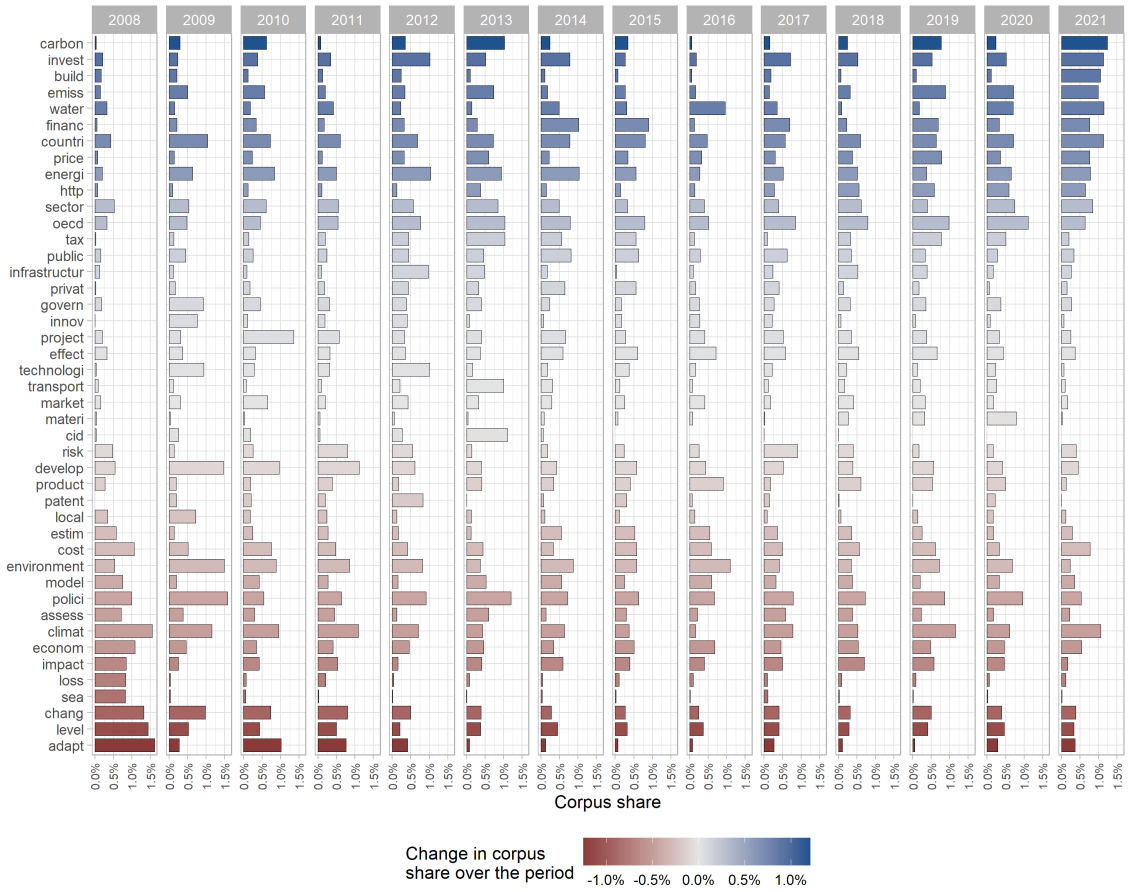
4.2 Evolution of the "Evidence-Based Policies" narrative in the three institutions corpus

In this section, I now turn to the analysis of the impregnation of the *evidence-based policies* paradigm in the corpus of reports from the OECD, the World Bank and the European Commission. As previously detailed in the methodology, I create a small corpus of tokens related to *evidence-based policies* based on eight articles.

Central to the following analysis is the assumption that the selected words are a good proxy for assessing the impregnation of the *evidence-based policies* paradigm. Words frequency in the three corpus, and their evolution through time, should therefore provide insights about the importance of this this narrative in the institutions discourses. In Figure 6, I plot for each political entity's publications, the cumulated share of these words per year. Comparison of the three graphs should be limited to the overlapping time period, which ranges from 2015 to 2021.

The European Commission and the World Bank both feature a slightly positive trend in the use of these words. The share that represents this semantic in the whole corpus is slightly greater in the European Commission documents than in the World Bank ones, respectively above 5% and around 4.5%. The trend is different in the OECD case. During the comparison time period, the use of words associated to *evidence-based policies* has declined from 7% to approximately

Figure 5: Evolution of Top Words in the OECD Corpus (2008-2021)

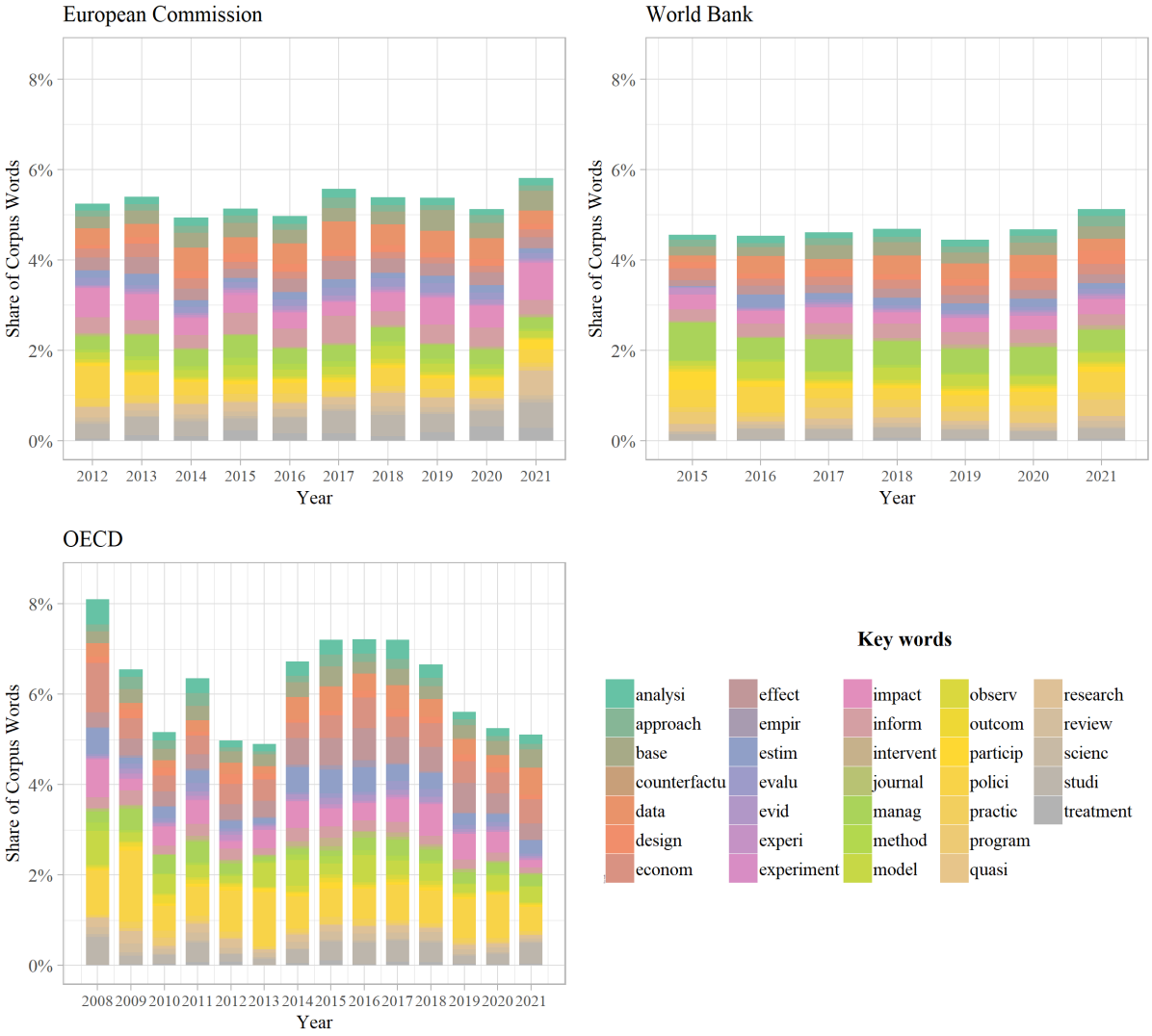


Note: This graph plots the share that each top word represents in the corpus of tokens for the World Bank reports, by year. These tokens are selected because they are amongst the top 10 most used words during at least one time period.

Reading: The token "carbon" shows two peaks in use in 2013 and 2021, where it respectively represented 1% and 1.3% of the corpus. The frequency of the word is irregular over the years, but features the greatest increase in popularity between 2008 and 2021

5%. However, if one considers the entire period over which documents could be retrieved, the impregnation of the paradigm has been constantly oscillating between a low 5% and a high 7%. Considering the year 2008 as an outlier, the OECD has infact experienced a temporary peak in the influence of the *evidence-based policies* paradigm from 2014 to 2018 before going back to its initial level, which is comparable to the European Commission and World Bank ones.

Figure 6: Evolution of the semantics associated to "Evidence-based" policies in the corpus



Note: The figures present for each institution corpus the cumulated share of *evidence-based policy* semantics, after stop words removal.

Reading: In 2008, 8% of the words used in OECD environmental policy reports were related to the the *evidence-based* paradigm.

The question is now to understand how substantial a 5% share is, once stop words have been removed? With no counterfactual, an answer to this question cannot be formulated convincingly. Indeed, on the one hand the approximate stability of this share across reports and years,

shows how implemented this semantics now is. Moreover, it implies a constant level of *evidence* input in these reports. On the other hand, 95% of the document narrative uses other kinds of semantics, such that the use of evidence may only be parsimonious and oriented to deliver a final argument in favour of a non-evidence based policy proposal. To make a sense of this share in the three institutions corpus, it would be interesting to compare the frequency of this list of words in environmental publications to their frequency in documents related to development, a policy field deeply embedded into an evidence-based approach.

Furthermore, the underlying meaning of words used above to analyse trends in the *evidence-based policy* paradigm is somewhat ambiguous. Indeed, evidence about environmental issues can be used to motivate and justify policy actions, but the cursor can also be put further as evidence can also be applied to policy evaluation and in turn guide policy decisions. Evidence can thus be invoked to discuss a problem's roots as well as its solutions. A complete *evidence-based* policy approach would thus imply relying on science both for understanding the stakes and assessing the efficiency of the different policy options. The later perspective is performed with the support of published policy impact evaluations.

Next, I therefore restrict the studied semantics to a set of words related to *policy impact evaluation*, as developed in the economics literature over the past twenty years. My goal is to understand how much the *credibility revolution* claimed by economists has persuaded these three institutions. These words constitute a proxy for how much the three institutions are relying on past evidence on the impact of different policies to guide their proposals. Given the World Bank is very close to the field of development economics, which is a major contributor to the methodological improvements realised in the field of impact evaluation over the past decades, a first guess approach would suggest that this institution is the most likely to be infused with this semantics.

However, as can be observed in Figure 7, the opposite is happening. Indeed, vocabulary related to impact evaluation of public policies is slowly growing over time in the World Bank reports but only represents between 0.5% and 0.75% of the whole corpus semantics. Particularly striking is the total absence of the word "experiment" which symbolises the golden standard method used in development economics to assess the efficiency of a program. In contrast, the European Commission features an increase in the use of impact evaluation semantics, which represents 1.5% of the tokens in documents published so far in 2021. Two words, "treatment" and "impact" drive this growth. The impact evaluation paradigm is therefore gaining more an more attention from the World Bank and European Commission environmental policy-makers, but the later entity has taken a substantial advance on that matter. Finally, the OECD graph shows an inverse u-shape trend similar to the one observed with the *evidence-based policy* set of words. From a peak of 1.5% of the corpus in 2018 to a low of 0.70% in 2021, this rapid decrease is mostly attributable to the decline of "impact", "treatment" and "evaluation".

In a nutshell, the European Commission seems to be increasingly and the more substantially relying on an evidence-based approach for assessing both the policy issues at stakes and the efficiency of policy candidates. The World Bank is also aligned in that direction, but the importance of policy impact evaluation in the institution semantics is surprisingly one percentage

Figure 7: Evolution of the semantics associated to counterfactual policy-evaluation in the corpus

Note: The figures present for each institution corpus the cumulated share of *policy impact evaluation* semantics by year, after stop words removal.
Reading: In 2008, 1.26% of the words used in OECD environmental policy reports were related to the the *policy impact evaluation* paradigm.

point lower than the European entity. Finally, the OECD reports feature a decline in the impregnation of the *evidence-based policy* narrative since 2018. This trend is concomitant to the rise of a bag of words related to infrastructure development, namely "invest", "finance", "energi" and "build", whom empirical impacts may not be easy to identify and quantify given the wideness of factors influencing them and that they affect in return. Because the OECD seems to be pushing for projects that are harder to evaluate quantitatively, they may mechanically reduce references to policy evaluation. This semantic and imperfect outlook at the trends in OECD, World Bank and European Commission reports suggests that the former, in contrast with the two later, is positioning itself on the action grounds with a decreasing interest in evidence.

At this stage of the analysis, the institutional differences in approach to environmental issues confirm the existence of political orientations. Entities do not neutrally base their policy agenda on the stakes identified within the academic literature. If this was the case, they would tackle similar topics under a very related semantics. Other factors therefore impact institutional prioritisation of policy issues. But what if albeit different topics covered, the three entities were still carefully relying on evidence when assessing the issue and potential actions related to each of them. One first answer to this question can be obtained by comparing the three entities prism on a similar topic. In the next section, I will thus move to a case-study approach in which I will compare the semantics and citations networks of three reports on *nature based solutions to coastal erosion*, each of them being released by one of the tree institutions.

4.3 Focus: Selection of evidence in three policy reports on *Nature-Based Solutions to Water Related Risks due to Climate Change*

The goal in this new section is to reproduce our former approach and expand it on a reduced corpus of three reports, each one of them published by one of the three institutions, tackling the potentials of nature-based solutions to water-related risks due to climate change. A starting point to the following analysis is the assumption that the three reports can be considered as approximately good counterfactual to each other given they cover a similar topic and have been released within a two-years period, in 2018 (World Bank) and 2020 (EC and OECD). To assess the strength of this comparability hypothesis, I compare words frequencies between reports in the following subsection. I then dissect the reports alongside three dimensions. First, I compare how much the semantics used in the reports is different from the underlying set of words used in their citations' abstracts. Second, I reproduce this approach with the *evidence-based policy* and *policy impact evaluation* group of tokens I used in the previous section. Finally, I examine citations networks that I consider being a proxy for institutions' selectivity in evidence.

4.3.1 Semantics Comparison: are the Reports comparable?

Before proper analysis, I test the veracity of the assumption that reports tackle the same topic. A first approach is obviously to compare titles, but this may be too generic to capture the inner orientations of each report. For this reason, I examine frequency of the most popular words in the reports, once stop words have been removed. I select the thirty most frequent words and plot them in Figure 8. Additionally, I fill bars by a colour capturing the normalised difference of

the word frequency to the mean frequency across reports[3]. When colours tones are very similar, this implies that the word use is very close between reports. In contrast, when a word's colour is very dark, it is an important element of distinction of the report corpus from the two others.

Distributions of words frequency are not precisely similar between reports, but orders of magnitude are often very close. If reports had been totally different, the plot would contain ninety different words, with positive frequency only for one report each time. Here, we retrieve a total of 46 words, which can presumably be interpreted as an approximate overlapping of the reports' key ideas.

Nonetheless, some words are very segmenting as they appear very dark on the graph. These words capture the decisive distinction in the narrative adopted by the three institutions. The European Commission is focused on an approach based on "resili"(ence), "mitig"(ation), "reduct"(ion) and "research". This is precisely the orientation of the source report, which is an assessment of the research and policy projects subsidised by the European Union to tackle these issues. The OECD is differentiating itself with the adoption of a macroscopic approach to the issue, symbolised by the use of tokens such as "nation", "countri", "ecosystem", "polici" and "adapt" which are aligned to the political orientations and role of the entity: generic and distant from concrete measures. Finally, the World Bank narrative particularity is in the employment of "fund", "invest", "financ" and "program" which perfectly fit with its development bank actions.

Even though the policy implications of the reports seems to be considered from a different angle by each institution, the words frequency analysis suggests that the definition of water-related risks is approximately identical. I will thus consider the hypothesis that the three publications tackle a similar topic as good enough to pursue my examination of policy-makers biases.

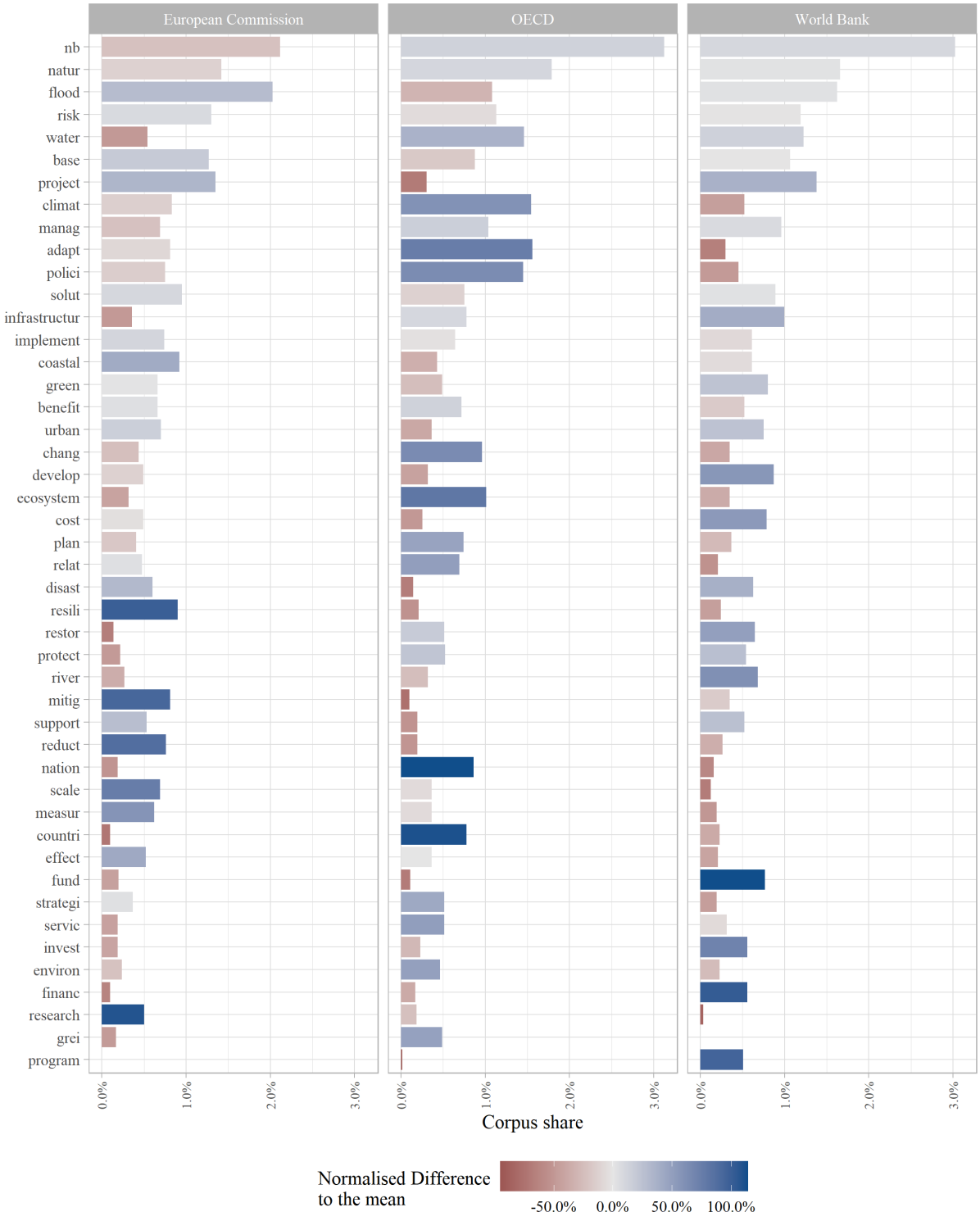
4.3.2 How Different are the Reports from their Citations Abstracts?

I now turn to the first pillar of the three reports cross-comparison: a study of the likely bias introduced during the evidence aggregating process. I propose a straight-forward approach to this question, as detailed in the methodology. Basically, I consider citations abstracts as a counterfactual for reports' semantics. The implicit assumption made here is that abstracts capture the semantics of the whole publication they summarise. If the three institutions reports consist in a pure and perfect synthesis of evidence, tokens frequencies should be very similar to those in abstracts. Comparing them allows to quantify closeness between reports and their sources.

Word by word results are presented in Figure 9, page 33. Additionally, I propose summary statistics of the graph in Table 5 to provide an aggregate picture of measured bias. The average absolute difference between words frequency in the report and in the abstracts, measured in percentage points, is of 0.53 for the European Commission, 0.4 for the World Bank and 0.37 for the OECD. This suggests that policy-makers introduce substantially more bias during the extrapolation of available evidence in the European Commission report than in the two other

[3]The formula is: $diff = \frac{freq_{word,report} - E[freq_{word}]}{freq_{word}}$ where the expected value is estimated using sample average.

Figure 8: Most Used Words and their Frequencies in the Reports



institutions reports.

Table 5: Average Difference in Word Frequency by Report

	World Bank	OECD	European Commission
Average Difference	0.4 p.p.	0.37 p.p.	0.53 p.p.
Standard Error	0.035 p.p.	0.037 p.p.	0.049 p.p.

Note: The table presents each report's average of the absolute value of the difference between words frequencies in the report and in the citations' abstracts
Reading: The average absolute difference in the World Bank report is of 0.4 percentage points, with a standard error of 0.035 - which implies that the difference is statistically different from 0.

If we turn to Figure 9 to understand the details of these statistics, we can identify the words generating this greater discrepancy between the European Commission report and the underlying abstracts. The words that are by far more frequently used in the report than in the abstracts are "flood", "natur", "climat", "risk", "base", "manag", "project", "polici", "adapt", "solut", "reduct", "resili" and "mitig". On the other hand, words that are much more frequently used in the abstracts than in the report are "water", "studi", "runoff", "wetland", "system", "flow", "river", "roof", "sediment", "groundwater". The clear pattern that this discrepancy illustrates is a difference in orientation between a report that is very much solution-oriented compared to abstracts that are descriptive of the issues at stake. This suggests that the policy proposed in the European Commission reports are not based on evidence, but only are the issues identified thanks to science. Considering words for which there exists an important gap in frequency between the report and abstracts in the World Bank and OECD publications, the former analysis seems to hold too but to a lesser extent. Indeed, the difference in frequency is for instance way less important in words such as "polici", "manag", "project", "adapt" - which could indicate that the OECD and the World Bank are putting forward solutions based on available evidence.

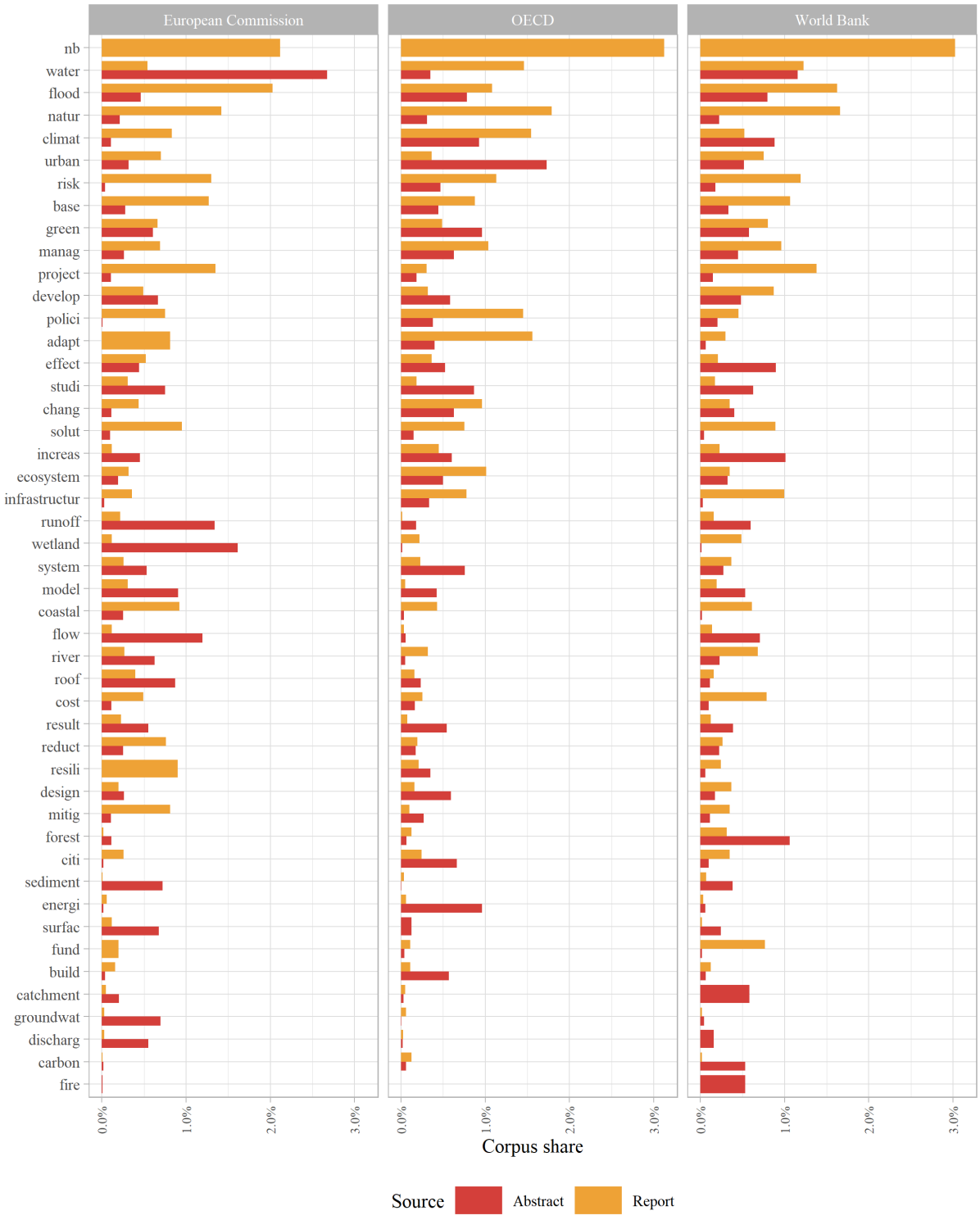
This first semantic analysis of the three reports hence indicates a bias in favour of solutions, whereas the underlying literature focuses more on stakes. In the next section, I will therefore examine if this signifies that the *evidence-based policies* narrative does not hold in reality.

4.3.3 Measure of the evidence-based paradigm penetration

To investigate the impregnation of the *evidence-based policy* and of the *policy impact evaluation* paradigms in the reports with regard to the abstracts, I replicate the methodology developed with all reports.

I start with the study of the *evidence-based policy* paradigm and use the same group of words as before to identify its importance in each corpus. Figure 10 compares the cumulated frequency of these group of words in the reports and in the abstracts. As mentioned in the previous subsection, the European Commission publication relies much more heavily on a semantic of "evidence" than the underlying abstracts. It is also surprising to note that the cumulated frequency of paradigm-related tokens is much smaller in the European Commission abstracts than in the

Figure 9: Difference in Words Frequencies Between Reports and Citations Abstracts

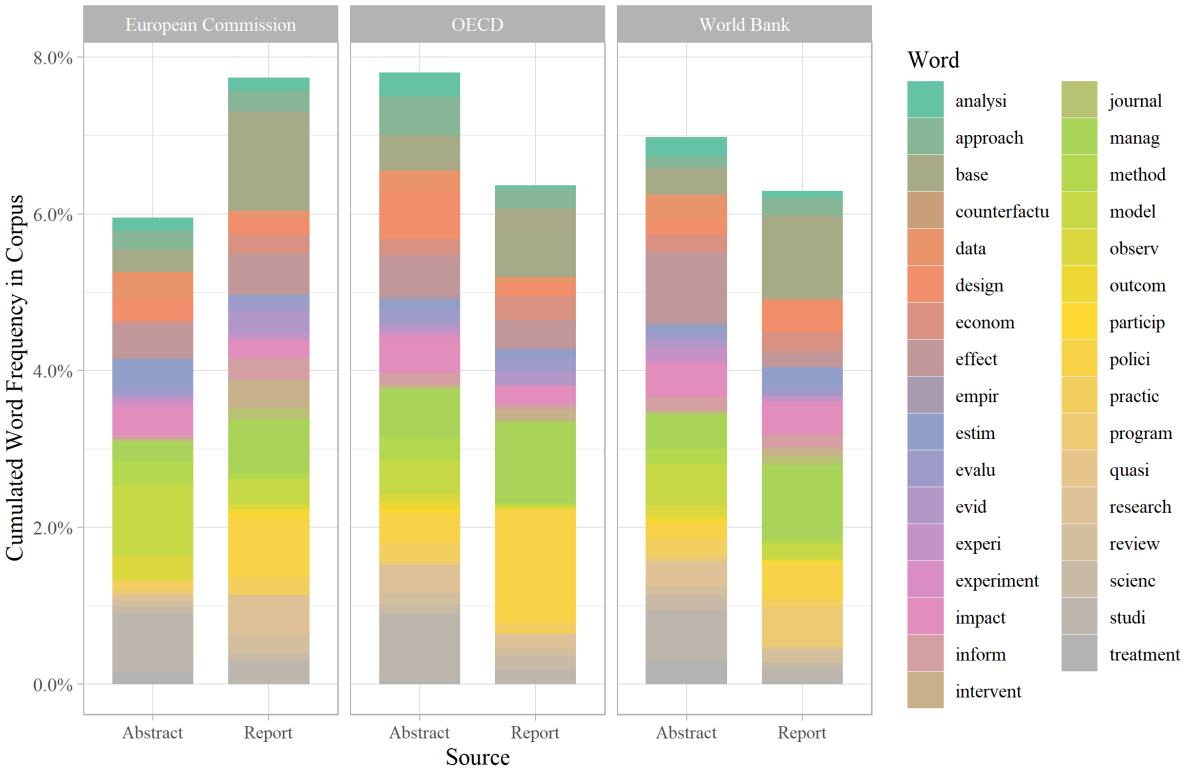


OECD and World Bank ones. This could be explained by a difference in underlying sources of evidence, a point I will cover further in the next subsection.

The European Commission corpus presents important differences between corpus and abstract in the use of "base", "polici" - much more frequent in the report, and "model" and "studi" - much more used in abstracts. In fact, this difference boils down to a difference in objective between the reports and its sources. Whereas the first is solution-oriented, the later are focused on the stakes. This finding could mean two things. Firstly, it could be that the report is only extracting information about solutions from cited sources and would leave aside discussions about causes. Secondly, it could also be that the report uses citations mostly to synthesise the issues but not so much to reuse their policy proposals. Under the second scenario, the *evidence-based policies* paradigm would break down.

In contrast, the World Bank and the OECD reports are less impregnated with an *evidence-based policies* semantics than the abstracts of sources they refer to. For these two institutions, reports puts much more emphasis on words like "program", "policy", "manag", and "base" than the abstracts. On the other hand, words like "model", "econom", "design" and "data" are more frequent in the abstracts. These semantics choices are consistent with the analysis proposed up to now. Indeed, institutions write these reports to come up with policy proposals, they may therefore only extract content from sources that is aligned with this objective.

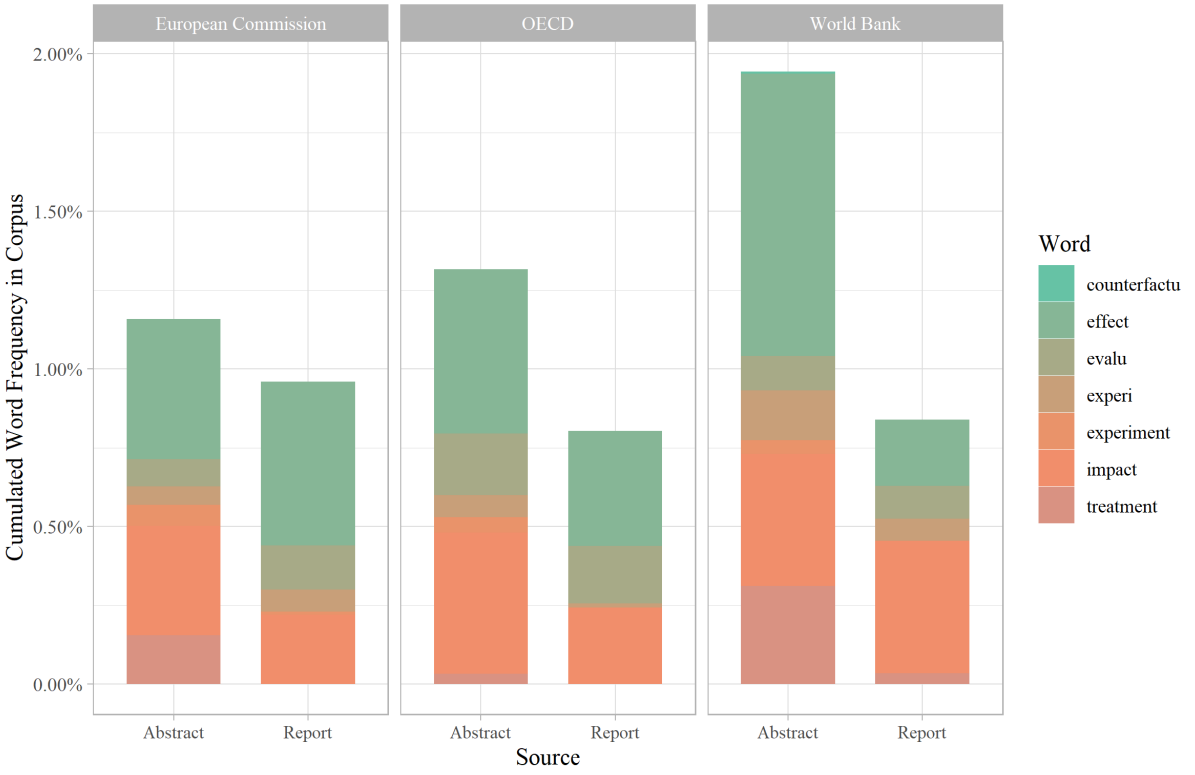
Figure 10: Difference in *Evidence-Based Policy* Semantics Between Reports and Citations Abstracts



Restricting attention to the importance of semantics related to *policy impact evaluations* in re-

ports now gives Figure 11. Results are different from the ones observed with *evidence-based* related words. Indeed, the three institutions now show a similar pattern ; where the use of tokens linked to impact evaluation is more prevalent in citations' abstracts than in reports. Regarding the rank of reports in terms of use of this semantics, the picture is similar to the conclusion from previous section on all reports. However, examination of abstracts' vocabulary frequency opens the room for reinterpretation of my former conclusions on the World Bank's publications. Indeed, the institution being, as a development bank, one of the standard bearer of impact evaluation, it was quite puzzling to previously observe that this do not translate into a preponderant use of related vocabulary. In the specific case of this report, it appears that the bank relies much more heavily on impact evaluations than the two other entities to support its publication. Nonetheless, the reduction by half of *impact evaluation* semantics from abstracts to report, primarily attributable to reduction in the use of "effect" and "treatment", suggests that the institution does not detail the methodology of underlying studies but directly discusses measured "impact".

Figure 11: Difference in *Policy Impact Evaluation* Semantics Between Reports and Citations Abstracts



How are these features of the semantics positioning the three reports in terms of potential bias in the use of evidence to build policy proposals? The three analytical perspectives adopted in this section show that policy-makers in charge of the three studied publications do not perfectly translate the content of underlying sources. Indeed, previous examinations suggest that *evidence* is primarily used to gather quality information on stakes. Nonetheless, the policy solutions proposed in reports may not rely as much on scientific evidence. Even though alternative theories may be better fitted to explain the different figures in this section, the clear conclusion

one can draw is that the institutions are not acting as pure aggregators of evidence. They apply an action-oriented filter to the sources they base their reports on. The main limitation of the proposed analytical approach is that it cannot clearly detect if policies proposed in reports are extracted from citations, or if only is the information on stakes gathered this way.

If doubts remain on the scientific rigour of policy proposals, it appears clear that sources of evidence are very extensive on details about the issues to be tackled. However, as Figure 9 illustrates, the semantic distribution of abstracts is very different from an institution to another, whereas difference in reports' vocabulary is not as important. This firstly indicates that the three institutions may be acting as vocabulary normalising filters ; where the filter is the action-oriented approach discussed before. The result is a surprisingly similar final report semantics. But how can investigations converge to similar conclusions starting from different perspectives on stakes? A potential answer is policy-makers biases stemming from their willingness to publish convincing policy proposals - that may therefore be leaning towards politically popular strategies. Differences in abstracts' words frequencies secondly raises another potential source of bias: selectivity in sources of evidence.

4.3.4 Selectivity in the network of citations

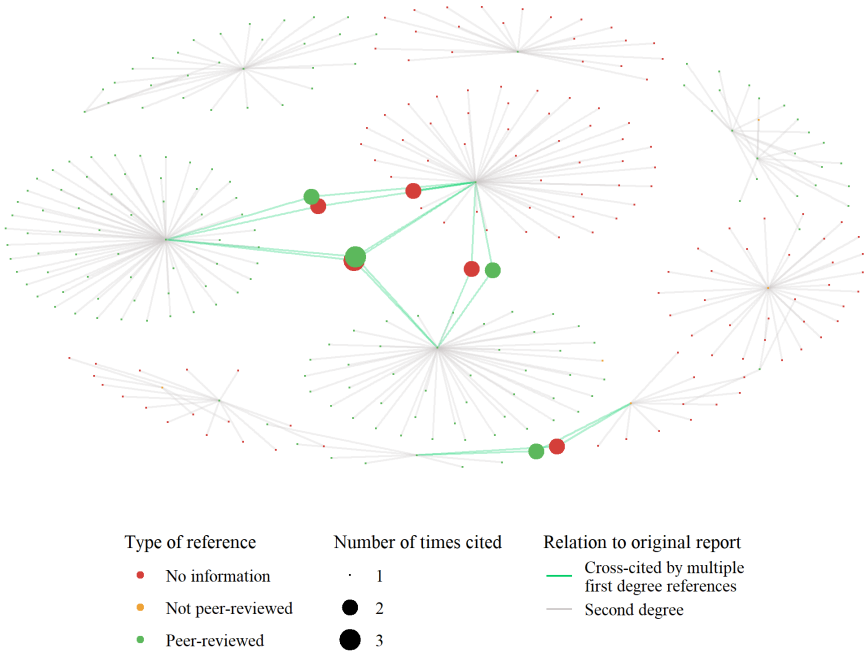
In this final subsection, I analyse the network of citations based on which the three reports have been constructed. My objective is to understand how representative these sources are of the true domain of evidence - which in reality is the comprehensive set of scientific literature focused on the policy issue. In a first best world, I would compare cited sources to an objective and comprehensive list of articles relevant to the discussed topic. Nonetheless, time constraints made this project unrealistic. As a second best approach, I therefore propose to compare the networks of references of the three reports, and look at how they interact.

If the reports are perfectly representative of the underlying literature, we may expect two network characteristics. First of all, the denser the network, the more comprehensiveness is to be expected, hence the closer to representativity of underlying science. Secondly, more connections in the network means that sources tackle a similar topic, such that the topic is clearly delimited, and ultimately that very connected sources are likely to be acknowledged for their quality. I thus carefully examine these two features of networks.

As a starting point to the analysis, I first plot each report's citation network independently in Figures 12, 13, and 14. The pictures show all existing interconnections between citations and how these citations may themselves cite similar references. The three networks are very different alongside the two dimensions of interest. Indeed, the World Bank one is characterised by a small number of citations, which implies a likely unrepresentative subset of evidence used to derive policy conclusions. Furthermore, sources are not very connected between one another, with an average 1.98 connection per node. As hypothesised above, this may indicate recent or low quality sources as well as a potentially loosely delimited topic. Therefore, characteristics of the World Bank network alongside the two dimensions of interest could indicate an unreliable evidence pool from which conclusions were drawn. This point is strengthened by the comparison of the World Bank network to the two other institutions'. The OECD and the European

Commission networks are much more complete in terms of density and interconnections. The average number of connection per node is of 2.1 for the OECD and 2.3 for the European Commission. Reports are hence more likely constructed on a cemented and well-defined pool of evidence. Furthermore, the two network sizes are much bigger than the World Bank one - 983 citations for the European Commission, 1,216 for the OECD, and 349 for the World Bank.

Figure 12: Network of the World Bank report inner references



To conclude the analysis, I look at connections between the three reports networks in Figure 15. The graph shows that the three reports are not built on the same evidence pool. Moreover, one observes a strong disconnection between the World Bank report and the European Commission one, whereas the OECD report seems to be positioned in-between the two others institutions. Interpretation of these relations between reports can take two directions. One possibility is that the initial objective and topic delimitation of each report does not perfectly overlap with the two others. The observed features would thus not be informative of any type of bias. The second scenario is that the three reports initially intended to tackle a similar topic but relied on different pools of evidence. In that case, the observed imperfect overlap between reports' citations would imply the existence of an important selection bias of evidence. Consequently, provided analysis and policy proposals would only be partial and unrepresentative of the entire stakes of nature-based solutions to water-related risks.

In this section, I have proposed a network approach to examine the magnitude of selection in references. From the graphs, one could observe an important difference between institutions in the *evidence* pool used as a foundation to the publications. This raises the question of the

Figure 13: Network of the OECD report inner references

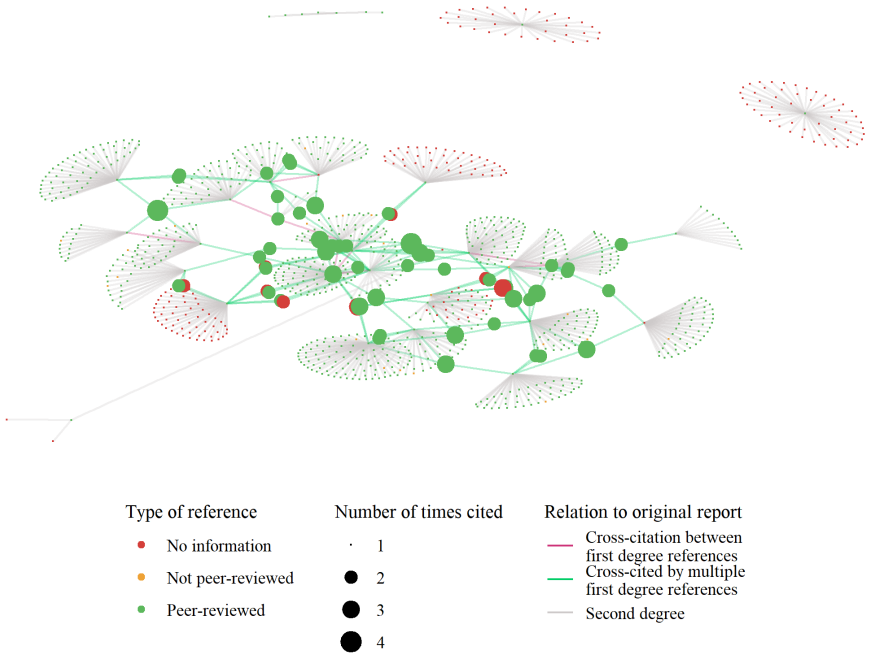


Figure 14: Network of the European Commission report inner references

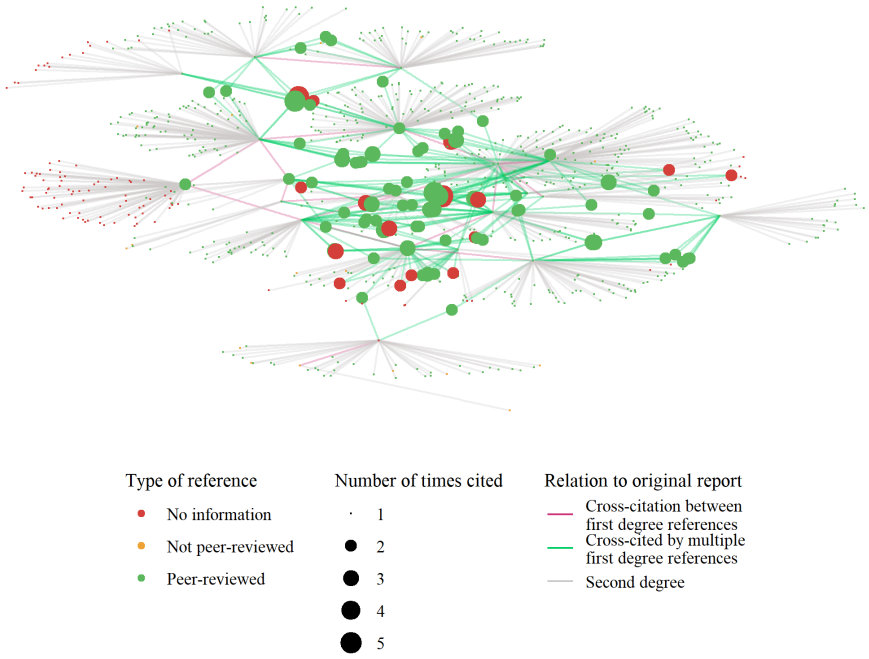
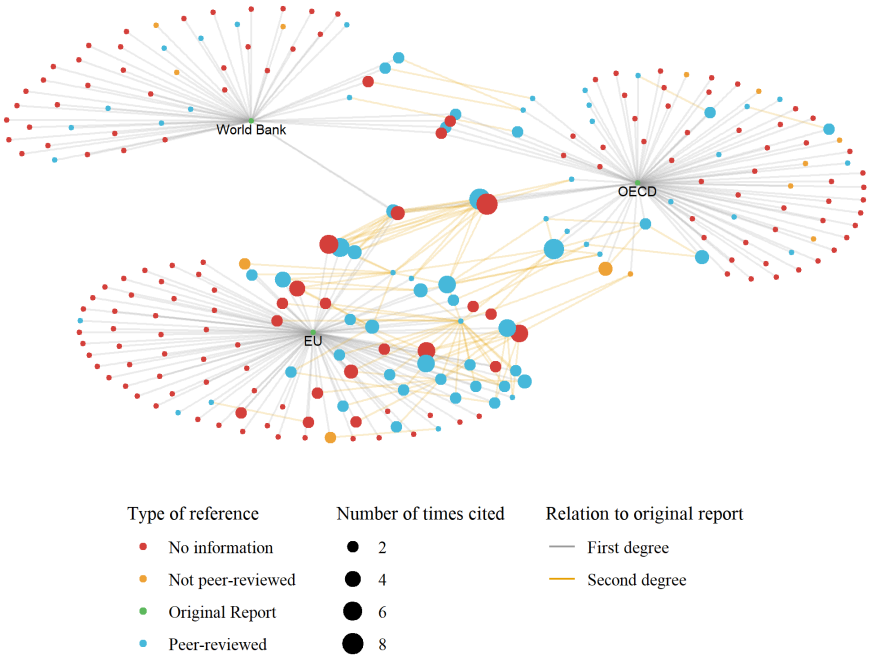


Figure 15: Shared network of references



representativity of used references with regard to the true domain of evidence actually existing. If these citations in fact encapsulate all the trends and results from the underlying literature, the issue of selectivity is mitigated. However, it is not clear with data at hand how much this may be. The *evidence-based policy* paradigm may therefore be weaker than the semantics used would suggest.

4.4 Limits to the analysis

Throughout the former analysis, I have assumed that my dataset was perfectly constructed and that the analysis that followed was therefore perfectly capturing the three institutions practices and potential biases. But a discussion on others' biases would be dishonest without transparent acknowledgment of the present material and conclusions limits. For this reason, I detail in this section all data assembling and cleaning steps, as well as visualisation choices that may bias my own results.

4.4.1 Data construction

I start with a review of the data construction steps that may be limiting the internal validity of the above conclusions.

All Reports database One first critic that can be formulated is the difference in number of reports that were retrieved by institution. The OECD sample hence cumulates the double par-

ticularity of being the smallest set and the set covering the longest time period. Consequently, some years are characterised by very few publications, as illustrated in Figure 1 - which in turn increases corpus semantics' sensitivity to the specific topic of each article. Concretely, this could mean that patterns identified in the European Commission and World Bank corpus are much stronger than the ones observed in the OECD. Another concern about the comparability of reports can be formulated. The semantic examination showed that topics covered were not the same, such that it may indicate political orientations of the institutions. However, this *effect* can only be isolated if everything else is held equal. But what if the three entities are just simply not meant to address same topics in a similar way given the very different nature of their respective political missions? I tried to limit this issue to the maximum by selecting filters on the three websites that would return somewhat similar kinds of documents. However, categories were not perfectly matching each other such that the European Commission Corpus contains documents relative to its own regulation bills, and the World Bank corpus contains working documents on ongoing infrastructure projects. I am therefore not comparing onions and eggs, but there is a little chance that it is about onions and shallots.

Three Reports database The big concern with the citation database is that it is incomplete. It does not contain all the references of references from initial reports - that is to say the second layer of citations is not exact. Unfortunately, not all sources cited by the three initial reports could be found with my automation technique. As a consequence, meta-data could not be downloaded and I did not have the time to manually seek for this information online. Table 6 shows that the problem is particularly strong for the World Bank report, and in turn partially explains the reason for a smaller network of citations. This limitation of meta-data availability extends to the set of abstracts used for the analysis. Indeed, abstract could not be found for all sources. To assess the magnitude of the resulting bias, the key question is the representativeness of retrieved abstracts with regard to the entire true population. A first element of answer is that retrieved abstracts may be biased towards scientific publication, because they are the most indexed on the two search engine used and are also the more likely to provide abstract. Grey literature, which includes working papers but also reports from NGOs, firms and public institutions may be underrepresented in the analysis. The final shortcoming of the citation database is the arbitrary choice to expand the network only up to the second degree connections. By doing so, I miss all the third degree connections that may occur between second degree references - which may wrongly bias my conclusion that the three reports are based on separated pools of evidence and surprisingly conclude to somewhat similar analyses. One could also argue that the network could have been expanded further to detect research topic clusters with natural language processing techniques and hence create a "counterfactual" network to the institutions' ones. This could be an interesting extension, that I will somewhat propose in my policy recommendations with the discussion of a research engine.

Table 6: Share of citations for which meta-data was retrieved, by report

OECD	European Commission	World Bank
44%	61 %	38%

4.4.2 Data cleaning

The data cleaning step is also very prone to bias creation. Nonetheless, very few steps were realised unless for mitigating these potential undesired effects.

All Reports database Two main data cleaning steps are performed on the reports database. The first of them is the triptych of tokenisation, stemming, and stop words removal. As explained in the methodology, these manipulations are performed using dedicated packages, and should not introduce bias to the analysis. The second step is the choice of both a metric for measuring words influence and of a threshold for selecting the more important of them. I decided to use token frequency - as in the number of occurrence of a word over the total number of words in the corpus - to rank words from the most influential to the least. The advantage of frequency in contrast with simple count is that it makes influence comparable across institutions and times, independently from the number of documents based on which the semantic analysis is made. However, the choice of the threshold above which words are considered the more influential is totally arbitrary and mostly answer to a concern of graph readability. Consequently, it could be that the inclusion of (epsilon) more words would have change the analysis. This is one of the main limitation of my approach: there exists no self-evident cutoff, either in terms of word frequency or of word rank, for inclusion in the analysis.

Three Reports database Data cleaning applied to the citation database follows the same procedure as presented just before for all the reports when constructing the counterfactual corpus semantics. It therefore suffers from similar shortcomings. In order to construct the networks, I matched citations based on names and therefore cleaned all titles to ensure that two articles would not be included in the database under different appellations. The manipulation of raw data here therefore reduces potential bias.

4.4.3 Visualisations and Analyses

The final element of the analysis during which bias may be introduced is through visualisation choices and proposed comments.

All Reports Throughout the examination of reports semantics, I assume that underlying reports are comparable. This assumption is in reality quite strong given the different nature of political role each entity is entitled to. The proposed analysis showed how each institution's mission was translating into its approach to environmental issues and more specifically into its policy positions. Another potential source of bias in the analysis of reports I did not account for is purely logistics. The lower number of reports by the OECD may signal a smaller human resource capacity to publish on a variety of topics. Either for quantitative reasons, the number of employee it takes to write a report, or for qualitative reasons, people need to understand the technicalities of an environmental topic to analyse it, this internal human capital issue may also be directing the institution choices in reports. The smaller the team and the more the semantics is prone to employee-introduced bias. Another implication is that even with the best willingness to be as unbiased as possible, the institution may be trapped in its incapacity to tackle many topics. In such scenario, the institution would be less "biased" than previously analysed, even

though another approach could argue that these topic choices under constraint may in fact be unveiling priorities and therefore institutional prism on stakes. A final limitation to my analysis is purely statistical. Indeed, I do not take into account confidence intervals around words frequencies. This omission was dictated by the willingness to keep the graphs that are already very loaded readable. Including a confidence interval would have made graphs very heavy. However, from the point of view of the analysis quality, this is a real shortcoming. It could indeed be that changes in words frequency across periods or reports are actually driven by few outlier sources. Hence, changes would potentially not be significant and the proposed analysis would break down.

Three Reports database In the analysis of the three reports, I use citations' abstracts as a counterfactual for the semantics in the reports. However, these abstracts may summarise parts of the underlying sources that are not of direct interest for policy-makers, and therefore avoid mentions to other sections holding policy-makers' attention. If that is the case, it would mean that the actual aggregation bias of evidence into policy reports is not as important as discussed in the analysis section. Moreover, I may have increased the bias towards scientific articles by including abstracts from the second-layer of the network. Stated differently, I included abstracts of second degree citations, which were most of the time academic publications. With this choice, I may be artificially inflating some words frequency in the corpus - given second-degree citations were not available for all reports' original citation. When proposing a network representation, I also propose a biased vision as the networks in fact continue way beyond the second degree connection. This implies I miss some connections between second-degree references - which would have potentially changed the picture on the root relationship between the pools of evidence based on which the three reports were written.

To conclude, the relevance of my choice to try to apply an imperfect *counterfactual* thinking approach to the issue at hand could be questioned. It actually is not straightforward how different networks can be compared under a counterfactual approach, neither how can corpus be examined through this lens. What should be the level of similarity to be expected between two sets of words for them to be considered as proposing aligned narratives on a topic? Furthermore, I could have opted for an analysis on the whole corpus instead of the more frequent words, to analyse how closely the entire distributions are.

5 Conclusion

This research is an attempt at measuring how much evidence-based are environmental policies from three supra-national institutions. The semantic analysis of reports from the OECD, the World Bank, and the European Commission shows that the three institutions are not acting as *pure aggregators* of scientific evidence. First of all, each institution proposes a different approach to environmental issues - a prism that is found perfectly aligned to the political role and agency of each entity. Secondly, the impregnation of an *evidence-based policies* paradigm has declined in the OECD publications, whereas it has risen for the two other institutions. Thirdly, the comparison of institutions semantics on a similar topic confirms the patterns observed at large scale. Moreover, it illustrates the action-oriented filter each institution applies to the underlying references used in the writing of reports. From very different source articles, the three institutions come up with very close final reports. Fourthly, the network analysis demonstrates the difficulty for policy-makers to build a set of evidence representative of the true underlying domain of scientific knowledge.

The research design was mainly focused on the demand-side, that is to say the processes through which policy-makers use science to build their policies. Nonetheless, the literature review discussed how academic knowledge is not particularly adapted in the first place for extrapolation into policy decisions. Consensus building is by nature an *asymptotic* process, and cannot be used as definitive truth. Methodological issues are particularly important in environmental policy evaluation, which limit the number of published studies and reduce the importance of scientific consensus on a majority of topics. For this reason, both the supply-side and demand-side are accounted for in the next chapter of policy recommendations.

As a conclusion to this research, it is important to remind that the reliability of its conclusions rely on the assumption that databases are representative of the phenomenon they intend to capture. Nonetheless, the data set may be prone to biases. For this reason, an important extension to this research could be performed by the improvement of the data quality. Another direction for extension would be with the creation of a counterfactual pool of scientific evidence on nature-based solutions. Hence, reports' semantics could be compared with academic knowledge. Finally, the set of techniques used to examine text corpus could be extended to more natural language processing method such as sentiment analysis. The latter would allow to analyse the neutrality and objectivity of institutions, as well as their level of optimism on environmental topics - potentially unveiling an ideologically oriented vision of the future.

6 Policy Recommendations

In this final chapter, I propose directions in which I believe there exists room for interesting and useful research and development in order to improve accessibility of science to policy-makers, with the ultimate goal of improving environmental policies. Recommendations' order is unrepresentative of their importance as they are all in fact complementary. Throughout my analysis, I have exposed the two main channels through which science may not prove useful to the improvement of environmental policies. The first is inadequacy in the knowledge proposed to policy-makers, the second is inappropriate approaches by policy-makers to summarising available academic results. My proposals are structured around these two pillars.

6.1 Recommendations for the improvement of evidence-building

The first pillar of measures proposed is focused on the supply-side with the structuring of a dynamic and evidence-accumulating academic field of environmental policies evaluation.

1. Incentivise and subsidise more environmental impact evaluations
The most pressing issue - partially illustrated by the low frequency of words associated to *impact evaluations* in the reports, but most rigorously discussed by Ferraro (2009) - is the lack of evaluations of implemented environmental policies. Many attempts at reducing human footprint have been implemented throughout the world, yet we miss the opportunity to understand the mechanisms that made them succeed or fail. This implies policy makers cannot effectively build on past experiences. An important research effort should be put in the field of environmental policies assessment, with an emphasis on the development of new methodologies to tackle hurdles mentioned in the literature review.

2. Harmonise concepts and create common analysis frameworks In their paper, Maki et al. (2018) notice the fragmentation of research on environmental policies between different disciplines who fail at interacting and connecting to provide a structured, complementary and thus complete overview of these program's impact. The authors propose the creation of a wiki-like platform where researchers would work hand in hand to come up with community-approved concepts. They argue that this collaboration could result in the publication of a dictionary that would simplify interdisciplinary collaboration as well as the sharing of knowledge with policy-makers. A starting point to this ambitious project could be to work on a systematic review of the key concepts used across papers and disciplines, along with their proposed definitions. This (important) study could result in the publication of a conceptual map proposing bridges between closely related concepts used in different disciplines.

6.2 Recommendations for a more accurate use of science in decision-making

The second pillar of measures proposed tackles the issues of the demand-side, and proposes the development of tools to facilitate policy-makers' work. Indeed, the literature and the proposed study of reports above show that practitioners' use of science may be altered by selectivity of evidence and biased understanding and aggregation of articles. The solution to this issue may

lie in the switch towards a systematic approach, that is to say a comprehensive screening and summary of scientific debates and consensus. However, policy-makers are not scientists so they do not have the resources to reach such precision. I argue that solutions coming from research should to be developed for them.

3. Develop a smart research engine to facilitate the screening process
The first direction in which a science could facilitate the screening work of policy-makers is obviously by structuring and classifying all existing works in policy-relevant categories. However, this may not be as easily implementable as said. Another possibility would be to work on the development of a machine-learning based research engine as suggested by Bannach-Brown et al. (2019). The tool could rely on semantic and network data, and would provide users with a list of articles and their probabilities of being relevant to the key words entered by the user. The user would then progressively pick the articles matching his/her request, and the research engine list would update accordingly, such that it would progressively delimit the policy boundaries the user is interested into. The research engine could first be developed and trained with academic articles openly available online, and centralised by core.ac.uk.

4. Develop Dynamic and Systematically Updated Meta-analyses, for results aggregation and quantification of uncertainty Another issue faced by policy-makers is to assess the external validity of studies. Is the positive impact of a policy observed in a neighbour country likely to be similar in their own context? One way to understand the factors that may affect the policy outcome is to use meta-analyses. The concept of meta-analysis is straightforward: the idea is to summarise quantitative results from different studies and to understand the factors that cause heterogeneous results. In their article, Maki et al. (2018) mention the need to develop what they call a dynamic and systematically updated repository of meta-analyses. Page & Moher (2016) talk about a "living cumulative network meta-analysis". According to them, the advantages are evident. First of all, it provides a global picture of the different effects a policy has in different contexts. If results are convergent or if variations are explained in the literature, it increases confidence in findings and their probability of being used by policy-makers to update their beliefs. Secondly, it provides policy-makers and citizens with a clear sense of the effects they can expect from a policy. Strategical choices are done in a much more informed way. Thirdly, local adaptations can be made thanks to the understanding of factors generating different results. Better policies are therefore implemented.

However, as enthusiastic as I am about these methods[4], it should be clearly said that this aggregation method is only as good as the underlying studies and most importantly relies on the very strong assumption of *exchangeability*. Stated simply, results are only informative if we assume that once identified factors are controlled for, policy outcomes should only be randomly varying across concepts. Furthermore, as Vivalt & Coville (2017) show, policymakers are often very interested in the programs details to design their own ; however as explained by Maki et al. (2018), articles often neglect reporting precise policy schemes, which imply they cannot be accounted as influencing factors in the meta-analysis.

[4]The academic work I realise to finalise my other degree is focused on these methods

Even though meta-analyses cannot provide definitive answers to policy-makers, they have the ability to provide with an interesting quantitative synthesis of academic impact evaluations. Furthermore, the adoption of Bayesian approach, as recently proposed by Meager (2019), Meager (2016), and Rubin (1981), can improve the readability and policy-relevance of meta-analyses. Indeed, the bayesian framework allows to simulate the (posterior) distribution of impact evaluations across contexts and therefore estimate the probability to observe a given impact magnitude in a new study.

As a continuation to the present thesis and to my Master's dissertation for the Paris School of Economics, I am working on the ambitious project of developing a prototype website that will provide users with a structured overview of environmental open data, through customisable dashboards, and of associated research, through a research engine of open science. The website will also propose policy-oriented meta-analyses[5]*. The objective is to create a hub, redirecting users to relevant evidence in the least biased way possible.*

[5] An alpha version is under development at the following link

References

Andrews & Kasy 2019
Andrews, Isaiah; Kasy, Maximilian: Identification of and correction for publication bias. In: *American Economic Review* 109 (2019), Nr. 8, pages 2766–94

Angrist & Pischke 2010
Angrist, Joshua D.; Pischke, Jörn-Steffen: The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. In: *Journal of economic perspectives* 24 (2010), Nr. 2, pages 3–30

Bannach-Brown et al. 2019
Bannach-Brown, Alexandra; Przybyła, Piotr; Thomas, James; Rice, Andrew S.; Ananiadou, Sophia; Liao, Jing; Macleod, Malcolm R.: Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. In: *Systematic reviews* 8 (2019), Nr. 1, pages 1–12

Bennear & Coglianese 2005
Bennear, Lori S.; Coglianese, Cary: Measuring progress: program evaluation of environmental policies. In: *Environment: Science and Policy for Sustainable Development* 47 (2005), Nr. 2, pages 22–39

Brodeur et al. 2020
Brodeur, Abel; Cook, Nikolai; Heyes, Anthony: Methods matter: p-hacking and publication bias in causal analysis in economics. In: *American Economic Review* 110 (2020), Nr. 11, pages 3634–60

Chang & Li 2015
Chang, Andrew C.; Li, Phillip: Is economics research replicable? Sixty published papers from thirteen journals say'usually not'. In: *Available at SSRN 2669564* (2015)

Ferraro 2009
Ferraro, Paul J.: Counterfactual thinking and impact evaluation in environmental policy. In: *New directions for evaluation* 2009 (2009), Nr. 122, pages 75–84

Ferraro & Pattanayak 2006
Ferraro, Paul J.; Pattanayak, Subhrendu K.: Money for nothing? A call for empirical evaluation of biodiversity conservation investments. In: *PLoS Biol* 4 (2006), Nr. 4, pages e105

Frondel & Schmidt 2005
Frondel, Manuel; Schmidt, Christoph M.: Evaluating environmental programs: The perspective of modern evaluation research. In: *Ecological Economics* 55 (2005), Nr. 4, pages 515–526

Fujimura 1996
Fujimura, Joan H.: *Crafting science: A sociohistory of the quest for the genetics of cancer*. Harvard University Press, 1996

Greenstone & Gayer 2009
Greenstone, Michael; Gayer, Ted: Quasi-experimental and experimental approaches to environmental economics. In: *Journal of Environmental Economics and Management* 57 (2009), Nr. 1, pages 21–44

Ingold & Gschwend 2014
Ingold, Karin; Gschwend, Muriel: Science in policy-making: Neutral experts or strategic policy-makers? In: *West European Politics* 37 (2014), Nr. 5, pages 993–1018

Jasanoff 2007
Jasanoff, Sheila: Technologies of humility. In: *Nature* 450 (2007), Nr. 7166, pages 33–33

Maki et al. 2018
Maki, Alexander; Cohen, Mark A.; Vandenbergh, Michael P.: Using meta-analysis in the social sciences to improve environmental policy. In: *Handbook of sustainability and social science research*. Springer, 2018, pages 27–43

Meager 2016
Meager, Rachael: Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. In: *Manuscript: MIT* (2016)

Meager 2019
Meager, Rachael: Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. In: *American Economic Journal: Applied Economics* 11 (2019), Nr. 1, pages 57–91

Muthukadan 2011
Muthukadan, Baiju: *Selenium with Python*. `https://selenium-python.readthedocs.io/`. Version: 2011

Newig & Rose 2020
Newig, Jens; Rose, Michael: Cumulating evidence in environmental governance, policy and planning research: towards a research reform agenda. In: *Journal of Environmental Policy & Planning* 22 (2020), Nr. 5, pages 667–681

Page & Moher 2016
Page, Matthew J.; Moher, David: Mass Production of Systematic Reviews and Meta-analyses: An Exercise in Mega-silliness? In: *The Milbank Quarterly* 94 (2016), Nr. 3, pages 515

Peplow 2014
Peplow, Mark: Social sciences suffer from severe publication bias. In: *Nature News* (2014)

Rubin 1981
Rubin, Donald B.: Estimation in parallel randomized experiments. In: *Journal of Educational Statistics* 6 (1981), Nr. 4, pages 377–401

Saltelli & Giampietro 2017
Saltelli, Andrea; Giampietro, Mario: What is wrong with evidence based policy, and how can it be improved? In: *Futures* 91 (2017), pages 62–71

Shwed & Bearman 2010
Shwed, Uri; Bearman, Peter S.: The temporal structure of scientific consensus formation. In: *American sociological review* 75 (2010), Nr. 6, pages 817–840

Sutherland et al. 2004
Sutherland, William J.; Pullin, Andrew S.; Dolman, Paul M.; Knight, Teri M.: The need for evidence-based conservation. In: *Trends in ecology & evolution* 19 (2004), Nr. 6, pages 305–308

Vivalt 2015
Vivalt, Eva: How much can we generalize from impact evaluations? In: *Journal of the European Economic Association* (2015)

Vivalt & Coville 2017
Vivalt, Eva; Coville, Aidan: How do policymakers update? In: *Unpublished manuscript, Berkeley, CA: University of California, Berkeley* (2017)

Ynnig 2020
Ynnig: *REST API - Crossref*. `https://www.crossref.org/education/retrieve-metadata/rest-api/`. Version: Apr 2020

A Codes used to create the report databases

Creation of the three reports dataset

EU reports

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import NoSuchElementException,
    TimeoutException
from bs4 import BeautifulSoup
import re
import requests
from tqdm.notebook import tqdm
import numpy as np
from pdfminer.high_level import extract_text
import time
import pandas as pd
import glob
import os
import pickle
from dataclasses import make_dataclass


options = webdriver.ChromeOptions()
options.add_experimental_option("prefs", {
  "download.default_directory": r"/home/yann/Documents/Projets/memoire/01.
    data/raweu",
  "download.prompt_for_download": False,
  "download.directory_upgrade": True,
  "plugins.always_open_pdf_externally": True,
  "safebrowsing.enabled": True
})
driver = webdriver.Chrome(options=options)
row_start = 1
result = 0
save_path = '/home/yann/Documents/Projets/memoire/01.data/raw_sources/
    eu_reports'


def latestFile():
    list_of_files = glob.glob('/home/yann/Documents/Projets/memoire/01.data
    /raweu/*.pdf') # * means all if need specific format then *.csv
    return(max(list_of_files, key=os.path.getctime))



reportRow = make_dataclass('Report',[('Title',str),
('pdfUrl',str),
('Year',int),
('Topics',int),
```

```
('Text',str)])

euReports = pd.DataFrame(columns=['Title','pdfUrl','Year','Topics','Text'])

while result < 582:
    path = f'https://op.europa.eu/fr/browse-by-subject?p_p_id=
    eu_europa_publications_portlet_pagination_PaginationPortlet_INSTANCE_eYu9jIuZAUpO
    &p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&facet.collection=EUPub&
    facet.collection=EUSummariesOfLegislation&facet.documentFormat=PDF&facet
    .studies=general&facet.author=RTD,ENV&facet.language=ENG&facet.eurovoc.
    domain=52&selectedSubjectId=52&elementType=0&sortBy=PUBLICATION_DATE-
    DESC&SEARCH_TYPE=BROWSE_BY_SUBJECT&QUERY_ID=199592665&&facet.language=
    ENG&facet.language=ENG&facet.language=ENG&facet.language=ENG&facet.
    language=ENG&facet.language=ENG&facet.language=ENG&facet.language=ENG&
    facet.language=ENG&resultsPerPage=50&startRow={row_start}&QUERY_ID
    =199592665'

    driver.get(path)
    html = driver.page_source
    clean_html = BeautifulSoup(html)

    nbResults = len(clean_html.find_all('li', {'class':'list-item first
    clearfix row'}))

    for row in tqdm(range(nbResults)):
        # Adding + 1 to the result for the while loop
        result += 1
        date = int(re.sub('\s+Publié:\xa0|-.+\n\s+','',clean_html.find_all(
    'time')[row].text))
        title  = clean_html.find_all('span',{'class':'result-name'})[row].
    text
        topics = ', '.join([subject.text for subject in clean_html.find_all
    ('li', {'class':'list-item first clearfix row'})[row].find_all('li',{'
    class':'hidden-xs list-item col-md-12 mt-2'})[0].find_all('a')])

        # Download
        link = clean_html.find_all('li',{'class':'list-item first filetype
    PDF'})[row].find_all('a')[0]['data-uri']
        path = 'https://op.europa.eu/'+link
        driver.get(path)

        # Importing pdf to python
        if row == 0:
            time.sleep(5)
            while glob.glob('/home/yann/Documents/Projets/memoire/01.data/
    raweu/*') is None and not bool(re.search(".crdownload$", latestFile())):
                time.sleep(2)

            latest_file = latestFile()

        else:
            wait = 0
            while bool(re.search(".crdownload$", latestFile())) or
    latestFile() == latest_file:
```

```
                time.sleep(1)
                wait += 1
                if wait == 60:
                    print('Too slow download or no download, jumped to next
    ')
                    pass
            latest_file = latestFile()

        # Importing text
        text = extract_text(latest_file)

        # Creating row
        row = pd.DataFrame([reportRow(title,link,date,topics,text)])
        euReports = euReports.append(row,  ignore_index = True)

        with open(save_path,'wb') as pckl:
            pickle.dump(euReports,pckl)


        # Setting a break
        kitkat = np.random.randint(5,10)
        # Taking the break
        time.sleep(kitkat)

    row_start += 51
    kitkat = np.random.randint(15,20)
    # Taking the break
    time.sleep(kitkat)
```

World Bank Reports

```
# Importing libraries
import requests
import pandas as pd
from dataclasses import make_dataclass
import pickle
from tqdm.notebook import tqdm
import re

# Defining the url for the api
def url(nb):
    return(f"http://search.worldbank.org/api/v2/wds?format=json&fl=
    abstracts,docdt,docna,docty,dois,txturl,pdfurl,subtopic,teratopic,theme&
    docty_exact=Report&lang_exact=English&teratopic_exact=Environment&rows={
    nb}&srt=docdt&order=desc")

# First retrieving the total number of results
nb = 1
response = requests.get(url(nb)).json()
nb = response['total']

# Downloading all meta-datas:
response = requests.get(url(nb)).json() # we've updated nb
print(response['rows'] )
```

```

reportRow = make_dataclass('Report',[('Title',str),
('TextUrl',str),
('Topic',str),
('Subtopic',str),
('Year', str),
('Text',str)])

def makeRow(row):
    try:
        text =  requests.get(row['txturl']).content
    except Exception:
        text = 'error'
    return(pd.DataFrame([reportRow(row['display_title'][0]['display_title'],
    row['txturl'],
    row['teratopic'],
    row['subtopic'],
    re.sub('-.+','',row['docdt']),
    text)]))


wb_reports = pd.DataFrame(columns=['Title','TextUrl','Topic','Subtopic','Text'])
save_path = '/home/yann/Documents/Projets/memoire/01.data/raw_sources/wb_reports'

for report in tqdm(response['documents']):

    if 'txturl' in response['documents'][report].keys():
        wb_reports = wb_reports.append(makeRow(response['documents'][report]), ignore_index = True)
    else:
        print(f'No text url for {report}')

    # Saving the file
    with open(save_path,'wb') as pckl:
        pickle.dump(wb_reports,pckl)
save_path = '/home/yann/Documents/Projets/memoire/01.data/raw_sources/wb_reports'

with open(save_path,'rb') as pckl:
    wb_reports = pickle.load(pckl)
```

OECD Reports

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import NoSuchElementException, TimeoutException
from bs4 import BeautifulSoup
```

```
import re
import requests
from tqdm.notebook import tqdm
import numpy as np
from pdfminer.high_level import extract_text
import time
import pandas as pd
import glob
import os
import pickle
from dataclasses import make_dataclass


options = webdriver.ChromeOptions()
options.add_experimental_option("prefs", {
  "download.default_directory": r"/home/yann/Documents/Projets/memoire/01.data/rawpdf",
  "download.prompt_for_download": False,
  "download.directory_upgrade": True,
  "plugins.always_open_pdf_externally": True,
  "safebrowsing.enabled": True
})
driver = webdriver.Chrome(options=options)



save_path = '/home/yann/Documents/Projets/memoire/01.data/raw_sources/oecd_reports'

# get total number of pages:

page = 1
path = f'https://www.oecd-ilibrary.org/environment-and-sustainable-development/oecd-environment-policy-papers_23097841?page={page}'
driver.get(path)
html = driver.page_source
clean_html = BeautifulSoup(html)
pages = int(clean_html.find_all("div", {"class": "paginator"})[0].find_all('a')[-2].text) # getting total number of pages

def latestFile():
    list_of_files = glob.glob('/home/yann/Documents/Projets/memoire/01.data/rawpdf/*') # * means all if need specific format then *.csv
    return(max(list_of_files, key=os.path.getctime))



reportRow = make_dataclass('Report',[('Title',str),
('pdfUrl',str),
('Year',int),
('Text',str)])


oecdReports = pd.DataFrame(columns=['Report','pdfUrl','Year','Text'])
```

```
57
58 for page in range(1, pages + 1):
59     print(page)
60     # Opening page
61     path = f'https://www.oecd-ilibrary.org/environment-and-sustainable-
   development/oecd-environment-policy-papers_23097841?page={page}'
62     driver.get(path)
63
64     # defining material for title/date extraction
65     html = driver.page_source
66     clean_html = BeautifulSoup(html)
67
68     # Getting the number of results:
69     nbResults = len(clean_html.find_all("div", {"class": "row panel"}))
70
71     # Downloading the results
72     for doc in tqdm(range(1, nbResults)):
73         # Downloading the file
74         driver.find_element_by_xpath(f'//*[@id="bellowheadercontainer"]/div
   /div[4]/div[3]/div[{doc}]/div[2]/ul/li/a').click()
75         # Importing pdf to python
76         # Waiting for the doc to be imported
77         if doc == 1:
78             while glob.glob('/home/yann/Documents/Projets/memoire/01.data/
   rawpdf/*') is None:
79                 time.sleep(2)
80
81             latest_file = latestFile()
82
83         else:
84             wait = 0
85             while latestFile() == latest_file:
86                 time.sleep(1)
87                 wait += 1
88                 if wait == 30:
89                     print('To slow download or no download, jumped to next'
   )
90                     pass
91             latest_file = latestFile()
92
93         # Importing text
94         link = clean_html.find_all("div", {"class": "row panel"})[doc].
   find_all('a',{'class':'action-pdf'})[0]['href'] # link
95         text = extract_text(latest_file)
96         title = clean_html.find_all("div", {"class": "row panel"})[doc].
   find_all('strong')[0].text
97         year = re.sub('^\d+ [a-zA-Z]+ ',"",clean_html.find_all("div", {"
   class": "row panel"})[doc].find_all('strong')[2].text)# year
98
99         # Creating row
100        row = pd.DataFrame([reportRow(title,link,year,text)])
101
102        oecdReports = oecdReports.append(row,  ignore_index = True)
103
```

```
104         with open(save_path,'wb') as pckl:
105             pickle.dump(oecdReports,pckl)
106 
107 
108         # Setting a break
109         kitkat = np.random.randint(5,10)
110         # Taking the break
111         time.sleep(kitkat)
112 
113         with open(save_path,'wb') as pckl:
114             pickle.dump(oecdReports,pckl)
```

Nature-Based solutions reports

References extraction from PDFs

The first step is to retrieve references in the three reports pdfs.

```
# Data extraction pdf
# install.packages("pdftools")
# install.packages("RJSONIO")
library(pdftools)
library(tidyverse)
library(tabulizer)
library(jsonlite)

folder_pdf <- "D:/OneDrive - sciencespo.fr/environmental_policy_tool/01. literature/07.oecd reports"

# 01. CLEAN FUNCTIONS -----------------------------------------------------

get_references <- function(page){
  t1 = unlist(str_split(str_remove_all(page, "References|(\\d{2} \\|)|(\\| \\d{2})"|"),"(\\.\r\n)"))
  t2 = sapply(1:length(t1), function(k) str_remove_all(t1[k],"\r\n"))
  t3 = sapply(1:length(t2), function(k) str_remove_all(t2[k],"\\[\\d+\\]"))
  t4 = sapply(1:length(t3), function(k){
    if(str_count(t3[k],"\\(\\d+\\)|\\(n\\.d\\.\\)|\\(Forthcoming\\)")>1){
      t5 = unlist(str_split(t3[k],"\\.[:blank:]{4,}"))
      t6 = c()
      for(k in 1:length(t5)){
        if(str_detect(t5[k],".[:alpha:]")){
          t6 = c(t6,t5[k])
        }
      }
      t6
    } else if(str_detect(t3[k],"^[:space:]*\\d+[:space:]*$")){
      # nothing
    } else {
      t3[k]
    }
  })
  return(unlist(t4))
}

get_author_eu <- function(source){
  str_split(str_trim(str_extract(source,"^.+(?=\\, (\\d{4}|Forthcoming),?)")),'\\.,| and | [:alpha:]{1},')
}
get_author <- function(source){
  str_split(str_trim(str_extract(source,"^.+(?=\\. \\d{4}\\.|\\(\\d+\\)|\\(Forthcoming\\)|Forthcoming|\\(n\\.d\\.\\)?)")),'\\., | and ')
}

get_date <- function(source){
  str_remove_all(str_extract(source,"\\. \\d{4}\\.|(\\(\\d+\\)|\\(Forthcoming\\))|\\(n\\.d\\.\\)|Forthcoming"),"\\(|\\)|\\.")
```

```
}
get_date_eu <- function(source){
  str_remove_all(str_extract(source,"\\, (\\d{4}|\\d{4}[:alpha:]{1}|
    Forthcoming),"),",")
}

get_title_long <- function(source){
  source = str_replace(str_trim(str_remove(source,"^.+(Forthcoming|\\d
    {4}|\\(Forthcoming\\)|\\(\\d{4}\\)|\\(n\\.d\\.\\))[:punctuation:] ")),"
    [:blank:]{2,}"," ")
  return(str_remove_all(source,"""|"))

}

get_title_long_eu <- function(source){
  source = str_replace(str_trim(str_remove(source,"^.+(\\, (\\d{4}|
    Forthcoming))[:punctuation:] ")),"[:blank:]{2,}"," ")
  return(str_remove_all(source,"""|"))

}


get_title <- function(source){
  loc = str_locate(source,"[^,.](Publishing|Reviews|Journal|Working Paper|
    https://|http://)" )[1]
  if(!is.na(loc)){
    source = str_sub(source, 1, loc[1])
    end_comma = str_locate(source,'[^,]+$')[1]
    if(end_comma>1){
      source = str_remove(str_remove(source,'[^,]+$'),',$')

    }
  }
  journal = get_journal(source)
  if(!is.na(journal)){
    loc = str_locate(source, paste0(", ",journal))[1]
    source = str_sub(source, 1, loc-1)
  }

  return(str_remove_all(source,"""|"))
}


get_journal <- function(source){
    str_extract(source, "(?<=, ).+(?=, Vol\\..+?)")
}

get_doi <- function(source){
  if(str_detect(source,"doi\\.org|DOI\\:|doi\\:")){
    str_remove(str_trim(str_remove(source,".+doi\\.org/|.+DOI\\:|.+doi\\:")
    ),
    "\\.$")
  }else{
    NA
```

```
  }

}

get_url <- function(source){
  str_remove(str_remove_all(str_remove(str_extract(source,
                                                   "(https://.+)|(http://.+)"),
                                       "\\(accessed on .+\\)"),
                            "[:blank:]*"),
  "\\.$")

}

create_list <- function(ref){
  source =   list("authors" = unlist(get_author(ref)),
       "date" = get_date(ref),
       "long_title" = get_title_long(ref),
       "raw"=ref)

  source = append(source,list("title"=get_title(source[["long_title"]])))
  journal = get_journal(source[["long_title"]])
  url = get_url(ref)
  doi = get_doi(ref)
  if(!is.na(url)){
    source = append(source, list("url"=url))
  }
  if(!is.na(doi)){
    source = append(source, list("doi"=doi))
  }
  if(!is.na(journal)){
    source = append(source,list("journal"=journal))
  }
  return(source)
}

create_list_eu <- function(ref){
  source =   list("authors" = unlist(get_author_eu(ref)),
                  "date" = get_date_eu(ref),
                  "long_title" = get_title_long_eu(ref),
                  "raw"=ref)

  source = append(source,list("title"=get_title(source[["long_title"]])))
  journal = get_journal(source[["long_title"]])
  url = get_url(ref)
  doi = get_doi(ref)
  if(!is.na(url)){
    source = append(source, list("url"=url))
  }
  if(!is.na(doi)){
    source = append(source, list("doi"=doi))
  }
  if(!is.na(journal)){
    source = append(source,list("journal"=journal))
```

```
  }
  return(source)
}

# 02. DATA CLEANING ----------------------------------------------------


# OECD
pages = c(26:29)
references = c()
for (page in pages) {
  out = extract_tables(file.path(folder_pdf, "Nature-based solutions to adapting to walter-related climate risks.pdf"),pages=page, guess = F,
                       area = list(c(90.17379,  45.44380, 787.96739, 547.94903)), encoding = 'UTF-8')
  if (page == 26) {
    vector = as.vector(out[[1]][,2])[-1]
  } else {
    vector = c(as.vector(out[[1]][,1]),as.vector(out[[1]][,3]))

  }



  for(el in 1:length(vector)){
    first = identical(vector[el-1], character(0))
    now = str_detect(vector[el+1],"(\\(\\d+\\)|\\(Forthcoming\\))|\\(n\\.d\\.\\)")
    # |^[:blank:]{0}$
    last = length(vector)

    if(first){
      # In the case where we need to start a new string
      string = vector[el]
    } else if(now | el==last){
      # In the case when now is the last
      string = str_trim(paste(string, vector[el]))
      if(string != ""){
        references = c(references, string) # Drop previous string in references
      }
      string = c() # Create new string for future iteration
    } else {
      # When in middle of a reference, just adds local string to reference string
      string = paste(string, vector[el])

    }
  }

}

list_nature_based = lapply(references, create_list)
```

```
write_json(list_nature_based,path=file.path(folder_pdf,"nature_based.json")
    , encoding = "UTF-8")
```

```
### WORLD BANK
```

```
pages = c(21:22)
raw_ref = c()
for (page in pages) {
  out = extract_tables(file.path(folder_pdf, "nature_based_world_bank.pdf")
    ,pages=page, guess = F,
                       area = list(c(90.17379,  45.44380, 787.96739,
    547.94903)), encoding = 'UTF-8')
    raw_ref = c(raw_ref,as.vector(out[[1]][,2]),as.vector(out[[1]][,4]))
}
```

```
references = c()
for(el in 1:length(raw_ref)){
  first = identical(raw_ref[el-1], character(0))
  before = ifelse(!first, raw_ref[el-1], "hehe")
  now = str_detect(raw_ref[el+1],"^$|\\. \\d{4}\\.|Forthcoming")
  # |^[:blank:]{0}$
  last = length(raw_ref)

  if(first){
    # In the case where we need to start a new string
    string = raw_ref[el]
  } else if(now | el==last){
    # In the case when now is the last
    string = str_trim(paste(string, raw_ref[el]))
    if(string != "" & before != "" ){
      references = c(references, string) # Drop previous string in
    references
    }
    string = c() # Create new string for future iteration
  } else {
    # When in middle of a reference, just adds local string to reference
    string
    string = paste(string, raw_ref[el])

  }
}
references
```

```
list_nature_based_wb = lapply(references, create_list)
write_json(list_nature_based_wb,path=file.path(folder_pdf,"list_nature_
    based_wb.json"), encoding = "UTF-8")
```

```
242
243 ### CITATIONS EUROPE
244 pages = c(42:49)
245 page = pages[1]
246 raw_ref = c()
247 for (page in pages) {
248   out = extract_tables(file.path(folder_pdf, "eu_nature_based.pdf"),pages=
      page, guess = F,
249                        area = list(c(90.17379,  45.44380, 787.96739,
      547.94903)), encoding = 'UTF-8')
250
251   is.matrix(out[[1]])
252   if(page == 42){
253     tempo = out[[1]][2:nrow(out[[1]]),]
254     remove(out)
255     out = list()
256     out[[1]]=as.matrix(tempo)
257
258   }
259   raw_ref = c(raw_ref,as.vector(out[[1]][,1]))
260
261
262 }
263 raw_ref
264
265 raw_ref[1:3]
266
267 references = c()
268 for(el in 1:length(raw_ref)){
269   first = identical(raw_ref[el-1], character(0))
270   before = ifelse(!first, raw_ref[el-1], "hehe")
271   now = str_detect(raw_ref[el+1],"^$|\\, (\\d{4}|\\d{4}[:alpha:]{1}|
      Forthcoming)\\,")
272   # |^[:blank:]{0}$
273   last = length(raw_ref)
274
275   if(first){
276     # In the case where we need to start a new string
277     string = raw_ref[el]
278   } else if(now | el==last){
279     # In the case when now is the last
280     string = str_trim(paste(string, raw_ref[el]))
281     if(string != "" & before != "" ){
282       references = c(references, string) # Drop previous string in
      references
283     }
284     string = c() # Create new string for future iteration
285   } else {
286     # When in middle of a reference, just adds local string to reference
      string
287     string = paste(string, raw_ref[el])
288
289   }
290 }
```

```
references


list_nature_based_eu = lapply(references, create_list_eu)
write_json(list_nature_based_eu,path=file.path(folder_pdf,"list_nature_
    based_eu.json"), encoding = "UTF-8")
```

I thus created a list/dictionnary of references that I export in a JSON format so that I can now open it in python to retrieve meta-data about it on the web.

Cross-Ref API code

The idea is first to retrieve meta-data from the cross-ref API. Below is the code for the functions and then the application.

Functions Definitions

```
import time
from crossref.restful import Works, Etiquette
import re
import json
import numpy as np
import pickle

class cross_ref():
    """ This class is used to retrieve data from crossref api """

    def __init__(self):
        agent = Etiquette('yann.collindavid@gmail.com')
        self.works = Works(etiquette=agent)

    def check_exists_doi(references):
        for index in range(len(references)):
            try:
                references[index]['doi']
            except KeyError:
                references[index]['doi']='no doi in oecd report'


    def check_exists_title(references):
        to_pop = list()
        for index in range(len(references)):
            if references[index]['title'][0] == '':
                to_pop = to_pop + [index]
        if len(to_pop) > 1 :
            to_pop.sort(reverse=True)
            [references.pop(poppy) for poppy in to_pop]
            print(f'Deleted elements {to_pop}, because of empty title')

    def valid_doi(reference,key):
        return(re.sub(r'\.$','',reference[key]))

    def valid_date(reference):
```

```
        date = reference['date'][0]
        if date in ['n.d.','Forthcoming']:
            return(2000)
        else:
            return(re.sub(' |[a-zA-z]','',date))

    def first_author(reference):
        return(re.sub(",.+","",reference['authors'][0]))

    def create_id(references):
        for i in range(len(references)):
            references[i]['id']=i

    def query_doi(self,reference, key):
        if reference[key] != 'no doi in oecd report':
            doi = cross_ref.valid_doi(reference, key)
            search = self.works.doi(doi)
            return(search)
        else:
            print('no do provided for {}'.format(reference['title']))
            return('no doi provided')

    def result_match_raw(raw,result,key_result):
        ti_res = result[key_result][0].lower()
        ti_raw =  raw['title'][0].lower()
        if ti_res in ti_raw or ti_raw in ti_res:
            return(True)
        else:
            return(False)

    def search_for_doi(self, reference):
        if reference['doi'] == 'no doi in oecd report':
            # searching for the doi
            title = reference['title'][0]
            author = cross_ref.first_author(reference)
            date =  cross_ref.valid_date(reference)

            searches = self.works.query(title).filter(from_online_pub_date=
date).sample(1).query(author=author)
            # keeping the first result of the search
            try:
                search = [item for item in searches]


                if len(search) >0:

                    if type(search) is list :
                        search = search[0]

                    # First scenario: we got it right (lucky us!)
                    if cross_ref.result_match_raw(reference, search,
key_result='title'):
                        doi = search['DOI']
                        print('direct match \n')
```

```
                        return(doi)

                    # Second scenario, we could not retrieve it directly, but there's a chance our source is within
                    # the references of the search results
                    else :
                        if 'reference' in search.keys():
                            for ref in search['reference']:
                                if 'title' in ref.keys():
                                    if cross_ref.result_match_raw(reference, ref, key_result='title'):
                                        print('match with title \n')
                                        if 'DOI' in ref.keys():
                                            return(ref['DOI'][0])
                                        else:
                                            return('no doi found')

                                elif 'volume-title' in ref.keys():
                                    if cross_ref.result_match_raw(reference, ref, key_result='volume-title'):
                                        print('match with volume title \n')
                                        if 'DOI' in ref.keys():
                                            return(ref['DOI'][0])
                                        else:
                                            return('no doi found')

                                elif 'unstructured' in ref.keys():
                                    if cross_ref.result_match_raw(reference, ref, key_result='unstructured'):
                                        print('match with unstructured  \n')
                                        if 'DOI' in ref.keys():
                                            return(ref['DOI'][0])
                                        else:
                                            return('no doi found')
                                else:
                                    print('no doi found after search  \n')
                                    return('no doi found')

                        else:
                            print('no doi found because no ref \n')
                            return('no doi found')
                else:
                    print('crossref returns nothing, no doi  \n')
                    return('no doi found')
            except json.JSONDecodeError:
                print('not a valid json file returned')
        else:
            print('{} already has DOI \n'.format(reference['title'][0]))


    def create_sourced_results(self, references):
        """
```

```
        This super function takes raw references from OECD reports as inputs,
        checks for existence of DOI, search for it if not existing,
        and finally returns complete metadata from CROSS-REF if DOI exists.
        """
        # 0 step 1: check doi existence and creation if not
        cross_ref.check_exists_title(references) # Gonna be needed for searches
        # 0 step create ids
        cross_ref.create_id(references)
        # 0 step 1: check doi existence and creation if not
        cross_ref.check_exists_doi(references)
        # First step: adding searched doi if existing
        for ref in references:
            kitkat = int(np.random.randint(1,10,1))
            time.sleep(kitkat) # Let's give cross-ref a little break
            if ref['doi'] == 'no doi in oecd report' and 'search_doi' not in ref.keys():
                print('Searching for DOI for {}'.format(ref['title'][0]))
                ref['search_doi'] = self.search_for_doi(ref)
            else:
                ref['search_doi'] = ref['doi'][0]
        # Second step: create sources
        info = dict()
        print('\n---------------------------------------------- \n')
        for ref in references:
            print('Adding source for {}  \n'.format(ref['title'][0]))
            kitkat = int(np.random.randint(5,10,1))
            time.sleep(kitkat) # Let's give cross-ref a little break
            id = ref['id']
            if ref['search_doi'] == 'no doi found' or ref['search_doi'] is None :
                info[id]= {'id': id,
                'result':'no doi found'
                }

            else :
                info[id]= {'id':id,
                           'result': self.query_doi(ref,'search_doi')}
        return(info)
```

Now that the functions are created, I execute them.

Execution

```
# Files directory
path = re.sub('/00.coding.+','',sys.path[0])
folder_raw_sources = path + '/01.data/raw_sources/'

with open(folder_raw_sources + 'nature_based.json', 'rb') as json_file:
    nature_based = json.load(json_file)

with open(folder_raw_sources + 'list_nature_based_eu.json', 'rb') as json_file:
```

```
    nature_based_eu = json.load(json_file)

with open(folder_raw_sources + 'list_nature_based_wb.json', 'rb') as
   json_file:
    nature_based_wb = json.load(json_file)

# Initialisation of our research environment
search = cross_ref()

print("""NATURE BASED SOLUTIONS REPORT \n
-----------------------------------------
""")
clean_nature_based = search.create_sourced_results(nature_based)

with open(folder_raw_sources + 'nb_oecd_source', 'wb') as f1:
    pickle.dump(nature_based, f1)

with open(folder_raw_sources + 'nb_oecd_crossref', 'wb') as f1:
    pickle.dump(clean_nature_based, f1)


print("""\n,
EU REPORT NATURE BASED \n
-----------------------------------------
""")
# Initialisation of our research environment
search = cross_ref()
clean_nature_based_eu = search.create_sourced_results(nature_based_eu)
with open(folder_raw_sources + 'nb_eu_source', 'wb') as f1:
    pickle.dump(nature_based_eu, f1)
with open(folder_raw_sources + 'nb_eu_crossref', 'wb') as f1:
    pickle.dump(clean_nature_based_eu, f1)

 print("""\n
WB REPORT NATuRE BASED \n
-----------------------------------------
""")
search = cross_ref()

clean_nature_based_wb = search.create_sourced_results(nature_based_wb)
with open(folder_raw_sources + 'nc_wb_source', 'wb') as f1:
    pickle.dump(nature_based_wb, f1)
with open(folder_raw_sources + 'nc_wb_crossref', 'wb') as f1:
    pickle.dump(clean_nature_based_wb, f1)
```

In a separate script I performed analysis of the result, and extracted references for which I had found no result on Cross-ref, such that I would scrap Web of Knowledge to find them.

```
path = re.sub('/00.coding.+','',sys.path[0])
folder = path + '/01.data/raw_sources/'

files = ['source_nature_based_pickle',
        'ref_nature_based_pickle',
        'source_nature_based_eu',
        'ref_nature_based_eu',
```

```
        'source_nature_based_wb',
        'ref_nature_based_wb']

databases = dict()
for file_ in files:
    with open(folder + file_, 'rb') as f1:
        databases[re.sub('_pickle','',file_)] = pickle.load(f1)

reports = ['nature_based',
           'nature_based_eu',
           'nature_based_wb']
data_to_wos = dict()
for report in reports:
    source = databases[names[0]+report]
    references = databases[names[1]+report]
    data_to_wos[report]={}
    data_to_wos[report]['nodoi'] = []
    data_to_wos[report]['noref'] = []

    for ref in references:
        if references[ref]['result']=='no doi found' or references[ref]['
    result'] is None:
            id = references[ref]['id']
            data_to_wos[report]['nodoi'].append(source[id])
        elif 'reference' not in references[ref]['result'].keys():
            data_to_wos[report]['noref'].append(references[ref])
        else :
            pass
```

Retrieving data on Web of Knowledge

Again, I start by setting up the functions before executing them.

Functions

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import NoSuchElementException,
    TimeoutException
import time
import re
import numpy as np
import pickle
from bs4 import BeautifulSoup
import codecs
import os
import pandas as pd


class scraping_wos():
```

```
    """
    This class is used to scrap things on databases from bib.cnrs using Selenium
    """

    def __init__(self):
        self.driver =  webdriver.Chrome("/usr/lib/chromium-browser/chromedriver") # Loading browser

    def first_author(reference):
        if reference['authors'][0] is not None:
            return(re.sub(",.+","",reference['authors'][0]))
        else:
            return(None)

    def clean_title(reference):
        return(re.sub(' +',' ',re.sub('\\?|(\\.)|-| and |&|(\\(.+\\))'-|,|\\||\\:||',' ',reference['title'][0])))


    def connect_cnrs(self, username, password):
        "This function takes username and password as input and logs into the bib.cnrs interface, on the database tab."


        self.driver.get("https://bib.cnrs.fr/") #going to bib.cnrs
        self.driver.find_element_by_tag_name('button').click() #click on the connect
        WebDriverWait(self.driver, 10).until(EC.element_to_be_clickable((By.XPATH,
        '/html/body/div[4]/div[2]/div/div/div[2]/button[1]'))).click() # Click on the janus connect button
        WebDriverWait(self.driver, 10).until(EC.presence_of_element_located((By.NAME,
        'j_username'))).send_keys(username) # It waits for the username tag to appear and then fills form
        self.driver.find_element_by_name('j_password').send_keys(password) # Fills password
        self.driver.find_element_by_tag_name('button').click() # Click on connect
        self.driver.implicitly_wait(5) # Wait for the page to load
        self.driver.find_element_by_xpath('//*[@id="ebsco_widget"]/div/div/nav/div/ul/li[3]/a').click() # Click on databases

    def connect_wos(self,username,password):
        "This function logs into the Web Of Knowledge database"

        wos_link = 'http://apps.webofknowledge.com/'
        self.driver.get(wos_link)
        time.sleep(3)
        self.driver.find_element_by_name('username').send_keys(username)
        self.driver.find_element_by_name('password').send_keys(password)
        time.sleep(2)
        self.driver.find_element_by_tag_name('button').click()
```

```
        try:
            new_session = self.driver.find_element_by_xpath('//*[@id="
    WoKerror"]/div/table[2]/tbody/tr/td[2]/p/a[1]')
            new_session.click()
        except NoSuchElementException:
            pass
        print('login successfull')

    def wos_get_advanced_search(self):
        self.find_element_by_xpath('/html/body/div[9]/div/ul/li[4]/a').
    click()

    def is_doi(reference):
        if 'result' in reference[0].keys():
            return(True)
        else:
            return(False)

    def clean_ref(ref, doi):
        if doi:
            title = ref['result']['title'][0]
            doi = ref['result']['DOI']
            id = ref['id']
            result = {'title': title, 'doi': doi, 'id': id}
            if 'author' in ref['result'].keys():
                if 'family' in ref['result']['author'][0].keys():
                    author = ref['result']['author'][0]['family']
                elif 'name' in ref['result']['author'][0].keys():
                    author = ref['result']['author'][0]['name']
                result['author'] = author
            return(result)
        else:
            title = scraping_wos.clean_title(ref)
            author = scraping_wos.first_author(ref)
            date = ref['date'][0]
            id =  ref['id']
            result = {'title': title, 'author': author,
                          'year': date, 'id': id}
            return(result)


    def item_search1(self, ref, field):
        self.driver.find_element_by_id('value(input1)').clear()
        self.driver.find_element_by_id('value(input1)').send_keys(ref[field
    ])
        self.driver.find_element_by_id('select2-select1-container').click()
        self.driver.find_element_by_css_selector('input.select2-
    search__field').send_keys(field)
        self.driver.find_element_by_id('select2-select1-results').click()

    def item_search2(self, ref, field, first=False):
        self.driver.find_element_by_partial_link_text('+ Add row').click()
        self.driver.find_element_by_id('value(input2)').clear()
        if not first:
```

```
110             self.driver.find_element_by_id('value(input2)').send_keys(ref[
    field])
111         else:
112             self.driver.find_element_by_id('value(input2)').send_keys(ref[
    field][0])
113         self.driver.find_element_by_id('select2-select2-container').click()
114         self.driver.find_element_by_css_selector('input.select2-
    search__field').send_keys(field)
115         self.driver.find_element_by_css_selector('ul#select2-select2-
    results > li:nth-child(1)').click()
116 
117     def launch_search(self,nb):
118         self.driver.find_element_by_xpath(f'//*[@id="searchCell{nb}"]/span
    [1]/button').click()
119 
120     def get_home(self):
121         self.driver.find_element_by_css_selector("body > div.EPAMdiv.main-
    container > h1 > div > a").click()
122         try:
123             session.driver.find_element_by_link_text('Reset').click()
124         except NoSuchElementException:
125             pass
126 
127     def get_list(soup, line):
128         name = [i for i in [re.sub('\n    ','',str(el)) for el in soup.
    find_all('tr')[line].find_all('td')[0].contents] if i != '<br/>'][0]
129         content = [i for i in [re.sub('\n    ','',str(el)) for el in soup.
    find_all('tr')[line].find_all('td')[1].contents] if i != '<br/>']
130         dic = {name: content}
131         return(dic)
132 
133     def results_page(soup):
134         results = dict()
135         for i in range(scraping_wos.range_info(soup)):
136             result = scraping_wos.get_list(soup, i)
137             results[next(iter((result.keys())) )] = next(iter((result.
    values())))
138         return(results)
139 
140     def load_download(file_):
141         filepath = f"/home/yann/Téléchargements/{file_}"
142         while not os.path.exists(filepath):
143             time.sleep(1)
144         if os.path.isfile(filepath):
145             file_ = codecs.open(f"/home/yann/Téléchargements/{file_}", "r",
     "utf-8")
146             return(BeautifulSoup(file_, 'html.parser'))
147         # read file
148         else:
149             raise ValueError("%s isn't a file!" % filepath)
150 
151 
152     def range_info(soup):
153         return(len(soup.find_all('tr'))-1)
```

```
154
155     def download_result_nb(self,result_nb):
156         self.driver.find_element_by_css_selector(f'#RECORD_{result_nb} >
    div.search-results-content > div > div:nth-child(1) > div > a').click()
157         # There are two ways in which this button may be called, so I try
    both:
158         try:
159             downloadfile = WebDriverWait(self.driver, 10).until(EC.
    element_to_be_clickable((By.CSS_SELECTOR,'#exportMoreOptions')))
160         except TimeoutException:
161             downloadfile = WebDriverWait(self.driver, 10).until(EC.
    element_to_be_clickable((By.CSS_SELECTOR,'#exportTypeName')))
162         downloadfile.click()
163         self.driver.find_element_by_css_selector('#saveToMenu > li:nth-
    child(3) > a').click()
164         # Selecting output
165         self.driver.find_element_by_css_selector('#select2-bib_fields-
    container').click()
166         dropdown = self.driver.find_element_by_css_selector('#select2-
    bib_fields-results')
167         dropdown.find_elements_by_tag_name('li')[3].click()
168         # Selecting format HTML
169         self.driver.find_element_by_css_selector('#select2-saveOptions-
    container').click()
170         dropdown = self.driver.find_element_by_css_selector('#select2-
    saveOptions-results')
171         dropdown.find_elements_by_tag_name('li')[2].click()
172         # Click download button
173         self.driver.find_element_by_css_selector('#exportButton').click()
174
175     def download_ref_results(self):
176         WebDriverWait(self.driver, 15).until(EC.element_to_be_clickable((By
    .CSS_SELECTOR,
177         '#cited-refs-full-record > div.cited-ref-separator > h3 > a'))).
    click()
178         try :
179             self.driver.find_element_by_css_selector('#exportMoreOptions').
    click()
180             self.driver.find_element_by_css_selector('#saveToMenu > li:nth-
    child(3) > a').click()
181             self.driver.find_element_by_css_selector('#numberOfRecordsRange
    ').click()
182         except Exception:
183             self.driver.find_element_by_css_selector('#exportTypeName').
    click()
184             self.driver.find_element_by_css_selector('#saveToMenu > li:nth-
    child(3) > a').click()
185             self.driver.find_element_by_css_selector('#numberOfRecordsRange
    ').click()
186
187         self.driver.find_element_by_css_selector('#page > div.ui-dialog.ui-
    widget.ui-widget-content.ui-corner-all.ui-front.ui-dialog-quickoutput.
    qoExcel > div.ui-dialog-content.ui-widget-content > form > div.
    quickoutput-content > div.quick-output-section > div > span > span.
```

```
selection > span').click()
188         self.driver.find_elements_by_css_selector('#select2-bib_fields-
    results > li')[1].click()
189         self.driver.find_element_by_css_selector('#excelButton').click()
190 
191 
192     def clean_ref_table():
193         filepath = '/home/yann/Téléchargements/savedrecs.xls'
194         while not os.path.exists(filepath):
195             time.sleep(1)
196         if os.path.isfile(filepath):
197             ref = pd.read_excel(filepath)
198             col_list = ['Authors','Article Title','Publication Year','DOI',
    'Abstract','Publication Type']
199             return(ref[col_list])
200         # read file
201         else:
202             raise ValueError("%s isn't a file!" % filepath)
203 
204 
205     def delete_file(file):
206         os.remove(f'/home/yann/Téléchargements/{file}')
207 
208     def is_error(self):
209         try:
210             WebDriverWait(self.driver, 5).until(EC.element_to_be_clickable
    ((By.ID,'noRecordsDiv')))
211             return(True)
212         except TimeoutException:
213             return(False)
214         except NoSuchElementException:
215             return(False)
216 
217     def search_reference(self, clean_ref, doi, first=False):
218         break_time = np.random.randint(10,20, size=2)
219         # Retrieving elements in the source
220         if doi:
221             self.item_search1(clean_ref,'doi')
222             self.launch_search(1)
223             if self.is_error():
224                 time.sleep(break_time[0])
225                 self.driver.find_element_by_link_text('Reset').click()
226                 self.item_search1(clean_ref,'title')
227                 if 'author' in clean_ref.keys():
228                     self.item_search2(clean_ref,'author',first)
229                     self.launch_search(2)
230                 else :
231                     self.launch_search(1)
232                 if self.is_error():
233                     print('no result')
234                     return({'wos_no_result':True})
235                 else:
236                     time.sleep(break_time[1])
237                     self.download_result_nb(1)
```

```
                    time.sleep(5)
                    self.download_ref_results()
                    soup = scraping_wos.load_download('savedrecs.html')
                    results = scraping_wos.results_page(soup)
                    scraping_wos.delete_file('savedrecs.html')
                    refs_tab = scraping_wos.clean_ref_table()
                    results['ref_wos'] = refs_tab
                    scraping_wos.delete_file('savedrecs.xls')
                    print('results found')
                    return(results)
            else:
                time.sleep(break_time[1])
                self.download_result_nb(1)
                time.sleep(5)
                self.download_ref_results()
                soup = scraping_wos.load_download('savedrecs.html')
                results = scraping_wos.results_page(soup)
                scraping_wos.delete_file('savedrecs.html')
                refs_tab = scraping_wos.clean_ref_table()
                results['ref_wos'] = refs_tab
                scraping_wos.delete_file('savedrecs.xls')
                print('results found')
                return(results)
        else:
            # First try with two components: title, first author
            self.item_search1(clean_ref,'title')
            if clean_ref['author'] is not None:
                self.item_search2(clean_ref,'author',first)
                self.launch_search(2)
            else:
                self.launch_search(1)
            if self.is_error():
                time.sleep(break_time[0])
                # Then try with title only
                self.driver.find_element_by_link_text('Reset').click()
                self.item_search1(clean_ref,'title')
                self.launch_search(1)
                if self.is_error():
                    print('no result')
                    return({'wos_no_result':True})
                else:
                    time.sleep(break_time[1])
                    self.download_result_nb(1)
                    soup = scraping_wos.load_download('savedrecs.html')
                    results = scraping_wos.results_page(soup)
                    scraping_wos.delete_file('savedrecs.html')
                    time.sleep(5)
                    try:
                        self.download_ref_results()
                        refs_tab = scraping_wos.clean_ref_table()
                        results['ref_wos'] = refs_tab
                        scraping_wos.delete_file('savedrecs.xls')
                    except Exception:
                        pass
```

```
                    print('results found')
                    return(results)

            else:
                time.sleep(break_time[1])
                self.download_result_nb(1)
                soup = scraping_wos.load_download('savedrecs.html')
                results = scraping_wos.results_page(soup)
                scraping_wos.delete_file('savedrecs.html')
                time.sleep(5)
                try:
                    self.download_ref_results()
                    refs_tab = scraping_wos.clean_ref_table()
                    results['ref_wos'] = refs_tab
                    scraping_wos.delete_file('savedrecs.xls')
                except Exception:
                    pass
                print('results found')
                return(results)

    def research_from_list(self,references,cleaned=False, first=False):
        doi = scraping_wos.is_doi(references)
        list_result = []
        kitkat = np.random.randint(15,35,1)
        for element in range(len(references)):
            if not cleaned:
                clean_el = scraping_wos.clean_ref(references[element], doi)
            else:
                clean_el = references[element]
            print('--------\nSearching result for {}'.format(clean_el['
title']))
            try:
                result = self.search_reference(clean_el, doi,first)
                result['id'] = clean_el['id']
                list_result.append(result)
            except NoSuchElementException:
                print('No element found, moving forward')
                result = {'error':True, 'id':clean_el['id']}
                list_result.append(result)
            except TimeoutException:
                print('Timeout, moving forward')
                result = {'error':True, 'id':clean_el['id']}
                list_result.append(result)
            time.sleep(int(kitkat))
            try:
                self.get_home()
            except NoSuchElementException:
                self.driver.get("http://apps.webofknowledge.com/")

        print('----------------------\nFinished \n----------------------'
)
        return(list_result)
```

Executing the functions

I now execute the functions to retrieve the data.

```
# Loading data
path = re.sub('/00.coding.+','',sys.path[0])
folder = path + '/01.data/tempo_sources/'
with open(folder + 'data_to_wos_v2', 'rb') as f1:
        data = pickle.load(f1)

del data['nature_based']['nodoi'][17] # deleting a bad entry

# Logging into the advanced search module
with open(path + '/mp', 'rb') as f1:
    mp = pickle.load(f1)
session = scraping_wos()
session.connect_wos(username = mp[0],
                    password = mp[1])

nature_based_noref = session.research_from_list(data['nature_based']['noref
    '])

nature_based_nodoi = session.research_from_list(data['nature_based']['nodoi
    '])

nb_oecd_wos = nature_based_nodoi + nature_based_noref

with open(folder + 'nb_oecd_wos', 'wb') as f1:
    pickle.dump(nb_oecd_wos, f1)

nature_based_eu_noref = session.research_from_list(data['nature_based_eu'][
    'noref'])

nature_based_eu_nodoi = session.research_from_list(data['nature_based_eu'][
    'nodoi'])
nb_eu_wos = nature_based_eu_noref + nature_based_eu_nodoi
with open(folder + 'nb_eu_wos', 'wb') as f1:
    pickle.dump(nb_eu_wos, f1)

nature_based_wb_nodoi = session.research_from_list(data['nature_based_wb'][
    'nodoi'])

nature_based_wb_noref = session.research_from_list(data['nature_based_wb'][
    'noref'])

nb_wb_wos = nature_based_wb_noref + nature_based_wb_nodoi
with open(folder + 'nb_wb_wos', 'wb') as f1:
    pickle.dump(nb_wb_wos, f1)
```

Assembling Retrieved Data

Once again, I start by setting up the functions needed before executing them. Here I add meta-data retrieved from Cross-Ref with the one retrieve from web of science. Furthermore, I retrieve meta-data about second-degree references from cross-ref when possible.

Functions definition

```
import pickle
import re
import seaborn as sns
import matplotlib.pyplot as plt
sns.set_theme(style="whitegrid")
import time
from crossref.restful import Works, Etiquette
import json
import numpy as np
from tqdm.notebook import tqdm as tqdm
path = re.sub('\\\\00.coding.+','',sys.path[0])

def get_position_id_source(source,id):
    pos = [i for i in range(len(source)) if 'id' in source[i].keys() and
   source[i]['id'] == id ][0]
    return(pos)

class cross_ref():
    """ This class is used to retrieve data from crossref api """

    def __init__(self):
        agent = Etiquette('yann.collindavid@gmail.com')
        self.works = Works(etiquette=agent)

    def check_exists_doi(references):
        for index in range(len(references)):
            try:
                references[index]['doi']
            except KeyError:
                references[index]['doi']='no doi in oecd report'


    def check_exists_title(references):
        to_pop = list()
        for index in range(len(references)):
            if references[index]['title'][0] == '':
                to_pop = to_pop + [index]
        if len(to_pop) > 1 :
            to_pop.sort(reverse=True)
            [references.pop(poppy) for poppy in to_pop]
            print(f'Deleted elements {to_pop}, because of empty title')

    def valid_doi(reference,key):
        return(re.sub(r'\.$','',reference[key]))

    def valid_date(reference):
        date = reference['date'][0]
        if date in ['n.d.','Forthcoming']:
            return(2000)
        else:
            return(re.sub(' |[a-zA-z]','',date))

```

```
    def first_author(reference):
        return(re.sub(",.+","",reference['authors'][0]))

    def create_id(references):
        for i in range(len(references)):
            references[i]['id']=i

    def query_doi(self,reference, key):
        kitkat = int(np.random.randint(2,4,1))
        time.sleep(kitkat)
        if reference[key] != 'no doi in oecd report':
            doi = cross_ref.valid_doi(reference, key)
            search = self.works.doi(doi)
            return(search)
        else:
            print('no do provided for {}'.format(reference['title']))
            return('no doi provided')

    def result_match_raw(raw,result,key_result):
        ti_res = result[key_result][0].lower()
        ti_raw =  raw['title'][0].lower()
        if ti_res in ti_raw or ti_raw in ti_res:
            return(True)
        else:
            return(False)

    def search_for_doi(self, reference):
        if reference['doi'] == 'no doi in oecd report':
            # searching for the doi
            title = reference['title'][0]
            author = cross_ref.first_author(reference)
            date =  cross_ref.valid_date(reference)

            searches = self.works.query(title).filter(from_online_pub_date=
date).sample(1).query(author=author)
            # keeping the first result of the search
            try:
                search = [item for item in searches]


                if len(search) >0:

                    if type(search) is list :
                        search = search[0]

                    # First scenario: we got it right (lucky us!)
                    if cross_ref.result_match_raw(reference, search,
key_result='title'):
                        doi = search['DOI']
                        print('direct match \n')
                        return(doi)

                    # Second scenario, we could not retrieve it directly,
but there's a chance our source is within
```

```
                    # the references of the search results
                    else :
                        if 'reference' in search.keys():
                            for ref in search['reference']:
                                if 'title' in ref.keys():
                                    if cross_ref.result_match_raw(reference
, ref, key_result='title'):
                                        print('match with title \n')
                                        if 'DOI' in ref.keys():
                                            return(ref['DOI'][0])
                                        else:
                                            return('no doi found')

                                elif 'volume-title' in ref.keys():
                                    if cross_ref.result_match_raw(reference
, ref, key_result='volume-title'):
                                        print('match with volume title \n')
                                        if 'DOI' in ref.keys():
                                            return(ref['DOI'][0])
                                        else:
                                            return('no doi found')

                                elif 'unstructured' in ref.keys():
                                    if cross_ref.result_match_raw(reference
, ref, key_result='unstructured'):
                                        print('match with unstructured  \n'
)
                                        if 'DOI' in ref.keys():
                                            return(ref['DOI'][0])
                                        else:
                                            return('no doi found')
                                else:
                                    print('no doi found after search  \n')
                                    return('no doi found')

                        else:
                            print('no doi found because no ref \n')
                            return('no doi found')
                else:
                    print('crossref returns nothing, no doi  \n')
                    return('no doi found')
            except json.JSONDecodeError:
                print('not a valid json file returned')
        else:
            print('{} already has DOI \n'.format(reference['title'][0]))


    def create_sourced_results(self, references):
        """
        This super function takes raw references from OECD reports as
    inputs,
        checks for existence of DOI, search for it if not existing,
        and finally returns complete metadata from CROSS-REF if DOI exists.
        """
```

```
        # 0 step 1: check doi existence and creation if not
        cross_ref.check_exists_title(references) # Gonna be needed for
    searches
        # 0 step create ids
        cross_ref.create_id(references)
        # 0 step 1: check doi existence and creation if not
        cross_ref.check_exists_doi(references)
        # First step: adding searched doi if existing
        for ref in references:
            kitkat = int(np.random.randint(1,10,1))
            time.sleep(kitkat) # Let's give cross-ref a little break
            if ref['doi'] == 'no doi in oecd report' and 'search_doi' not
    in ref.keys():
                print('Searching for DOI for {}'.format(ref['title'][0]))
                ref['search_doi'] = self.search_for_doi(ref)
            else:
                ref['search_doi'] = ref['doi'][0]
        # Second step: create sources
        info = dict()
        print('\n---------------------------------------------- \n')
        for ref in references:
            print('Adding source for {}  \n'.format(ref['title'][0]))
            kitkat = int(np.random.randint(5,10,1))
            time.sleep(kitkat) # Let's give cross-ref a little break
            id = ref['id']
            if ref['search_doi'] == 'no doi found' or ref['search_doi'] is
    None :
                info[id]= {'id': id,
                'result':'no doi found'
                }

            else :
                info[id]= {'id':id,
                           'result': self.query_doi(ref,'search_doi')}
        return(info)

    def clean_from_crossref(self,source,key=None,id=None,ref=True):
        if id is not None:
            pass
        elif 'id' in source.keys():
            id = source['id']
        else :
            id = '999'
        if source is not None:
            if key:
                tempo_results = source[key]
            else:
                tempo_results = source


            if 'article-title' in tempo_results.keys():
                title = ''.join(tempo_results['article-title']).lower()
            else:
                title = ''.join(tempo_results['title']).lower()
```

```
203
204
205             results = {'id': id,
206                     'title':title,
207                     'type':tempo_results['type'].lower()}
208
209             if 'abstract' in tempo_results.keys():
210                 results['abstract'] = tempo_results['abstract']
211
212
213             if 'author' not in tempo_results.keys():
214                 results['author'] = tempo_results['publisher'].lower()
215             else:
216                 results['author'] = cross_ref.clean_author_crossref(
    tempo_results)
217
218             if 'reference' in tempo_results.keys() and ref is True:
219                 results['reference'] = self.clean_reference_crossref(
    tempo_results)
220
221             if 'DOI' in tempo_results.keys():
222                 results['doi'] = tempo_results['DOI']
223         else:
224             results = {'id':id,'noresult':True}
225         return(results)
226
227
228     def clean_author_crossref(source):
229         if 'name' in source['author'][0].keys():
230             return([author['name'].lower() for author in source['author']])
231         else:
232             return([author['family'].lower() for author in source['author']
233                     if 'family' in author.keys()])
234
235     def clean_reference_crossref(self, source):
236         references = source['reference']
237         references_clean = list()
238         length = len(references)
239         title= source['title'][0]
240         print(f'Retrieving inner references for {title}')
241         for id in tqdm(range(length)):
242             ref = references[id]
243             # Cleaning the reference dictionnary
244             to_del = ['key','doi-asserted-by']
245             remove = [key for key in ref.keys() if key in to_del]
246             for k in remove: del ref[k]
247             if 'DOI' in ref.keys():
248                 try:
249                     searched = self.clean_from_crossref(source = self.
    query_doi(ref,'DOI'),id=id, ref=False)
250                 except:
251                     searched = {'id': id, 'error_doi':True, 'doi':ref['DOI'
    ]}
252                 references_clean.append(searched)
```

```
            else:
                tempo = dict()
                if 'author' in ref.keys():
                    tempo['author'] = re.sub('( |^)[A-Z]{1,2}( |$)|\\.| ',' ',ref['author']).lower()
                if 'year' in ref.keys():
                    tempo['year'] = ref['year']
                if 'journal-title' in ref.keys():
                    tempo['journal'] = ref['journal-title']
                if 'type' in ref.keys():
                    tempo['type'] = ref['type']
                if len(tempo)==0:
                    tempo['unstructured'] = ref['unstructured']
                tempo['id'] = id
                references_clean.append(tempo)

        return(references_clean)

    # 1rst degree reference
    def get_wos_ref(self, reference, id):
        ref = dict()
        if 'AU ' in reference.keys():
            author = [re.sub(',.+','',author).lower() for author in reference['AU ']]
            ref['author'] = author

        if 'TI ' in reference.keys():
            title = reference['TI '][0].lower()
            ref['title'] = title

        if 'DT ' in reference.keys():
            type_ = reference['DT '][0].lower()
            ref['type'] = type_

        if 'PY ' in reference.keys():
            date = reference['PY '][0]
            ref['date']=date

        if 'DI ' in reference.keys():
            doi = reference['DI '][0]
            ref['doi'] = doi

        if 'AB ' in reference.keys():
            ref['abstract'] = reference['AB '][0]

        # Inner references
        if 'CR ' in reference.keys():
            reference = self.get_inner_wo_ref(reference['CR '])
            ref['reference'] = reference

        ref['id'] = id

        return(ref)

```

```
305
306     def get_inner_wo_ref(self, references_list):
307         ref_clean = list()
308         for id in tqdm(range(len(references_list))):
309             ref = references_list[id]
310             # If I find the doi, i search for info on crossref about
    article
311             if re.search(' DOI ',ref):
312                 doi = re.sub('^.+ DOI ','', ref)
313                 try:
314                     kitkat = int(np.random.randint(3,8,1))
315                     time.sleep(kitkat)
316                     searched = self.clean_from_crossref(source = self.works
    .doi(doi),id=ref,ref=False)
317                     ref_clean.append(searched)
318                 except Exception:
319                     split = ref.split(',')
320                     author = re.search('[a-zA-Z]+', split[0])[0].lower()
321                     date = re.sub(' ','',split[1])
322                     title = re.sub('^ ','',split[2]).lower()
323                     ref_clean.append({'title':title,'author':author,
324                                       'date':date, 'id':id, 'doi':doi})
325             else:
326                 split = ref.split(',')
327                 author = re.search('[a-zA-Z]+', split[0])[0].lower()
328                 date = re.sub(' ','',split[1])
329                 title = re.sub('^ ','',split[2]).lower()
330                 ref_clean.append({'title':title,'author':author,
331                                   'date':date, 'id':id})
332         return(ref_clean)
```

Execution

It's now time to execute this.

```
path = re.sub('\\\\00.coding.+','',sys.path[0])
folder = path + '\\01.data\\tempo_sources\\'

data_final = dict()


files = [
    'nb_oecd',
    'nb_eu',
    'nb_wb'
        ]
extensions = [
              '_source',
              '_crossref',
              '_wos',
                '_wos_clean'
                ]

search = cross_ref()

```

```
for _file in files:
    file_name = re.sub('nb_','',_file).upper()
    print(f'\nWorking on the {file_name} references\n--------------')
    for extension in extensions:
        if (extension != '_wos_clean' and _file != 'nb_oecd' ):
            with open(folder + f'{_file}{extension}', 'rb') as f1:
                tempo = pickle.load(f1)
            if extension == '_source':
                print('added source')
                data_final[_file] = tempo
            elif extension == '_crossref':
                print('starting adding crossref results')
                for el in tempo:
                    print(el, end=" ")
                    if type(tempo[el]['result']) is dict:
                        try :
                            id = tempo[el]['id']
                            pos_source = get_position_id_source(data_final[
    _file],id)
                            data_final[_file][pos_source] = search.
    clean_from_crossref(source = tempo[el]['result'],key=None,id=id)
                        except:
                            print('something went wrong here, needs to be
    checked')
            elif extension == '_wos' or extension == '_wos_clean' :
                print('\n starting adding wos references, with crossref
    check for inner ref')
                for el in range(len(tempo)):
                    print(el, end=" ")
                    if 'wos_no_result' in tempo[el].keys() or 'error' in
    tempo[el].keys():
                        pass
                    else:
                        try:
                            id = tempo[el]['id']
                            pos_source = get_position_id_source(data_final[
    _file],id)
                            if 'reference' not in data_final[_file][
    pos_source].keys():
                                data_final[_file][pos_source] = search.
    get_wos_ref(tempo[el],id=id)
                        except:
                            print('something went wrong, moving forward')
        else:
            pass

    with open(folder + 'nb_cleaned_v3','wb') as f1:
        pickle.dump(data_final, f1)
    print('\n----------------------------\n\n')
```

Finally, I clean it and export it - it's ready for analysis!

```
import pickle
import re
import numpy as np
import pandas as pd
```

```
from dataclasses import make_dataclass

path = re.sub('\\\\00.coding.+','',sys.path[0])
folder = path + '\\\\01.data\\\\tempo_sources\\\\'
with open(folder + 'nb_cleaned_v3','rb') as f1:
    data = pickle.load(f1)


# Cleaning author / title
files = ['nb_oecd','nb_eu','nb_wb']
for file_ in files:
    print(file_)
    fold = data[file_]
    items_to_clean = [item for item in range(len(fold)) if 'long_title' in fold[item].keys()]
    if file_=="nb_eu" or file_=="nb_oecd":
        for i in items_to_clean:
            print(i)
            source = data[file_][i]['long_title'][0]
            data[file_][i]['title'] = re.sub('\..+|,.+','',source).lower()
            if data[file_][i]['authors'][0] is not None:
                data[file_][i]['author'] = [re.sub('^ |(,|) ([A-Z]{1,2}(\.|$))+','',author).lower() for author in data[file_][i]['authors']]
    if file_ == 'nb_wb':
        for i in items_to_clean:
            print(i)
            raw = data[file_][i]['raw'][0]

            # Authors part
            authors = [re.sub('^( )+','',re.sub('(^| )([A-Z]{1,2}\.)+|(^| )and|\.','',author)).lower() for author in re.sub('\. Forthcoming.+| \d{4}.+','',raw).split(",")]
            authors = [author for author in authors if author != ""]

            data[file_][i]['author']=authors

            # Title part
            if re.search('""|',raw):
                title = re.sub('""^.+|\..+$','',raw).lower()
            elif re.search("Forthcoming| \d{4}", raw):
                title = re.sub('\..+|^ ','',re.sub('.+ Forthcoming(\.|)|.+\d{4}(\.|)','',raw)).lower()

            data[file_][i]['title'] = title

# Transformation in article network database
information_reports = {
    'nb_oecd':{
        'title':'nature-based solutions for adapting to water-related climate risks',
        'author':['oecd'],
        'date':2020,
        'type':'report'
    },
```

```
    'nb_eu':{
        'title':'nature-based solutions for flood mitigation and coastal
    resilience',
        'author':['european commision','vojinovic'],
        'date':2020,
        'type':'report'
    },
    'nb_wb':{
        'title':'nature-based solutions for disaster risk management',
        'author':['world bank','ozment','ellison','jongman'],
        'date':2019,
        'type':'report'
    }
}
# Two functions: needed:
connect = make_dataclass("Connection", [("Citing", str), ("Cited", str), ("
    Report",str),('Type',str),('Level',str)])

def new_connect(citing, cited, report, type_, level):
    return(pd.DataFrame([connect(citing, cited, report, type_, level)]))

# Creating the dataframe of papers citations
papers_cite = pd.DataFrame(columns=['Citing','Cited'])
for report in information_reports:
    report_title = information_reports[report]['title']

    for reference in range(len(data[report])):
        ref_title = data[report][reference]['title']
        if 'type' in data[report][reference].keys():
            ref_type = data[report][reference]['type']
        else:
            ref_type = 'na'
        level = 1
        papers_cite = papers_cite.append(new_connect(report_title,ref_title
    ,report,ref_type,level), ignore_index=True)

        if 'references' in data[report][reference].keys():
            data[report][reference]['reference'] = data[report][reference][
    'references']
            del data[report][reference]['references']

        if 'reference' in data[report][reference].keys():
            for sub_ref in range(len(data[report][reference]['reference']))
    :
                if 'title' in data[report][reference]['reference'][sub_ref
    ].keys():
                    sub_ref_title = data[report][reference]['reference'][
    sub_ref]['title']
                    if 'type' in data[report][reference]['reference'][
    sub_ref].keys():
                        subref_type = data[report][reference]['reference'][
    sub_ref]['type']
                    else:
                        subref_type = 'na'
```

```
                    level = 2
                    papers_cite = papers_cite.append(new_connect(ref_title,
    sub_ref_title,report,subref_type, level), ignore_index=True)

papers_cite.to_csv(folder+'ntk_papers.csv')
```