Using NLP to build causal maps from emissions trading policy analysis literature: A more comprehensive means of analysis

Rory Hooper, 5311446

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of Master of Science in Engineering and Policy Analysis Faculty of Technology, Policy and Management. To be defended publicly on 20.07.22

ARTICLE INFO

Graduation committee:

Prof. df. K Blok – Chair
Dr. N Goyal – 1st supervisor
Dr. L Scholten – 2nd supervisor

Keywords:

Policy analysis
Emissions trading schemes
Natural language processing
Relation extraction
Causal relations
Causal maps

ABSTRACT

Despite their best intentions, policy interventions often fail to adequately address the challenges they were designed to tackle. Disparities in scope, taxonomy and performance perspectives employed by different policy studies, make it difficult to obtain a system perspective of the policy effects of a certain domain. Additionally, a prohibitively large body of literature often makes manual methods of review infeasible. To combat these shortcomings of existing policy analysis methods and provide an insightful way of informing policy evaluation and development, this project proposes a novel five-step policy analysis method that can semi-automatically derive and aggregate causal relations from policy literature into a causal map of policy effects. The method has been applied to a collection of 28 emissions trading scheme (ETS) policy analysis literature sources, producing a causal map consisting of 159 causal links with a recall of 38% and precision of 84%. The results demonstrate that applying this approach produces an aggregated system perspective of the analysed policy's effects, and garners relevant insights for policy evaluation and development that would otherwise be difficult to obtain. This approach can therefore provide analysts from across different policy domains with a more comprehensive understanding of the factors and relations affecting policy and so provide a more comprehensive evidence base from which to inform policy development.

1. Introduction

In a world of wicked problems and grand societal challenges, we turn to policy interventions to provide a solution. Unfortunately, however, such policies often fail to adequately address the challenges they were designed to tackle (Bovens & 't Hart, 2016; Howlett et al., 2015). For climate-related policies in particular, given the growing threat of global warming, the urgency of the energy transition, and the seriousness of so many other grand challenges, we must endeavour to remedy past policy failures and maximise the utility of future interventions. To this end, policy analysis exists to inform policy development that may ultimately produce better policy outcomes. Yet as this paper will explain, existing methods of policy analysis suffer from several shortcomings.

1.1. Policy analysis in practice

The first shortcoming of current means of policy analysis relates to the disparate scope of analysis employed by different studies. Looking at the 'level' of analysis, there are numerous small-n or case-specific 'low-level' studies into the performance of individual policies i.e. (Jiali, 1995; Warner & van Buuren, 2011). These studies allow specific insights into how the analysed policy could be developed, but broader insights into how similar policies could be improved are limited because it is unclear how the identified factors and dynamics transfer to other contexts. There are also 'high-level' studies of entire policy areas i.e. (Esty & Porter, 2005; Thow et al., 2010). These studies frequently miss more granular case-specific factors which could be influential but aim instead to understand the higher-level factors which prove relevant in a diverse population of cases. At each level of scope, there is also a question of performance perspectives. Policy analysis is often conducted concerning only a single performance dimension, or a certain aspect of a certain dimension i.e. (Wakabayashi & Kimura, 2018) (Marsh & McConnell, 2010). For example, one study may only analyse the performance of a policy with respect to its political success whilst another may look at its achievement of programmatic aims. Whilst studies from each level of scope, and each performance perspective undoubtedly contribute valuable insights individually, viewing them in isolation will provide only a limited perspective of performance as a whole.

This shortcoming is compounded by the disparity in taxonomy employed by different studies. Some papers might refer to "competitive distortions" in certain industries following the implementation of a policy. Another study on the same policy may instead refer to the same dynamic as "disruptions to market merit order". In more complex cases, such taxonomical differences can cause difficulties in making

comparisons between studies and important connections may be lost. These issues of scope and terminology can therefore encumber the achievement of a more comprehensive, aggregate understanding of a given policy domain.

Finally, it is also important to consider the format of analysis. As with most academic fields, policy analysis is often made public in the form of scientific literature. Whilst the exponential growth in scientific literature provides an abundance of relevant source material, the sheer quantity of information means that manual methods of review are often not feasible (Bornmann & Mutz, 2015; Larsen & Von Ins, 2010; Nunez-Mir et al., 2016). Manual synthesis approaches such as systematic reviews and meta-analyses do exist but are very effort-intensive. As the literature space continues to expand, adequately encompassing all relevant studies within a review will become increasingly difficult and relevant material is likely to be missed.

Existing methods of policy analysis evidently leave room for improvement. Research into the reasons for policy performance is usually constrained in scope, utilising distinct taxonomies, and presented in a prohibitively large body of academic literature. As a result, more comprehensive insights into the factors and relations which contribute to policy success or failure can be difficult to obtain.

1.2. Towards a new method of analysis

Given these limitations, there exists an opportunity relating to how policy-relevant insights can be derived by combining the lessons learned from across disparate sources of analysis. To combat the limitation imposed by the large body of text, there exists an opportunity to leverage automated or semi-automated language processing tools. With this in mind, this section will now explore two specific concepts which hold particular promise in addressing these opportunities.

Conceptual models – This umbrella term refers to models which represent a system and can be used to better understand the modelled subject. One such example is causal maps, these models aggregate the causal relations contributing to the behaviour of a system (Eden et al., 1992). The benefits of using a causal map for policy analysis are two-fold. Firstly, given that their representation is comprised of individual factors and their causal connections, they are well suited for combining analysis of different scopes. The more granular features found from a specific case study can be included alongside high-level, case-independent behaviour, identified in a study on an entire policy area. In the same way, features relating to, say, process success can be included alongside programmatic success features. If there is no relationship between these features then they will exist in unconnected parts of the map but if evidence suggests some aspect of their structure is related, then they will be connected with a causal link. The second benefit is that this format allows for easy expansion when new information is introduced, for example, additional concepts and connections could be added to a causal map if a new policy case study supports it. The use of causal map modelling for structuring policy insights, therefore, appears promising. The success of their application in literature, albeit limited, supports this idea (H. Kim & Andersen, 2012). To further test their applicability for policy analysis, their use was evaluated in Appendix A.

Natural language processing – Natural language processing (NLP) encompasses a range of "computational techniques for analysing and representing naturally occurring texts" (Liddy, 2001, p. 1). NLP methods can be categorised according to their function, for example, Information retrieval (IR) which is concerned with retrieving relevant information sources, and information extraction (IE) which is concerned with extracting structured information (i.e. named entity recognition, or sentiment analysis) (Chowdhary, 2020). Information extraction methods are particularly relevant to this study given their potential to derive policy-relevant insights from policy analysis in an automated fashion. In this way, they may be able to address the policy analysis shortcoming stemming from an inability to manually review the burgeoning literature landscape. It is important to note, however, that a fully automatic application is unlikely in this case. Despite significant NLP advances in recent years, sophisticated applications involving complex source material are rarely of sufficient quality to be used without some degree of human interpretation. Accordingly, NLP methods are often used in a semi-automatic process whereby automatically derived outputs can be reviewed and refined by human coders. Such an approach is more likely to be applicable in this project given the complex nature of the source language.

In summary, it is clear that existing approaches to evaluating policy performance suffer from several shortcomings. There exists a research gap in exploring a method that can derive causal relations, in a semi-automated fashion, from policy analysis source material, which can then be aggregated into a causal map to represent the policy effects as a whole. Such a method promises to provide a more comprehensive evidence base to inform policy development.

1.3. Research objective

This research aims to develop a semi-automated method to aggregate causal relations from policy analysis literature into a causal map of policy behaviour. As explained in the next section, this method will concurrently be applied to emissions trading policy to aid in the development of the method and to determine its value/potential value in aiding the analysis and design of policy interventions (refer to Appendix H for EPA relevance). With this objective, this project aims to contribute to the field of policy analysis by presenting an initial exploration of a novel approach that can better synthesise factors and relations influencing policy, and which may ultimately be used to inform policy developments. By applying this method to emissions trading policy, this project aims to contribute to the understanding of the factors and relations that influence emissions trading systems which may ultimately be used to inform future developments in this policy domain. From this objective, the research question of this project is:

To what extent can causal insights from disparate sources of policy analysis literature, be semi-automatically derived and aggregated into a causal map to help in analysing and ultimately developing policy?

2. Emissions trading policy analysis

Emissions trading schemes, also known as cap-and-trade systems are one policy under the broader area of carbon pricing, which aims to impose a cost on the emission of greenhouse gasses. As the name suggests, this policy involves setting a 'cap' on certain types of emissions, within certain sectors. Regulating bodies divide the total allowable emission quantity into tradeable 'allowances that are auctioned or allocated to the entities covered by the system. Polluters can buy and sell these allowances but must surrender the number of allowances corresponding to their respective emissions at the end of pre-defined periods. Entities for whom abatement is relatively cheap have a financial incentive to reduce emissions, and entities for whom abatement is relatively expensive have the option to purchase allowances to satisfy their pollution demand. In this way, emissions trading schemes provide a mechanism to reduce emissions, to specified levels, in a flexible manner, whilst also encouraging the most cost-efficient abatement (ICAP, 2022a).

Given that emissions trading schemes are increasingly becoming a core policy tool to drive decarbonisation their adoption is only expected to increase in the coming years (European Commission, 2021; ICAP, 2022b). Consequently, it is imperative that they can deliver on their promise to reduce emissions in a cost-effective, efficient manner. Worryingly, however, empirical analysis has so far indicated only limited policy success. Looking at ex-post examinations of ETS emission abatement performance, Green (2021) finds that aggregate reductions are limited, generally between 0-2% per year. In other success dimensions as well schemes have experienced mixed results, with political outcomes often proving to be controversial (Leining et al., 2020). It is for this reason that emissions trading schemes have been chosen as the example policy analysis domain to be used in this project.

ETS policy analysis exhibits many of the limitations discussed in section 1.1: Analysis is conducted at various levels of scope from case studies exploring the impacts of one ETS policy on a small number of firms, to high-level reviews seeking to summarise the wholesale abatement impact of carbon pricing. Studies frequently only consider a single performance dimension. The taxonomy employed varies considerably. And finally, there is an enormous literature base, Google Scholar returns 23,100 results for a preliminary search of "ETS" and "policy analysis". As suggested for policy analysis in general, these features encumber the discovery of more comprehensive insights into the factors and relations which contribute to ETS policy performance.

3. Methods

This section presents the proposed method to go from raw policy analysis literature to an aggregated causal map of policy effects, as applied to the ETS source material. The inputs and outputs of each step are shown in Figure 1. After selecting relevant articles, the causal relations present are extracted, similar factors are grouped and an aggregate causal map can be built. The model can then be analysed in a variety of ways to derive policy-relevant insights.

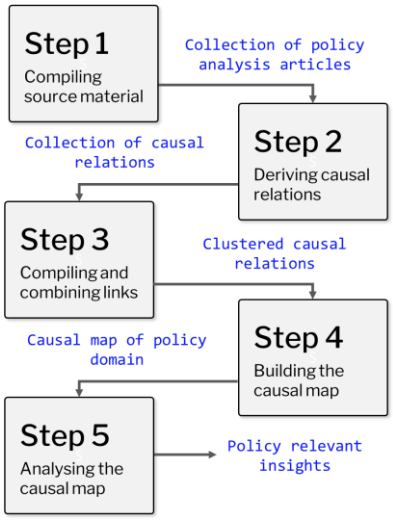

Figure 1 - Input-output diagram of the five-step process

3.1. Step 1: Compiling source material

First, the specific policy analysis literature sources must be selected. This is an important step as it defines the boundaries of what will be included in the causal map. In principle a causal map can be constructed from any subset of literature, to reflect as narrow, or broad, of an analysis perspective as desired. The selection criteria used in *this* project are described below.

This study examines ex-post ETS policy analysis literature encompassing multiple ETS jurisdictions, and including programmatic, process and political performance dimensions (Marsh & McConnell, 2010). Including ex-ante studies would likely obfuscate the results by giving a false equivalence to projected causal relations and those identified from empirical evidence. The global scope and performance dimension criteria were chosen to maximise the breadth of included policy effects. To compile source material I have first consulted relevant meta-reviews of ETS literature and then selected all relevant articles within. This was done to benefit from the possibility of comparing the conclusions of this study with those presented in the meta-reviews of the same papers.

To find meta-reviews, an initial broad search was conducted using Google Scholar and Scopus comprising the following terms in combination with snowball searching.

- ETS
- Emission trading
- Cap-and-trade
- Carbon pricing
- Emission pricing

+

- Review
- Analysis
- Meta-review
- Meta-analysis

The results were then filtered to only include ex-post evaluations. These search conditions yielded 14 meta-review articles spanning periods from 1990 – 2021 and covering almost all ETS jurisdictions. The full list of articles can be found in Appendix B: ETS meta-reviews. From this list, two meta-reviews were selected to contribute the individual source papers:

Green (2021) – Chosen due to its explicit ex-post and quantitative direction encompassing studies from all major ETS jurisdictions. Additionally, it is a recent review and so includes studies not present in other review papers. One notable deficiency is that, given its stated focus on quantitative assessment and evaluation against carbon emissions as opposed to other success criteria, this review can be considered to have a solely programmatic performance perspective. Using only the source papers from this review will likely give limited insights into more process and political factors and dynamics that may be present in the wider literary landscape.

Schmalensee & Stavins (2017) - Included because of its specific discussion on political considerations and system design factors, which can be considered as a proxy for more process and political perspectives. This helps to address the gap from the Green (2021) review. It is also a recent review and covers the most significant ETS jurisdictions.

From the Green (2021) paper, the 18 papers relating solely to carbon taxes were excluded given their irrelevance to emissions trading. From Schmalensee & Stavins (2017), carbon tax papers were omitted, Burtraw (2006) due to its ex-ante analysis, and Goulder & Stavins (2011) due to their focus on other energy-related policies. Collating the papers within both reviews and removing those deemed irrelevant therefore yielded a final total of 28 source papers, as shown in Appendix C: Final ETS policy analysis source papers.

3.2. Step 2: Deriving causal relations

The next step involves deriving the individual causal relations present in the source material. Each sentence within the policy analysis source material is examined to determine whether it exhibits causality. Take, for example, the causal sentence "The higher emissions allowance price caused a decrease in coal power generation". Step 2, seeks to extract the causal factor, in this case 'emissions allowance price' the effect factor, 'coal power generation', and the direction of their causal relation, either positive or negative. A positive direction indicates that increasing the cause factor would increase the effect factor, conversely, a negative direction indicates that increasing the cause factor would decrease the effect factor. In this case then 'emissions allowance price' has a negative causal relation with 'coal power generation'. This is the most important step in the process given that the derived causal relations and their associated factors constitute the structure of the causal map. Each factor (cause and effect) is represented as a node in the causal map, and the derived causal relations represent the connections between these nodes.

This is no easy task given the complexity of causal relations; however, as described in Appendix D: Approaches to derive causal insights using NLP, there is a growing NLP field addressing this challenge. For this task, a deep-learning-based relation extraction NLP method was deemed most suitable for deriving causal relations. Relation extraction was chosen over co-occurrence-based methods because of the improved granularity of extracted factors, its explicit connection to the source material and the ill-suited relation conditions of co-occurrence methods. A deep-learning method within the relation extraction category was selected because of its high benchmark performance and improved portability over statistical machine learning and knowledge-based methods. The specific algorithm employed is SCITE - Self-attentive BiLSTM-CRF wIth Transferred Embeddings, presented by Li et al., (2021). The algorithm was selected given that it is open-source with well-documented code, has an explicit focus on causal relations and performs highly on benchmark datasets. More details on the selection criteria can be found in Appendix D: Approaches to derive causal insights using NLP. To derive causal relations from ETS literature, the 28 source papers have been processed in four stages:

Selection of relevant textual data – Involves the selection of relevant textual data from the papers to include as input, this includes the abstract, results, discussion, and conclusion sections (and equivalent). Methodology and literature review sections were excluded because causal relations derived from these sections often returned irrelevant methodological links, and relations suggested in other literature which did not necessarily represent a finding from the analysed paper.

Data preparation – the raw text data was automatically segmented into sentences and cleaned (removal of non-alphanumeric characters, separation of punctuation, empty sentences etc.) and then segmented into words. Next, layered embeddings were generated for each sentence using existing libraries (Akbik et al., 2018). Embeddings are essentially computer-readable representations of language which generally take the form of a vector, these are necessary for the algorithm to learn a representation of the input data (Levy & Goldberg, 2014). SCITE requires a matrix of character, word, and flair embeddings for every input sentence.

SCITE processing – The sentences and corresponding embeddings for each sentence in each relevant section in each of the papers are run through the pre-trained SCITE model. Text files including the output causal sentences and their suggested cause-effect factors are generated for each paper. An example causal relation outputted by SCITE is shown in Figure 2.

Manual review – When applied to scientific literature, the recall performance of SCITE is significantly below that achieved on benchmark datasets (as explained in Appendix E: SCITE output, recall and precision metrics). Given the degree of inaccuracy of SCITE outputs, and the lack of explicit causal relation direction provided, it was, therefore, deemed necessary to manually review each SCITE output sentence to determine the true causal relations, causal pairs and direction. This stage was also necessary to remove irrelevant causal sentences, i.e. those which contained a causal relation, but one not related to the emissions trading schemes.

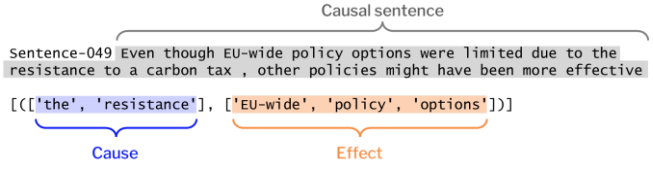

Figure 2 - Example SCITE output, paper 4 sentence 49

3.3. Step 3: Compiling and combining causal links

The previous step has obtained a collection of cause-effect pairs, two factors connected by either a positive or negative causal relation. It is possible to construct a causal map using these pairs, however, a problem arises given the high degree of similarity between many of the factors. For example, when talking about the use of coal power, one causal pair may include the factor 'coal power generation' whereas another could use the factor 'coal utilisation'. Such disparities in naming quickly cloud the causal map as it is often unclear when terms can be combined, or whether similarly named factors should be included in linkages. To combat this issue, some degree of semantic clustering is required. This involves grouping semantically similar factors so that one term can be used which defines their collective meaning. In this way, a more cohesive causal map can be generated with fewer factors and more inter-connections.

Compilation of causal pairs – First the individual causal pairs, derived from all source papers, are concatenated into a table including their contributing sentence and manually labelled cause-effect factors and relationship direction. Additionally, a unique cause-effect id is given to each so that they can easily be distinguished. The id is in the form [paper number] – [sentence number], and when a sentence includes multiple causal pairs, an additional letter is used to differentiate, i.e. id '2-033' indicates the first causal pair from sentence 33 in paper 2, and id '4-137B' indicates the second causal pair from sentence 137 in paper 4.

Semantic clustering – To cluster semantically similar factors we again turn to the concept of 'embeddings' (computer-readable representations of words) in this case, however, 'sentence embeddings' are used which aim to capture a representation of the semantic meaning of a sentence in a high dimensional vector. SBERT, presented by Reimers & Gurevych (2019), was used to generate sentence embeddings for each of the factor phrases present in the manually labelled causal pairs. From here a clustering algorithm is applied to group the factors with similar embedding values. In the data, many similar terms are expected, i.e., the numerous ways to express emission allowance price, but also many unique terms, i.e. case-specific factors found in a single paper. From these requirements, a density-based method was considered most suitable given that they deal well with uneven cluster sizes and outliers. As a result, a DBSCAN clustering algorithm was chosen (Ester et al., 1996) using a cosine distance metric, with a minimum cluster size of 2. The output of this stage is a collection of factor clusters consisting of cause and effect factors with similar semantic meanings.

Manual review - While the previous stage does adequately cluster many factors, semantically dissimilar terms do occasionally still appear within the same cluster. For example, one cluster includes concepts of gas utilisation, gas price and coal to gas price ratio. It is perhaps understandable that these terms would be grouped given their shared relation to natural gas however in the system of emissions trading they necessitate separate factors. A manual review of the automatic clustering outputs was therefore deemed necessary to refine the factors included in each group, and the overarching term used to describe the group. With the semantically clustered factors, the earlier causal pair table can be updated by overwriting factors from the original cause-effect pairs when they have been included in a cluster. An example of the compiled and clustered causal pairs is shown in Table 1.

3.4. Step 4: Building an aggregated causal map

Having compiled and combined the causal links, the final step is to generate the aggregate causal map. Using the clustered cause-effect pairs from the causal link table, individual factors and links can be added to the model as they occur. Slowly the causal map will grow in complexity as more factors appear and connections between structures emerge.

While based on the causal links identified in earlier stages, this map generation stage inevitably involves some degree of coder interpretation. Despite clustering like-terms, some additional grouping may prove warranted during the building process. Implicit structures are also likely to present themselves, the inclusion of which allows a richer model visualisation. For example, many papers discuss the impact that the free allocation of emission allowances and subsequent sales had on power-generating firm profits. Visualising this connection explicitly from allocations → sales → profit is possible, but it overlooks the implicit stock-flow nature of profits, that profits are a product of revenues less costs. By including a stock-flow structure whereby revenues are an inflow to profit, and costs are an outflow, a clearer and more accurate model is achieved. Figure 3 demonstrates how the raw model structure relating to firm profits looks clouded and confusing, including the implicit stock-flow structure allows demarcation of factors influencing revenues and costs, and how each impacts profits.

Another issue that arises is that of 'intermediate variables' whereby one causal link argues for a connection from A → C and another states the path from A → B → C. In many cases, it is unclear whether the link A → C implies the existence of intermediate factor B, whether it is unaware of factor B,

	id	manual_pairs	clustered_manual_pairs
14	2-028A	[european level commitments, +, 20-20-2010 targets]	[european level commitments, +, 2020 EU CLIMATE ENERGY PACKAGE]
15	2-028B	[20-20-2010 targets, +, renewable energy directive]	[2020 EU CLIMATE ENERGY PACKAGE, +, RENEWABLE ENERGY DIRECTIVE]
16	2-028C	[20-20-2010 targets, +, energy efficiency directive]	[2020 EU CLIMATE ENERGY PACKAGE, +, energy efficiency directive]
17	2-029	[renewable energy directive, +, renewable energy utilisation]	[RENEWABLE ENERGY DIRECTIVE, +, RENEWABLE ENERGY UTILISATION]
18	2-031A	[carbon price, +, fuel-switching]	[EMISSIONS ALLOWANCE PRICE, +, FUEL-SWITCHING]

Table 1 - Example of compiled and clustered causal pairs

whether it considers B irrelevant, or indeed whether it argues that A → B → C may only represent one of multiple paths to C, thereby partially contributing to the causal impact of A on C. Addressing this issue requires careful interpretation to avoid any potential coder bias. In situations where an intermediate variable is suggested but not explicitly mentioned by a causal link, the coder can refer to the contributing text segment to gain more context, in this way it is possible in many cases to find support for a connection. In other cases, support cannot be found for an intermediate link, in these cases, an additional path is added to the causal map alongside the intermediate path, i.e. including A → B → C alongside A → C.

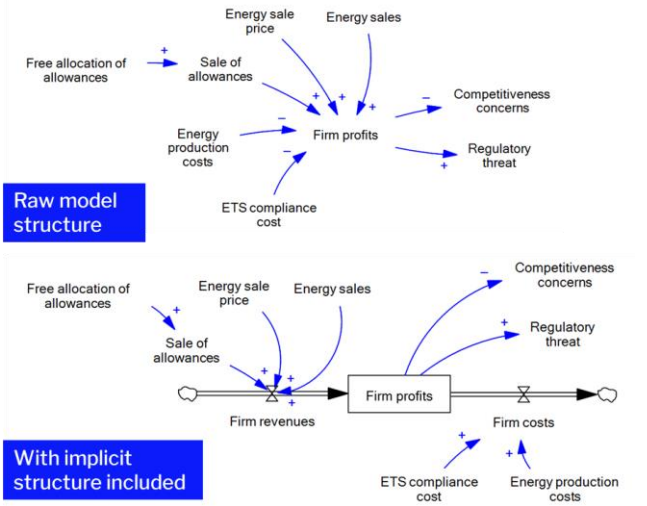

Figure 3 - Example of including implicit structures

What should be apparent is that the model generation stage is a highly iterative process. As new links are added, the overall structure must be reorganised, as the map evolves implicit structures emerge, and intermediate variables appear. There is no objective complete causal map, rather it should be refined until all causal links have been incorporated, and the overall structure can be used to garner relevant insights.

Finally, moving from the explicit causal links in the table to a complex graph structure, the connection with the contributing source material is easily lost. Connection to contributing text segments is necessary to provide additional context when examining the causal map. To do so, each connecting arrow in the causal map has been labelled. A reference table then describes which individual causal link ids were used to support the inclusion of that specific connection. In this way, a reader can identify a connection of interest from the causal map, find the contributing causal links from the reference table, and if desired, use the causal link table to analyse each link in more detail.

3.5. Step 5: Analysing the causal map

The output of step 4 is an aggregated causal map, step 5 now explains how to analyse this map and how each type of analysis may be used for policy analysis and development.

Topographic analysis

This kind of analysis seeks to gain insights from consideration of the structural layout of the cause map. In this category, insights can be gained by looking at the strength of certain causal linkages, by considering the cause and effect trees of factors, and by conducting centrality analysis of the causal network.

Strength of linkages – Involves consideration of the relative strength of linkages in the map (Montibeller & Belton, 2006). Although the generated map provides no explicit strength value, the number of text segments from unique papers supporting a link can be used as a proxy. The occurrence of a causal linkage in multiple papers indicates that it has been studied by multiple authors, accordingly, it can be seen as an indication of the degree of shared knowledge, in the analysed policy domain. Conversely, those links identified in only a few papers have not been studied as extensively and so may represent unique knowledge. This type of analysis can inform decisions on policy interventions by focusing on areas with a high degree of shared knowledge, and so confidence in the given effect. It may also be used to inform future research focus by identifying interesting areas of unique knowledge that may warrant further investigation.

Cause and effect trees – The effect tree of a given factor is the branching collection of downstream factors affected by that factor (Eden et al., 1992). Examining these branches shows the different mechanisms through which that factor affects the wider system. Conversely, the cause tree is the branching collection of upstream factors which have affected that factor. Examining these branches shows the different mechanisms through which the system has influenced it. Analysis of these structures, therefore, gives insight into the various mechanisms of behaviour in the system. An analyst can use this knowledge to: alter the behaviour along one or multiple of the causal branches, attempt to sever one or multiple branches, or introduce a new branch to alter a target factor. Such information would be hard to attain from traditional policy analysis text given that the numerous upstream and downstream effects are spread throughout various texts and are not made explicit.

Centrality analysis – Centrality analysis involves consideration of the positioning and connections of factors in the network structure of the causal map to gain insights into their influence and importance in the system. These kinds of insights cannot be easily gained through traditional review methods given that the network structure of concepts is not apparent in natural language text. There are various centrality metrics proposed in literature however in this project degree and betweenness centrality are most relevant:

- *Degree centrality* refers to the number of connections (degree) of a factor. Those factors with the most connections are deemed the most 'central' in the network. In the context of this project, a high degree centrality means that a certain factor has been mentioned in literature as impacting, or being impacted by, many other factors and therefore gives an indication of its 'influence'. Importantly, however, consideration must be given to the fact that literature may highlight numerous unimportant connections of a factor thereby giving a false sense of high influence, and that very influential factors may only have a small number of connections to other important factors.

- *Betweenness centrality* refers to how much a certain factor is in-between others. Its value is determined as the number of shortest paths between factors, that pass through a given factor. In the context of this project, high betweenness centrality means that a certain factor is involved in many shortest causal paths between factors which suggests it plays a role in many causal mechanisms. As such it gives an indication of the 'importance' of a factor. Again, this metric requires careful consideration, a factor may be located along many *shortest* causal paths but longer paths may also prove very relevant, in this way potentially very important factors may not be highlighted.

Measures such as closeness centrality were excluded because they were not deemed relevant for analysis in this context. Closeness refers to the average distance from a factor to all other factors in the network but given the presence of isolated structures in this network, this metric is not applicable. Nevertheless, in other domains or applications metrics other than degree and betweenness may be insightful.

Causal inference analysis

Causal inference relates to the determination of the causal impacts of one factor on another (Axelrod, 1976), and can be used to derive insights into the causal effect of certain paths as well as the total causal influence of one factor on another. This type of analysis is best suited for a causal map depiction given that trying to determine causal chains through manual text review alone would be a difficult and time-consuming process. Two indices are used in this type of analysis:

Partial effect – This index determines the causal impact of one factor on another, along a certain path. It is obtained by multiplying the signs along each linkage in the chosen path. For example, take the path from A → C → E in Figure 4, there are two positive linkages and so the partial effect is positive. Examining the partial effect of certain linkage paths gives a clear indication of the current literature-based expected impact (positive or negative) that a causal chain will have. This may not be readily apparent in the contributing literature sources given that constituent linkages in a chain can be spread between different sections of a paper, or from different sources entirely. This analysis, therefore, provides a clearer picture of the wider causal implications of certain factors which can be used for policy development efforts.

Total effect – This index determines the total impact of one factor on another, along all connecting paths. The total effect is positive if all paths between those two concepts have a positive partial effect; it is negative if all paths have a negative partial effect, and it is undetermined otherwise. Consider the total effect of factor A on factor E, there are two paths between these nodes, ACE and ADE. The partial effect of ACE is positive and ADE is negative, the total effect is therefore undefined. Looking instead at factors B and E, the total effect is positive. Examining the total effect of a factor on a target factor can give an indication of the *overall* impact that the former has on the latter, across multiple causal chains. This type of analysis is particularly interesting when different causal paths to the target are sourced from different papers. In such cases, one source may have concluded that a factor had a positive impact, whereas another source could have found, through another path, that that same factor in fact had a negative impact. Examination of only the paths suggested by a limited number of sources could therefore give the false impression that a factor has an undisputed impact on a target factor.

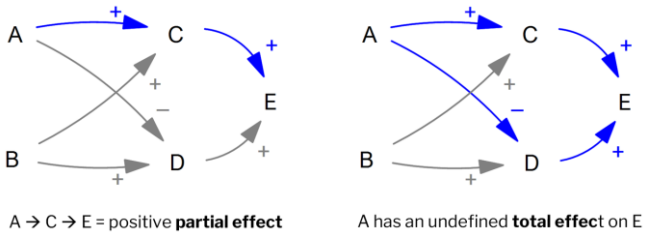

Figure 4 - Types of causal inference

Causal loop analysis

Finally, causal loop analysis involves the identification and description of feedback mechanisms within the causal map. These are circularly connected causal paths that represent either reinforcing or balancing behaviour for the associated factors. A reinforcing loop occurs when the net direction of linkage within the loop is positive, i.e., in Figure 5a, in this case increasing factor A will increase B and C and in turn reinforce the increase of A. Conversely, a balancing loop occurs when the net direction is negative, i.e., in Figure 5b, in this case increasing A will decrease B which in turn decreases C and A. As a result, factor A is considered to be balanced by this loop. This type of analysis is relevant because these loops can convey important feedback dynamics within the ETS system.

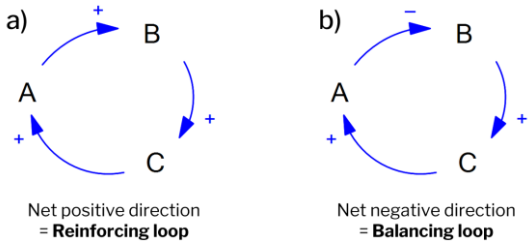

Figure 5 - Types of causal loop

4. Results

The five-step method has been applied to ETS policy analysis literature, yielding the following results: 28 ETS source papers were selected in Step 1. In Step 2, 4542 input sentences were cleaned and provided as input for SCITE. 317 were deemed causal by the algorithm, of which 154 were manually verified as causal and relevant. These sentences contained 284 final causal pairs. Further analysis within this step reveals that the results are obtained with a precision of 84% and a recall of ~38%, details provided in Appendix E: SCITE output, recall and precision metrics. In Step 3 the causal pairs were first compiled which identified that the causal pairs contained 300 unique cause-effect factors. Semantic clustering informed manual grouping of 230 factors into 49 clusters, the remaining 70 factors were deemed sufficiently unique to warrant their own factor. In an iterative process, Step 4, produced the final causal map shown in Figure 6. The complementary reference table is provided in Appendix F.

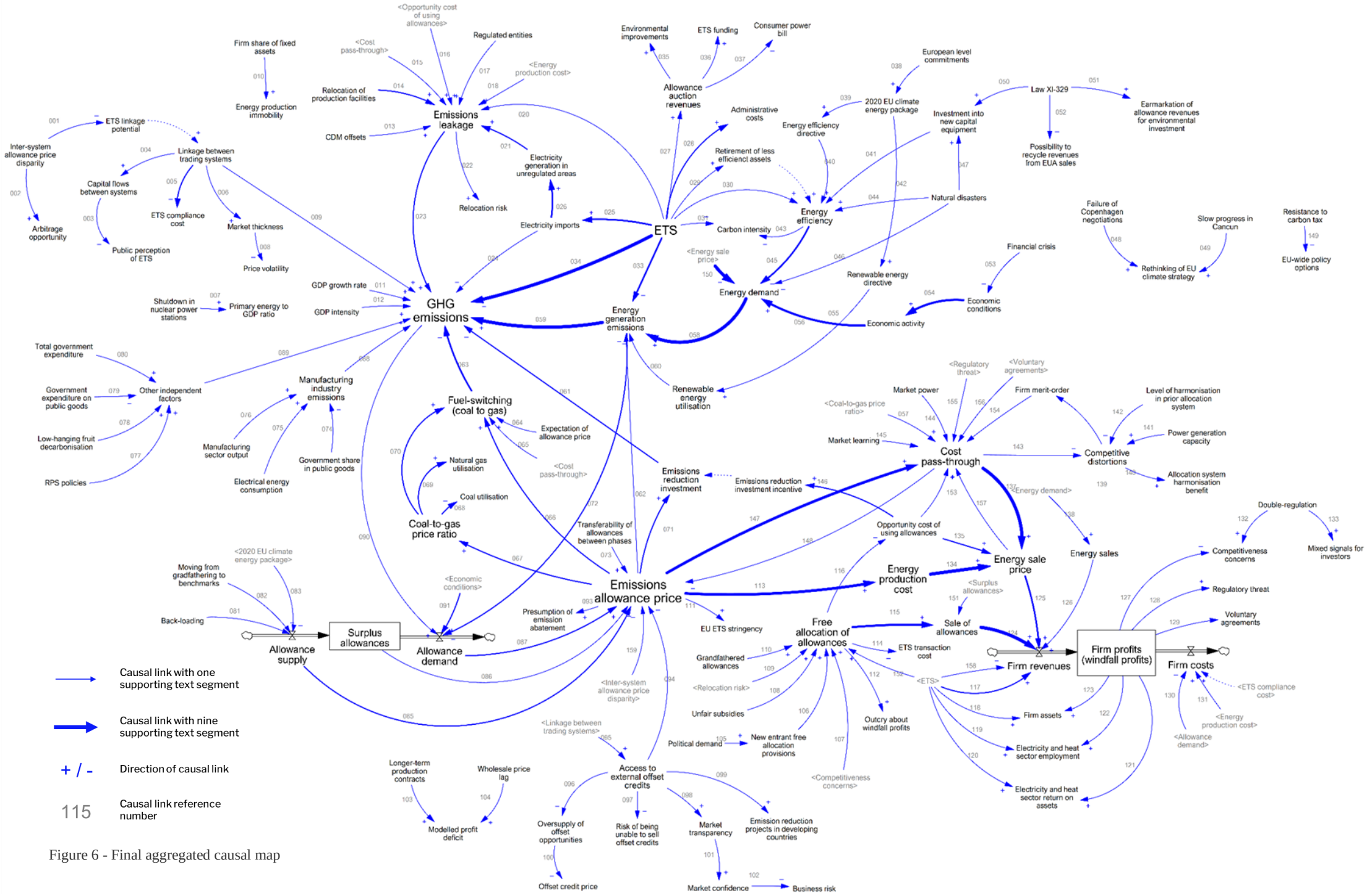

Figure 6 - Final aggregated causal map

4.1. Description of the aggregated ETS causal map

Before describing the results of the Step 5 analysis, in section 4.2, it is worth describing the core structural elements of the graph as a whole in order to provide a better understanding of the representation and build confidence in its reflection of emissions trading policy.

Emissions allowance price inducing GHG emission reductions – This structure concerns the factors of 'Emissions allowance price' and 'GHG emissions' and the various connections between them, representing the various ways in which allowance price impacts emissions. From the map, there is a clear influence of allowance price on emissions via fuel switching, this is elaborated by the coal-to-gas price ratio factor. These linkages are consistent with the idea that allowance price will impact the cost of coal more than gas (as coal is more carbon-intensive), thereby inducing fuel switching. There is also the connection from allowance price to emissions via cost pass-through/production cost → energy sale price → energy demand. This causal chain reflects the idea of price increases being passed on to consumers who then reduce their energy demand as a result. Finally, there is also the path through emissions reduction investment, as higher allowance prices encourage polluters to enact measures to reduce their emissions. Besides these mechanisms, the causal map also conveys the expected impact that changes in any of these factors would have, through the direction of linkage. A lower allowance price would reduce the coal-to-gas price ratio and curtail fuel switching, it would limit the energy sale price increases thereby mediating the energy demand reductions and it would lessen the emissions reduction investment.

Emissions leakage diminishing ETS efficacy – This structure concerns the factor 'Emissions leakage', as well as its causes and effects, representing the factors contributing to leakage, and the different ways in which leakage impacts the system. This is an important structure given that leakage is a key concern for policy developers. Firstly, one can observe the obvious direct positive connection between 'ETS' regulation and 'Emissions leakage' and from 'Emissions leakage' to 'GHG emissions'. This reflects the core idea of emissions leakage, that ETS regulation incentivises polluting activity to relocate outside of regulatory jurisdictions, so avoiding ETS-induced emission reductions. Taking a closer look at the structural location of emissions leakage also reveals its negative role in some important causal loops. Consider again the path from allowance price through cost pass-through/production cost, energy demand to GHG emissions, the idea that a higher allowance price reduces emissions. Given that both cost pass-through and energy production cost also have connections to emissions leakage, the effect of allowance price on emissions is mitigated somewhat by leakage.

The role of complementary policies – This structure relates to the factor '2020 EU climate-energy package', and its two constituent factors 'Energy efficiency directive' and 'Renewable energy directive'. These represent the laws passed by the EU to help ensure that it can meet its climate targets. Given that these were the most significant climate policies, alongside the EU ETS, and that many studies have highlighted their impact on the ETS system, it is important to reflect on their inclusion in the causal map. As expected, the causal map displays a positive connection from the efficiency directive to energy efficiency, representing the energy efficiency improvements induced by this policy. Similarly, the renewable energy directive is positively linked with renewable energy utilisation, representing increased utilisation induced by this policy. The straightforward representation of the impact of these policies perhaps reflects the lack of elaboration in the analysed source material of this study. What is interesting, however, is how the causal map allows consideration of the subsequent connections, for example how increased renewable generation and subsequent emissions reductions may in turn reduce the allowance price.

Free allocation of allowances leading to windfall profits – This concerns the 'Firm profit' stock-flow structure and its connections with 'Free allocation of allowances'. This is a well-publicised sub-system within the ETS system, particularly in the earlier phases of the EU ETS whereby firms profited off the sale of freely distributed emissions allowances. The basic mechanism of this behaviour is clearly present with a path from free allocation → sale of allowances → revenues → firm profit. Several factors elaborate the causes of free allocation, including relocation risks, competitiveness concerns and political demand for new entrant provisions. Additionally, the impacts of the firm 'windfall' profits can be seen, i.e., the reduction in competitiveness concerns, and an increasing regulatory threat concerning these profits. Other studies also note the impact profits have had on firm assets and employment.

4.2. Analysis of the aggregated ETS causal map

Analysis of the causal map has been conducted in each of the three categories presented in section 3.5. While many interesting insights were revealed, it is important to note that the results in this section are by no means exhaustive, they are only meant to demonstrate some insights that can be obtained by applying step 5 analysis to the generated causal map.

Topographic analysis

Looking first at the strength of linkages, the most supported connections on the map are perhaps not surprising. Link #34, ETS → GHG emissions is supported by 9 unique papers. The impact of ETS on emissions is the core conclusion from the majority of papers and so it is heavily represented in the map. The linkage is negative in all contributing papers, which supports the conclusion that ETS has contributed to emissions reductions. Links #58 and #59, Energy demand → energy generation emissions → GHG emissions are also supported by 9 unique papers. These linkages are highly supported as a result of their central role within several important mechanisms: efficiency impacts, allowance price via energy sale price, economic conditions etc. The high degree of shared knowledge in literature for this linkage suggests that it is very relevant in explaining the ETS system, and so may be a fruitful area for policy intervention. Some presumed highly relevant links are not highly supported: #71, #61 allowance

price → emissions reduction investment, with only 2 unique papers supporting them. There are some potential reasons for this lack of support: their causal effect may be captured by other links, the linkage may not have been examined so far in the literature, or the linkage may not be relevant. In any case, the apparent lack of shared knowledge of these connections can help to encourage future research in this area.

Consider the cause and effect trees for the 'energy sale price' factor, shown in Figure 7. In the effect tree, one can observe the three ways in which sale price impacts the system. There is an obvious connection whereby energy demand is influenced by price, this, in turn, impacts generation emissions, allowance demand and sales. The connection with firm revenues indicates how energy-generating firms profit from higher sale prices. Interestingly there is also a connection to cost pass-through, supported by a causal sentence in paper 20, which argues that the degree of cost pass-through can be influenced by fluctuations in sale price. With this connection, energy sale price now has an indirect influence on numerous other important factors including emissions leakage, allowance price etc. Consider instead the cause and effect trees of ETS system linkage. If an analyst was exploring the impact of linkage between ETS systems, the effect tree would highlight that linkage is expected to reduce compliance costs and lessen price volatility as a result of increased market thickness, however amongst the numerous other effects are also capital flows between systems and associated negative public perception. The analyst could use this knowledge to take measures to mitigate the negative downstream effects, to try and sever the negative effect branches, or to introduce reinforcing branches which promote the desired behaviour.

Figure 17 (Appendix G) shows a network graph of the causal map, with nodes sized according to their degree centrality and coloured according to their betweenness. Looking first at the degree, those factors with a higher degree are largely as expected. Factors such as ETS, GHG emissions, allowance price and firm profits are known to be influential factors in the system. GHG emissions, for example, are positioned at the end of many causal paths which contribute to its high degree, allowance price is a central mechanism within trading schemes, being affected by and affecting numerous other factors. Some less obvious factors also have high degree, cost pass-through for example. Evidently many papers have analysed the factors influencing the level of pass-through and its impacts on energy pricing, competitive distortions etc. Looking at betweenness, allowance price, allowance demand, pass-through and firm profits have the highest values. This is likely a reflection of their position in the key causal paths within the system. Allowance demand has much higher betweenness than centrality indicating that it is important in determining system behaviour, but that its influence is fairly narrow. This is consistent with the idea that allowance demand has a large impact on allowance price, but does not directly influence other factors outside of this mechanism. Conversely, ETS has high degree centrality and low betweenness. This is a result of its peripheral position in the network, it influences many other factors but is not itself influenced by others, it is therefore far from the shortest path of any causal path.

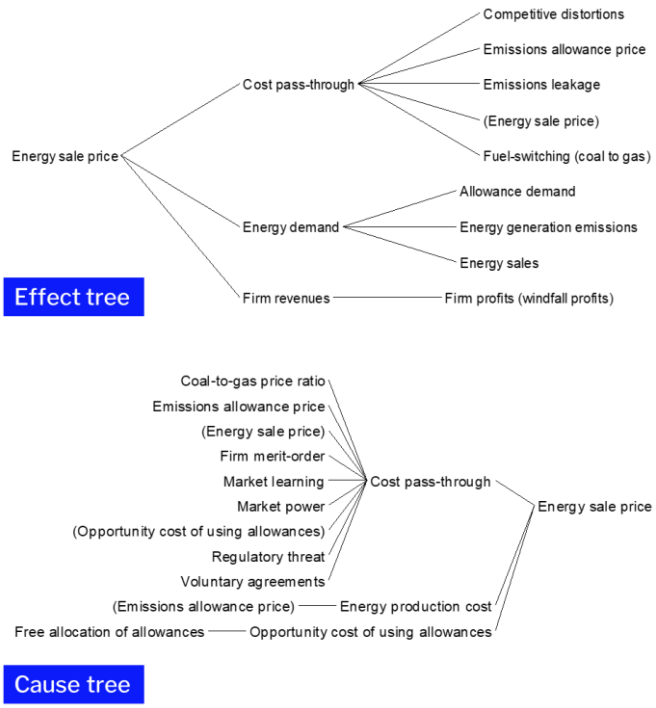

Figure 7 - Energy sale price cause and effect trees

Causal inference analysis

Consider the path from 'Emissions allowance price' +→ coal-to-gas price ratio +→ Fuel switching (coal to gas) -→ GHG emissions. The partial effect of this path is negative, consistent with the idea that a greater allowance price will lead to a price disparity between coal and gas fuel sources (given that coal is more carbon-intensive) which in turn leads to fuel switching from coal to gas thereby reducing emissions. Paper 2 supports the link between price ratio and fuel switching but was only mentioned briefly and the connection with emissions price was not included. Conversely, paper 16 notes the connection with emissions price but does not argue for the link to fuel-switching, rather noting the impact of coal and natural gas utilization respectively. If a policy developer sought to encourage fuel-switching, an understanding of the aggregated path shows that increasing the allowance price would likely contribute to this end. While a fairly simple example, this demonstrates that explicit representation of the partial effect can be useful and relevant.

Consider the total effect from 'Cost pass-through' to 'GHG emissions'. There is an obvious path via 'Energy sale price', 'Energy demand', 'Energy generation emissions' to 'GHG emissions' which has a negative partial effect indicating that greater cost pass-through would lead to emissions reductions. This is consistent with the idea that passing the cost on to consumers would reduce consumer consumption and associated emissions. Such an effect is well studied and understood, with aspects of this path covered by 15 unique papers. But taking instead the path from 'Cost pass-through' to 'Emissions leakage' to 'GHG emissions' yields a positive partial effect, consistent with the idea that greater pass-through costs contribute to greater leakage and associated net

emissions. The connection pass-through → leakage is only suggested in paper 27. In this case, then the total effect between these two factors is undetermined given the unknown relative magnitude of each path. If targeting emissions reductions, a policy developer may examine the degree of cost pass-through. If they were to only examine the energy sale price pathway then they may conclude that encouraging cost pass-through would be fruitful, examining the total effect of cost pass-through on emissions would instead reveal that carbon leakage can mitigate the reduction effect somewhat and should warrant further investigation. In other cases, the total effect *is* determined, at least to the extent of the literature coverage. Here the analysis is also useful in that it provides analysts with confidence in the causal impact of certain factor pairs. Consider the multiple paths from 'Free allocation of allowances' to 'GHG emissions'. There is the path through → energy sale price → energy demand → GHG emissions, with a positive partial effect, but also the path via Opportunity cost of using allowances → Emission reduction investment → GHG emissions also with a positive effect. In this case, then there appears to be an agreement in (current) literature that allocating free allowances appears to contribute to a relative increase in GHG emissions, through several different avenues. Policy analysis of the total effect of this connection would provide confidence in the impact of system linkage of emissions.

Causal loop analysis

Free allocations as a means to combat carbon leakage – This balancing loop demonstrates the rationale behind the free allocation of allowances as a means to combat emissions leakage. As can be seen in Figure 8, emissions leakage induces relocation risk, free allocations, therefore, are implemented, reducing the opportunity cost of allowances thereby lessening cost pass-through and so the degree of leakage. As a result, the free allocation method balances emissions leakage behaviour. What this loop also demonstrates is how this policy mechanism involves a negative partial effect from free allocation to cost pass-through, which, as has already been explained, will contribute to mitigating emissions abatement mechanisms. Causal loop analysis, in this case, has demonstrated the feedback mechanism central to the free-allocation policy intervention and has also highlighted some of its implicit negative impacts.

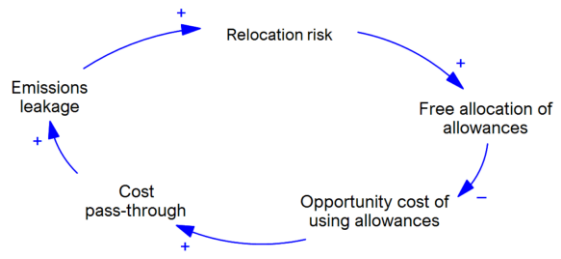

Figure 8 - Free allocation of allowances to combat leakage, causal loop

Consideration of allowance price collars – Figure 9 shows a balancing loop relating to feedback from changes in GHG emissions. Given that emissions reductions will lessen allowance demand, there is a potential dampening effect on the allowance price, which in turn impacts emissions, if allowance supply cutbacks do not keep pace with these reductions. Indeed a similar issue was experienced in the first phase of the EU ETS where an allowance oversupply contributed to an allowance price collapse (Schmalensee & Stavins, 2017). If future allowance supply caps are not set sufficiently low, then successful abatement efforts may reduce the allowance price thereby inducing a balancing effect on GHG emissions, which is undesirable when the goal is to maximise abatement. Consideration of this causal loop indicates that, alongside stringent allowance caps, a price collar could work to mitigate this issue. By imposing a price collar, such a policy intervention could limit the balancing effect of this loop thereby promoting greater emission reductions. In this case then, causal loop analysis has helped to identify potential future policy interventions, based on consideration of concerning feedback behaviour in the causal map.

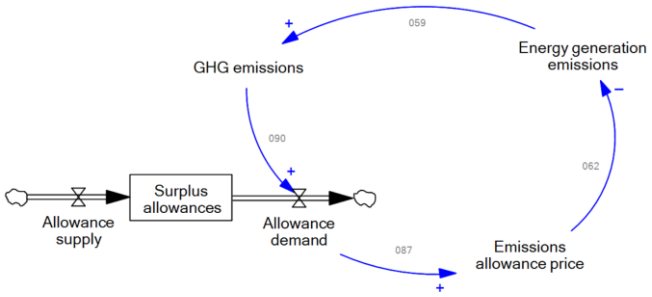

Figure 9 - Consideration of allowance price collars, causal loop

4.3. Comparisons with meta-reviews

In this section we now look again at the meta-reviews, from which the source material was selected, to examine whether their review insights are reflected in the causal map. In doing so it is possible to evaluate the extent to which an aggregated causal map representation of policy analysis insights can replace or supplement a traditional review.

Green (2021) – First, it is important to discuss the types of insights that the causal map representation struggles to reflect. Two key insights from this review, relating to emissions trading are: that the quantitative impact of emission trading, on the whole, is limited, and that the EU-ETS has only contributed to 0-1.5% annual emissions reduction. Both of these insights are difficult to ascertain from the causal map representation – the first because it is a counterfactual result that causal maps struggle to represent. The second insight is not clearly represented because this causal map is not quantified, however it may be possible to determine the emission reduction percentage by using the reference table and tracing the original text segments. In the results section, Green (2021) also notes that emission reductions vary by sector. Such insights are not visible in the current map because factors are not sub-categorised according to their respective sectors. Where the causal map performs better is in the communication of the granular features contributing to high-level behaviour. In the results conclusion section, the author notes: "Studies of emissions intensity find marginal improvements, suggesting that the ETS promotes some

degree of fuel switching". Also, that the "modest reductions can be attributed to incremental solutions: fuel switching, enhanced efficiency and reduced consumption of fuels" (Green, 2021, p. 11). Both of these points are conveyed clearly in the causal maps – in the causal chain from emissions allowance price to fuel switching, and the links between ETS, energy efficiency, and energy demand respectively. In the discussion section of the review, the author goes on to provide a broader context for understanding carbon pricing, not referring explicitly to the papers. They discuss that the consistently low carbon price has mitigated the impact of ETS on emissions and that there are many issues associated with using offset credits. Each of these discussion points is present to some degree in the causal map – by tracing the effect of reducing allowance price, the negative impact of emissions reductions can clearly be seen. Looking at the 'Access to external offset credits' factor, it can be seen that issues arise relating to market confidence, and financial risks, and that they can negatively impact allowance price.

Schmalensee & Stavins (2017) – Unlike in the Green review, Schmalensee & Stavins do not present quantified insights, but rather discuss the performance of various ETS systems and make recommendations for future implementations. These types of reflections are well suited for representation in a causal map, and indeed the insights presented in this review are largely all reflected in the causal map representation. They explain the potentially large revenues that can be generated through allowance auctions but also highlight the importance of free allowance allocations in gaining political support for ETS policy, and that these allocations are also motivated by concerns of adverse competitiveness impacts. Within the causal map, all of these structures are apparent: From ETS there is a link to allowance auction revenues (and its impacts), there is a causal path from political demand to free allocations, and a structure that indicates how firm profits impact competitiveness, and how this in-turn motivates free allocations. The review conclusions go on to explain the importance of reducing price volatility to facilitate emission abatement, noting how financial conditions led to price instability. This too is interpretable from the causal map. First, the adverse impacts of the financial crisis can be seen through its reduction of allowance demand and the subsequent impact on allowance price. Through examination of the various causal paths from allowance price to GHG emissions, it can be interpreted that fluctuating price would in turn fluctuate emission abatement, although there is no explicit connection with the price volatility factor. The insights offered in the review which are not immediately apparent in the causal map relate to the reflections on the differences in performance of different ETS implementations. For example, the authors note how carbon leakage is a significant concern, particularly for subnational systems. Given that the causal map aggregates insights from across the various ETS systems, the differences in behaviour between different systems may not be immediately obvious. Referencing the contributing text segments for each link would make this distinction possible but requires quite some review.

The comparison results have highlighted some strengths and weaknesses of this causal map. It can be seen that the generated causal map can faithfully represent the granular features contributing to ETS system behaviour and that the majority of behaviour highlighted in the reviews is captured in this representation. It performs well in distilling the factors and dynamics present in the disparate source material, making their relations clear and explicit. Where it can supplement traditional methods is in the presentation of an explicit, succinct, aggregated systems perspective of a domain. In doing so, and as described in section 3.5, it is possible to apply new analysis methods which could not easily be conducted on a purely textual review. The shortcomings of this representation relate to the conveyance of contextual information and counterfactuals. Quantified insights and insights separated according to the sector were not represented explicitly in this causal map but could be uncovered by referencing the contributing text segments. The map also overlooked some important counterfactual findings which proved to be important in one review paper

5. Discussion

5.1. Contribution toward the field of policy analysis

By synthesising derived policy effects into an aggregated causal map, this project has demonstrated a way to harmonise previously disconnected information from various levels of scope, different performance perspectives and taxonomies into one homogenous model. To my knowledge, no other method exists in literature that achieves this ambition. The five-step method proposed therefore contributes to the field of policy analysis by providing a novel perspective of the current literature-based understanding of policy effects. Not only does this approach address shortcomings of existing means of policy analysis, as explained in section 1.1, but it also provides an opportunity for new methods of analysis (presented in section 3.5) which are not easily available in purely textual policy literature. In doing so it provides analysts with a more comprehensive understanding of the factors and inter-relations which have been identified as relevant in literature and so a more comprehensive evidence base to inform their decision making. This method therefore represents a useful additional tool that analysts may use to inform policy evaluation and development.

The successful utilisation of semi-automated NLP techniques has helped to support the contribution of NLP methods (particularly relation extraction methods) to the field of policy analysis. Their application in this project reduced the manual effort required to derive causal relations from policy literature from ~3 hours to ~20 minutes which represents a significant improvement. Such improvements can allow a greater number of sources to be included in a study, thereby increasing the pool of contributing information. Additionally, the results of the chosen relation extraction algorithm were achieved with high precision, this helps to alleviate fears that automated methods may necessarily reduce the quality of analysis. The comparison of the generated causal map against the ETS meta-reviews (in section 4.3) demonstrates that

semi-automatically derived results can capture many of the same insights as a traditional manual review article. Concerns remain, however, regarding recall as the algorithm performed poorly against this metric.

Regardless of empirical algorithmic performance and the quality of final model insights, the method and implementation presented in *this* study only represents a first attempt at utilising causal relation extraction methods for the derivation of an aggregated model of policy effects. This attempt has demonstrated that usable results can be achieved in this way and that they provide value for policy analysts. As methods improve, more data becomes available and other refinements are made, future developments are likely to experience better outcomes.

5.2. Contribution toward ETS policy evaluation and development

By building a causal map from ex-post sources of emissions trading policy analysis, this project has provided an explicit systems perspective of the contributing literature-based understanding of ETS policy effects. Despite the factors and dynamics being derived from existing sources, their novel representation in a causal map format provides a clearer understanding of how ETS relevant factors identified in different sources are interrelated and how different ETS policy areas are connected. To my knowledge, such a perspective has not been previously presented in literature, it, therefore, constitutes a contribution toward ETS policy analysis in its own right.

Another contribution is in the analysis that can be conducted on the aggregated ETS causal map. As described in section 4.2 various insights can be obtained by using topographic, causal inference and causal loop analyses. Amongst other results, the causal map indicates that allowance demand is a very important factor in the system but that its influence is narrow, restricted to its impact on allowance price. Causal loop analysis has suggested that an allowance price collar may help to maintain emissions reductions despite allowance demand fluctuations and cause-effect tree analysis has made explicit the multitude of factors influencing and being influenced by the energy sale price factor. Additional analysis, conducted with a particular policy ambition in mind, and with deeper consideration of the causal map (and contributing text segments) is likely to reveal further and more specific insights.

5.3. Limitations and future research

The limitations of the approach proposed in this project can be broadly categorised into two groups, methodological limitations – those limiting the *quality* of the final output causal map, and format limitations – those limiting the *value* of the final output causal map. Future research can focus on addressing these limitations or on other development areas.

Methodological limitations and future research

Recall – The primary methodological limitation of this approach is the low recall (~38%) achieved by the implemented causal relation extraction algorithm. This means that currently only ~38% of causal relations present in the source material will be included in the final causal map. This severely limits confidence in the sufficiency of the systems perspective presented, and subsequent analysis. Nevertheless, this limitation only reflects the poor empirical performance of the chosen algorithm on this dataset and is not an inherent issue of the method. As has been discussed in Appendix E: SCITE output, recall and precision metrics, the current low recall is likely a result of insufficient training data. To improve confidence in the results obtained using this method, future work should focus on improving the recall of the causal relation extraction algorithm.

Structural gaps – Another issue is that of structural gaps existing in the final causal map, some expected structural connections are missing from the causal map. For example, there is an obvious connection between 'ETS system linkage' and allowance demand, as linkage effectively increases the number of market participants, and so the number of participants demanding allowances. This connection is partially captured in the link from linkage → market thickness, however, there is no connection from here to allowance demand. Again, this is not an inherent drawback of the method, but rather an artifact of the low recall and/or lack of description in the source material. Much like the recall limitation, the existence of structural gaps limits confidence in the conclusions drawn from the model because, without an accurate representation of the system, certain behaviour is likely to be missed. The solution to this issue is less clear, one avenue to explore in future research is the potential of first building a foundational model of widely accepted system behaviour from which extraction relations could be added. In this way, core structural elements can be guaranteed whilst still generating insights from the derived model structure.

What constitutes a causal relation – The inclusion of a causal link is contingent on textual argumentation for a cause-effect relationship between factors. Difficulties arise, however, given that it can often be difficult to determine what constitutes a strictly causal relation. For example, in the link: carbon leakage → relocation risk, the supporting text segment states "In order to protect industry from potential relocation risks … sectors deemed at risk of carbon leakage … qualify for free allowances". The expectation of relocation risk, amongst other factors, was used as one rationale for providing free allowances, yet in phases I and II of the EU ETS free allocations were provided regardless of relocation risk, and at the discretion of member states. Should there then be a link: relocation risk → free allowances? Only focusing on strictly causal links may give the impression that other relations are not important to the behaviour of the policy. Addressing this issue is would require careful consideration of the contrasting views of what constitutes causality (discussed briefly in Appendix E: SCITE output, recall and precision metrics), but must also recognise differences in language used by different authors. Some authors naturally present more assertive language which can more easily be defined as causal, whilst others may use a more passive voice which can lead to ambiguity over a causal categorisation. Future work to clarify what notions of causality should be included would be a valuable development for this approach.

Insufficient data – This method relies on existing policy analysis literature from which to derive the causal relations and build the model structure. In cases where there is limited literature, or literature of insufficient quality, both the quality

of the causal map and the quality of the subsequent analysis, will be negatively impacted. The quality of the policy analysis data, therefore, imposes a significant limitation on the application of this method.

Bias – The application of this method for ETS policy was conducted by a single researcher. There exists a potential limitation in the degree of bias presented by the researcher, however, this remains to be tested. Future research could include triangulation to determine the impartiality of the method.

Format limitations and future research

Specificity – By aggregating insights from across disparate sources, the causal map inevitably lacks specificity. As a result, quantified analysis of specific outcomes in specific cases is not easily attainable. This presents a limitation in the use of the method, it should not be seen as a quantified tool to evaluate policy options, but rather as a qualitative means to understand the current literature-based perspective of the policy landscape.

Lack of context – A core issue of causal map representations is the limited context conveyed in their representation of the system. For example, they struggle to communicate quantified relations, differences in behaviour between sub-categories of the system, and the constraints imposed on the insights by the methodology utilised. By supplementing the causal map with a reference table of the contributing text segments, this limitation has been addressed somewhat, however, this process is not convenient and often the text segments do not capture the full context. Future research on how to augment the causal map with additional context is therefore warranted.

Counterfactuals – Finally, it is apparent that policy analysis frequently reaches counterfactual conclusions, conclusions that state the *lack* of relationship between two factors. Given that causal maps only represent causal behaviour such insights are overlooked. This can lead to the false conclusion that causal relations between factors included in the map represent the current literary understanding of their relationship. In reality, there may be a more supported belief that they are unrelated. Addressing this limitation would require adjustment to the format of causal maps so that counterfactuals can, in some way, be included with causal relations.

Other development areas

Alongside these improvements, the application of the current method to new policy domains would be a valid avenue for future research. Doing so would help to further test the efficacy of this approach and its portability. Additionally, causal maps provide an ideal foundation for the generation of system dynamics models. Future research could therefore test the suitability of causal maps generated from policy analysis source material as a basis for more quantified models of policy behaviour.

6. Conclusions

Given the growing threat of global warming, the urgency of the energy transition, and the seriousness of so many other grand challenges, we must endeavour to remedy past policy failures and maximise the utility of future interventions. Unfortunately, however, existing means of policy analysis suffer from several shortcomings which limit their ability to achieve this ambition. Due to the disparities in scope, taxonomy and performance perspectives employed by different studies, it is difficult to obtain a comprehensive system perspective of the policy effects of a certain policy domain. Additionally, a prohibitively large body of literature often makes manual methods of review infeasible. To help combat these issues, the objective of this research project was to explore the development of a semi-automated method to aggregate causal relations from policy analysis literature into a causal map of policy behaviour so that it may help inform policy analysis and development. This project has demonstrated one such method which uses a deep learning-based relation extraction algorithm to derive causal relations, a semantic clustering approach to group similar factors, a process to build a final aggregated causal map, and a description of the various types of analysis that may be conducted. The application of this method to a collection of 28 emissions trading policy analysis literature sources has helped to prove its efficacy and demonstrated that the analysis of the resulting causal map can deliver policy-relevant insights, including those not readily available in traditional forms of analysis. As a result, I propose that this novel method represents a promising development in the field of policy analysis.

Acknowledgements

My sincerest thanks to everyone who helped and supported me throughout my thesis journey. Thanks to my graduation committee for their valuable guidance: Nihit for your weekly wisdom, Lisa and Kornelis for your helpful comments and enthusiasm. I would also like to thank all my friends and family, particularly everyone from Simonsstraat who have made these last two years a joy.

References

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., & Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, *9*(11), S2. https://doi.org/10.1186/1471-2105-9-S11-S2

Amer, M., Daim, T. U., & Jetter, A. (2013). A review of scenario planning. *Futures*, *46*, 23–40. https://doi.org/10.1016/j.futures.2012.10.003

Anderson, B., & Di Maria, C. (2011). Abatement and Allocation in the Pilot Phase of the EU ETS. *Environmental and Resource Economics*, *48*(1), 83–103. https://doi.org/10.1007/s10640-010-9399-9

Arimura, T. H., & Abe, T. (2021). The impact of the Tokyo emissions trading scheme on office buildings: What factor contributed to the emission reduction? *Environmental Economics and Policy Studies*, *23*(3), 517–533. https://doi.org/10.1007/s10018-020-00271-w

Asghar, N. (2016). *Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey*. https://arxiv.org/abs/1605.07895v1

Axelrod, R. (1976). *Structure of Decision: The Cognitive Maps of Political Elites*. Princeton University Press.

Bach, N., & Badaskar, S. (2007). A review of relation extraction. *Literature Review for Language and Statistics II*, *2*, 1–15.

Barik, B., Marsi, E., & Öztürk, P. (2016). Event Causality Extraction from Natural Science Literature. *Research in Computing Science*, *117*(1), 97–107. https://doi.org/10.13053/rcs-117-1-8

Bayer, P., & Aklin, M. (2020). The European Union Emissions Trading System reduced CO2 emissions despite low prices. *Proceedings of the National Academy of Sciences*, *117*(16), 8804–8812. https://doi.org/10.1073/pnas.1918128117

Beamer, B., Rozovskaya, A., & Girju, R. (2008). Automatic Semantic Relation Extraction with Multiple Boundary Generation. *AAAI*, 824–829.

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The Oxford handbook of causation*. Oxford University Press.

Bel, G., & Joseph, S. (2015). Emission abatement: Untangling the impacts of the EU ETS and the economic crisis. *Energy Economics*, *49*, 531–539. https://doi.org/10.1016/j.eneco.2015.03.014

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Bontis, N., & Fitz-enz, J. (2002). Intellectual capital ROI: A causal map of human capital antecedents and consequents. *Journal of Intellectual Capital*, *3*(3), 223–247. https://doi.org/10.1108/14691930210435589

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Bovens, M., & 't Hart, P. (2016). Revisiting the study of policy failures. *Journal of European Public Policy*, *23*(5), 653–666. https://doi.org/10.1080/13501763.2015.1127273

Bui, Q.-C., Nualláin, B. Ó., Boucher, C. A., & Sloot, P. M. (2010). Extracting causal relations on HIV drug resistance from literature. *BMC Bioinformatics*, *11*(1), 101. https://doi.org/10.1186/1471-2105-11-101

Burtraw, D., Kahn, D., & Palmer, K. (2006). CO2 Allowance Allocation in the Regional Greenhouse Gas Initiative and the Effect on Electricity Investors. *The Electricity Journal*, *19*(2), 79–90. https://doi.org/10.1016/j.tej.2006.01.001

Chowdhary, K. R. (2020). Natural Language Processing. In K. R. Chowdhary (Ed.), *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

Convery, F. J. (2020). Reflections—The emerging literature on emissions trading in Europe. *Review of Environmental Economics and Policy*.

Convery, F. J., & Redmond, L. (2007). Market and Price Developments in the European Union Emissions Trading Scheme. *Review of Environmental Economics and Policy*, *1*(1), 88–111. https://doi.org/10.1093/reep/rem010

Copley, B., & Wolff, P. (2014). Theories of causation should inform linguistic theory and vice versa. *Causation in Grammatical Structures*, 11–57.

Cullenward, D. (2014). Leakage in California's Carbon Market. *The Electricity Journal*, *27*(9), 36–48. https://doi.org/10.1016/j.tej.2014.09.014

Dechezleprêtre, A., Nachtigall, D., & Venmans, F. (2018). *The joint impact of the European Union emissions trading system on carbon emissions and economic performance*. OECD. https://doi.org/10.1787/4819b016-en

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Dong, J., Ma, Y., & Sun, H. (2016). From Pilot to the National Emissions Trading Scheme in China: International Practice and Domestic Experiences. *Sustainability*, *8*(6), 522. https://doi.org/10.3390/su8060522

Drury, B., Oliveira, H. G., & Lopes, A. de A. (2022). A survey of the extraction and applications of causal relations. *Natural Language Engineering*, *28*(3), 361–400. https://doi.org/10.1017/S135132492100036X

Duan, M., Pang, T., & Zhang, X. (2014). Review of Carbon Emissions Trading Pilots in China. *Energy & Environment*, *25*(3–4), 527–549. https://doi.org/10.1260/0958-305X.25.3-4.527

Eden, C., Ackermann, F., & Cropper, S. (1992). The Analysis of Cause Maps. *Journal of Management Studies*, *29*(3), 309–324. https://doi.org/10.1111/j.1467-6486.1992.tb00667.x

Egenhofer, C., Alessi, M., Georgiev, A., & Fujiwara, N. (2011). *The EU Emissions Trading System and Climate Policy Towards 2050: Real Incentives to Reduce Emissions and Drive Innovation?* (SSRN Scholarly Paper No. 1756736). Social Science Research Network. https://papers.ssrn.com/abstract=1756736

Ekici, A., & Onsel, S. (2013). How Ethical Behavior of Firms is Influenced by the Legal and Political Environments: A Bayesian Causal Map Analysis Based on Stages of Development. *Journal of Business Ethics*, *115*(2), 271–290. https://doi.org/10.1007/s10551-012-1393-4

Ellerman, A. D., & Buchner, B. K. (2007). The European Union Emissions Trading Scheme: Origins, Allocation, and Early Results. *Review of Environmental Economics and Policy*, *1*(1), 66–87. https://doi.org/10.1093/reep/rem003

Ellerman, A. D., & Buchner, B. K. (2008). Over-Allocation or Abatement? A Preliminary Analysis of the EU ETS Based on the 2005–06 Emissions Data. *Environmental and Resource Economics*, *41*(2), 267–287. https://doi.org/10.1007/s10640-008-9191-2

Ellerman, A. D., Marcantonini, C., & Zaklan, A. (2016). The European Union Emissions Trading System: Ten Years and Counting. *Review of Environmental Economics and Policy*, *10*(1), 89–107. https://doi.org/10.1093/reep/rev014

Ellerman, A. D., & McGuinness, M. (2008). *CO2 Abatement in the UK Power Sector: Evidence from the EU ETS Trial Period* [Working Paper]. https://dspace.mit.edu/handle/1721.1/45654

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*(34), 226–231.

Esty, D. C., & Porter, M. E. (2005). National environmental performance: An empirical analysis of policy results and determinants. *Environment and Development Economics*, *10*(4), 391–434. Scopus. https://doi.org/10.1017/S1355770X05002275

European Commission. (2021). *EU Emissions Trading System (EU ETS)*. https://ec.europa.eu/clima/eu-action/eu-emissions-trading-system-eu-ets_en

Fell, H., & Maniloff, P. (2018). Leakage in regional environmental policy: The case of the regional greenhouse gas initiative. *Journal of Environmental Economics and Management*, *87*, 1–23. https://doi.org/10.1016/j.jeem.2017.10.007

Fleuren, W. W. M., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, *74*, 97–106. https://doi.org/10.1016/j.ymeth.2015.01.015

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2009). Classification of semantic relations between nominals. *Language Resources and Evaluation*, *43*(2), 105–121. https://doi.org/10.1007/s10579-009-9083-2

Gloaguen, O., & Alberola, E. (2013). Assessing the factors behind CO2 emissions changes over the phases 1 and 2 of the EU ETS: An econometric analysis. *CDC Climat Research, Paris, France*.

Goulder, L. H., & Stavins, R. N. (2011). Challenges from State-Federal Interactions in US Climate Change Policy. *American Economic Review*, *101*(3), 253–257. https://doi.org/10.1257/aer.101.3.253

Green, J. F. (2021). Does carbon pricing reduce emissions? A review of ex-post analyses. *Environmental Research Letters*, *16*(4), 043004. https://doi.org/10.1088/1748-9326/abdae9

Haites, E., Maosheng, D., Gallagher, K. S., Mascher, S., Narassimhan, E., Richards, K. R., & Wakabayashi, M. (2018). Experience with carbon taxes and greenhouse gas emissions trading systems. *Duke Envtl. L. & Pol'y F.*, *29*, 109.

Han, H., Wang, Q., & Chen, C. (2019). Policy Text Analysis Based on Text Mining and Fuzzy Cognitive Map. *2019 15th International Conference on Computational Intelligence and Security (CIS)*, 142–146. https://doi.org/10.1109/CIS.2019.00038

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., Pennacchiotti, M., Romano, L., & Szpakowicz, S. (2019). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *ArXiv Preprint ArXiv:1911.10422*.

Hibbard, P. J., Okie, A. M., Tierney, S. F., & Darling, P. G. (2015). The economic impacts of the regional greenhouse gas initiative on nine northeast and mid-Atlantic states. *Analysis Group, July*.

Howlett, M., Ramesh, M., & Wu, X. (2015). Understanding the persistence of policy failures: The role of politics, governance and uncertainty. *Public Policy and Administration*, *30*(3–4), 209–220. https://doi.org/10.1177/0952076715593139

ICAP. (2022a). *About Emissions Trading Systems*. International Carbon Action Partnership. https://icapcarbonaction.com/en/about-emissions-trading-systems

ICAP. (2022b). *Emissions Trading Worldwide – International Carbon Action Partnership (ICAP) Status Report 2022*. 240.

Jaraite-Kažukauske, J., & Di Maria, C. (2016). Did the EU ETS make a difference? An empirical assessment using Lithuanian firm-level data. *The Energy Journal*, *37*(1).

Jiali, L. (1995). China's one-child policy: How and how well has it worked? A case study of Hebei Province, 1979-88. *Population & Development Review*, *21*(3), 563–585. Scopus.

Kahn, H., & Wiener, A. J. (1967). *A framework for speculation on the next thirty-three years*. NewYork, Macmillan.

Khoo, C. S., & Na, J.-C. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*(1), 157–228.

Kim, H., & Andersen, D. F. (2012). Building confidence in causal maps generated from purposive text data: Mapping transcripts of the Federal Reserve. *System Dynamics Review*, *28*(4), 311–328. https://doi.org/10.1002/sdr.1480

Kim, J., Han, M., Lee, Y., & Park, Y. (2016). Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. *Expert Systems with Applications*, *57*, 311–323. Scopus. https://doi.org/10.1016/j.eswa.2016.03.043

Kotnik, Ž., Maja, K., & Škulj, D. (2014). The effect of taxation on greenhouse gas emissions. *Transylvanian Review of Administrative Sciences*, *10*(43), 168–185.

Kruger, J., Oates, W. E., & Pizer, W. A. (2007). Decentralization in the EU Emissions Trading Scheme and Lessons for Global Policy. *Review of Environmental Economics and Policy*, *1*(1), 112–133. https://doi.org/10.1093/reep/rem009

Kyriakakis, M., Androutsopoulos, I., Ametllé, J. G. i, & Saudabayev, A. (2019). Transfer Learning for Causal Sentence Detection. *ArXiv:1906.07544 [Cs]*. http://arxiv.org/abs/1906.07544

Laing, T., Sato, M., Grubb, M., & Comberti, C. (2014). The effects and side-effects of the EU emissions trading scheme. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(4), 509–519.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Larsen, P., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*(3), 575–603.

Leining, C., Kerr, S., & Bruce-Brand, B. (2020). The New Zealand Emissions Trading Scheme: Critical review and future outlook for three design innovations. *Climate Policy*, *20*(2), 246–264. https://doi.org/10.1080/14693062.2019.1699773

Li, Z., Li, Q., Zou, X., & Ren, J. (2021). Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, *423*, 207–219. https://doi.org/10.1016/j.neucom.2020.08.078

Liddy, E. D. (2001). *Natural language processing*.

Löfgren, Å., Burtraw, D., Wråke, M., & Malinovskaya, A. (2015). *Architecture of the EU emissions trading system in phase 3 and the distribution of allowance asset values*.

Marsh, D., & McConnell, A. (2010). Towards a Framework for Establishing Policy Success. *Public Administration*, *88*(2), 564–583. https://doi.org/10.1111/j.1467-9299.2009.01803.x

Martin, G., & Saikawa, E. (2017). Effectiveness of state climate and energy policies in reducing power-sector CO2 emissions. *Nature Climate Change*, *7*(12), 912–919. https://doi.org/10.1038/s41558-017-0001-0

Martin, R., Muûls, M., & Wagner, U. (2012). An evidence review of the EU Emissions Trading System, focussing on effectiveness of the system in driving industrial abatement. *Department of Eenrgy and Climate Change*.

Martin, R., Muûls, M., & Wagner, U. J. (2020). The impact of the European Union Emissions Trading Scheme on regulated firms: What is the evidence after ten years? *Review of Environmental Economics and Policy*.

Mascher, S. (2018). Striving for equivalency across the Alberta, British Columbia, Ontario and Québec carbon pricing systems: The Pan-Canadian carbon pricing benchmark. *Climate Policy*, *18*(8), 1012–1027.

Montibeller, G., & Belton, V. (2006). Causal maps and the evaluation of decision options—A review. *Journal of the Operational Research Society*, *57*(7), 779–791. https://doi.org/10.1057/palgrave.jors.2602214

Murray, B. C., & Maniloff, P. T. (2015). Why have greenhouse emissions in RGGI states declined? An econometric attribution to economic, energy market, and policy factors. *Energy Economics*, *51*, 581–589. https://doi.org/10.1016/j.eneco.2015.07.013

Murray, B., & Rivers, N. (2015). British Columbia's revenue-neutral carbon tax: A review of the latest "grand experiment" in environmental policy. *Energy Policy*, *86*, 674–683.

Nadkarni, S., & Narayanan, V. K. (2005). Validity of the Structural Properties of Text-Based Causal Maps: An Empirical Assessment. *Organizational Research Methods*, *8*(1), 9–40. https://doi.org/10.1177/1094428104271999

Narassimhan, E., Gallagher, K. S., Koester, S., & Alejo, J. R. (2017). Carbon pricing in practice: A review of the evidence. *Climate Policy Lab: Medford, MA, USA*.

Nunez-Mir, G. C., Iannone III, B. V., Pijanowski, B. C., Kong, N., & Fei, S. (2016). Automated content analysis: Addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, *7*(11), 1262–1272. https://doi.org/10.1111/2041-210X.12602

Ohlendorf, N., Jakob, M., Minx, J. C., Schröder, C., & Steckel, J. C. (2021). Distributional Impacts of Carbon Pricing: A Meta-Analysis. *Environmental and Resource Economics*, *78*(1), 1–42. https://doi.org/10.1007/s10640-020-00521-1

Pakray, P., & Gelbukh, A. (2014). An Open-Domain Cause-Effect Relation Detection from Paired Nominals. In A. Gelbukh, F. C. Espinoza, & S. N. Galicia-Haro (Eds.), *Nature-Inspired Computation and Machine Learning* (pp. 263–271). Springer International Publishing. https://doi.org/10.1007/978-3-319-13650-9_24

Petrick, S., & Wagner, U. J. (2014). *The Impact of Carbon Trading on Industry: Evidence from German Manufacturing Firms* (SSRN Scholarly Paper No. 2389800). Social Science Research Network. https://doi.org/10.2139/ssrn.2389800

Psillos, S. (2007). Causal explanation and manipulation. In J. PERSSON & P. YLIKOSKI (Eds.), *Rethinking*

Explanation (pp. 93–107). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5581-2_6

Ranson, M., & Stavins, R. N. (2012). *Post-Durban climate policy architecture based on linkage of cap-and-trade systems*. National Bureau of Economic Research.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. http://arxiv.org/abs/1908.10084

Sartor, O., Pallière, C., & Lecourt, S. (2014). Benchmark-based allocations in EU ETS Phase 3: An early assessment. *Climate Policy*, *14*(4), 507–524. https://doi.org/10.1080/14693062.2014.872888

Schmalensee, R., & Stavins, R. N. (2017). Lessons Learned from Three Decades of Experience with Cap and Trade. *Review of Environmental Economics and Policy*, *11*(1), 59–79. https://doi.org/10.1093/reep/rew017

Sijm, J., Neuhoff, K., & Chen, Y. (2011). CO2 cost pass-through and windfall profits in the power sector. *Climate Policy*, *6*(1), 49–72. https://doi.org/10.1080/14693062.2006.9685588

Son, C., Kim, J., & Kim, Y. (2020). Developing scenario-based technology roadmap in the big data era: An utilisation of fuzzy cognitive map and text mining techniques. *Technology Analysis & Strategic Management*, *32*(3), 272–291. https://doi.org/10.1080/09537325.2019.1654091

Thow, A. M., Swinburn, B., Colagiuri, S., Diligolevu, M., Quested, C., Vivili, P., & Leeder, S. (2010). Trade and food policy: Case studies from three Pacific Island countries. *Food Policy*, *35*(6), 556–564. Scopus. https://doi.org/10.1016/j.foodpol.2010.06.005

Venmans, F. (2012). A literature-based multi-criteria evaluation of the EU ETS. *Renewable and Sustainable Energy Reviews*, *16*(8), 5493–5510.

Wagner, U. J., Muûls, M., Martin, R., & Colmer, J. (2014). The causal effects of the European Union Emissions Trading Scheme: Evidence from French manufacturing plants. *Fifth World Congress of Environmental and Resources Economists, Instanbul, Turkey*.

Wakabayashi, M., & Kimura, O. (2018). The impact of the Tokyo Metropolitan Emissions Trading Scheme on reducing greenhouse gas emissions: Findings from a facility-based study. *Climate Policy*, *18*(8), 1028–1043. https://doi.org/10.1080/14693062.2018.1437018

Warner, J., & van Buuren, A. (2011). Implementing room for the river: Narratives of success and failure in Kampen, the Netherlands. *International Review of Administrative Sciences*, *77*(4), 779–801. Scopus. https://doi.org/10.1177/0020852311419387

Wing, I. S., & Kolodziej, M. (2008). *The Regional Greenhouse Gas Initiative: Emission Leakage and the Effectiveness of Interstate Border Adjustments*.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Wu, R., Yao, Y., Han, X., Xie, R., Liu, Z., Lin, F., Lin, L., & Sun, M. (2019). Open Relation Extraction: Relational Knowledge Transfer from Supervised Data to Unsupervised Data. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 219–228. https://doi.org/10.18653/v1/D19-1021

Yang, J., Han, S. C., & Poon, J. (2021). A Survey on Extraction of Causal Relations from Natural Language Text. *ArXiv:2101.06426 [Cs]*. http://arxiv.org/abs/2101.06426

Zhao, S., Liu, T., Zhao, S., Chen, Y., & Nie, J.-Y. (2016). Event causality extraction based on connectives analysis. *Neurocomputing*, *173*, 1943–1950. https://doi.org/10.1016/j.neucom.2015.09.066

Appendix A: Causal maps for policy analysis

This appendix evaluates the extent to which a causal map, constructed from a subset of the emissions trading policy analysis sources, is sufficient to represent the policy causal insights present in the contributing paper.

Causal maps as a means of reflecting on policy analysis insights

As mentioned in the introduction, causal maps are a useful form of conceptual model which seek to capture and graphically represent the causal functioning of a system. Multiple studies have examined the many ways in which they can be used for analysis (Eden et al., 1992; Montibeller & Belton, 2006; Nadkarni & Narayanan, 2005) and yet more have demonstrated their successful application in a multitude of domains (Bontis & Fitz-enz, 2002; Ekici & Onsel, 2013; H. Kim & Andersen, 2012).

To test their efficacy in capturing and representing the causal insights of policy analysis, this section will now manually build and examine causal maps from the ETS sources and subsequently compare them against the papers themselves. By using a manual method, it is possible to test the efficacy independent of any NLP-related issues and using a highly cited literature-based method helps to prevent any bias that may occur by methodological issues.

The manual construction method used is that presented by Kim & Andersen (2012) which consists of five steps, described in Table 2. The source papers are first examined using open coding to discover themes in the data and to select relevant data segments. From here data is axially coded into individual 'coding charts' which denote specific arguments consisting of cause-and-effect relations which are then generalised and built into a final causal map. This method was used to create four causal maps from a random selection of source papers (2,7,10 and 13).

Step	Description of the process	Input	Output
1	Discovering themes in the data	Raw text data	Definition of problem and system boundary; selection of relevant data segments
2	Identifying variables and their causal relationships	Data segments (each segment = one argument + supporting rationales)	Coding charts
3	Transforming text into word-and-arrow diagrams	Coding charts	Simple words-and-arrow diagrams
4	Generalising structural representations	Simple words-and-arrow diagrams	Final causal map
5	Linking maps to the data source	Coding charts and final cause map	Data source reference table

Table 2 - Manual causal map construction process, adapted from (H. Kim & Andersen, 2012)

Analysis of manual causal map results

Overview of manual causal maps

Figure 10 shows the manually derived causal maps. Each of the maps is evidently quite detailed with many contributing factors and directional causal connections.

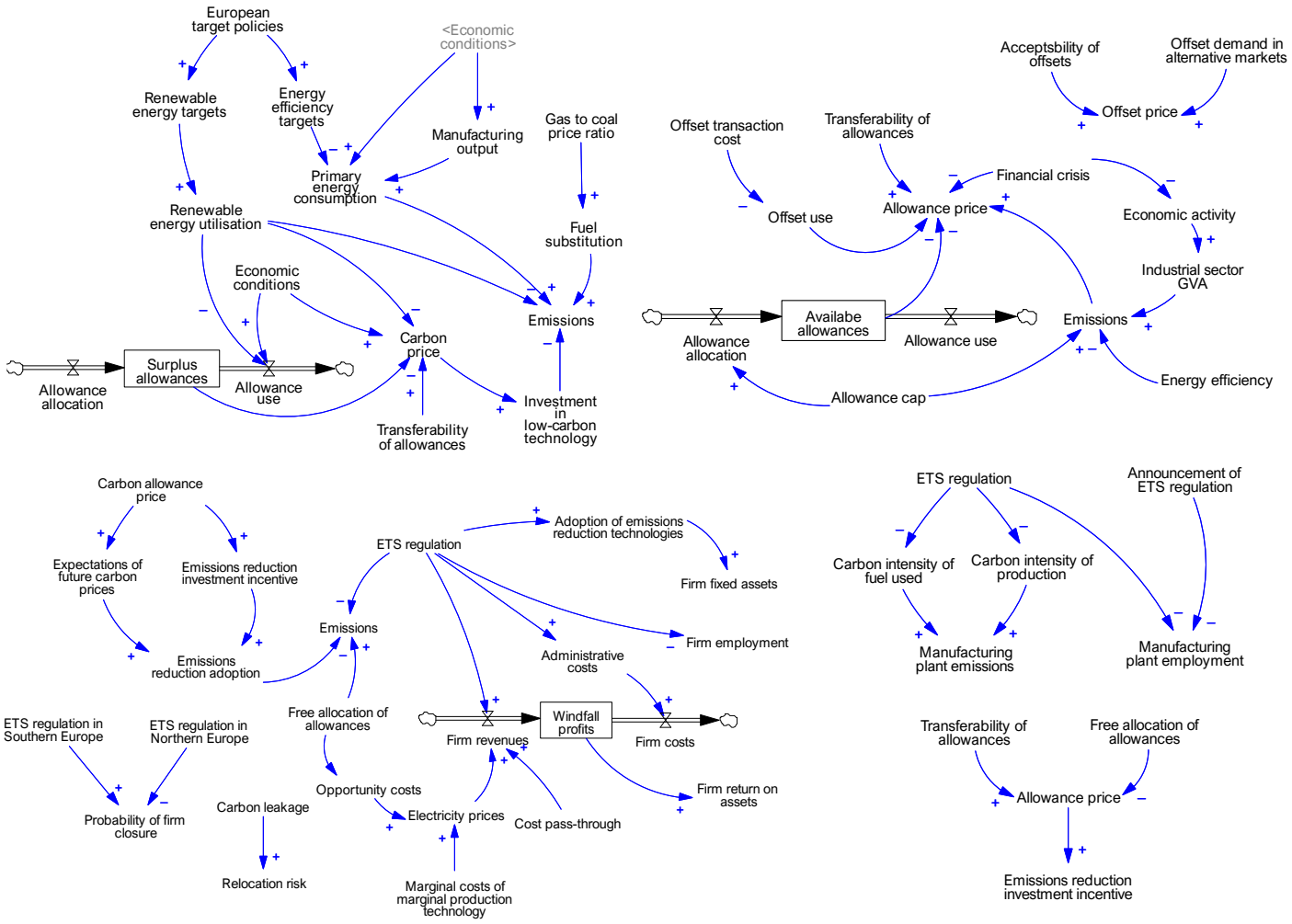

Figure 10 - Manually derived causal maps for papers 2, 7, 10 and 13

Analysis of manually derived causal map for paper 2

To explore the coverage and relevance of each causal map, a detailed comparison of each map and source paper was conducted. The comparison for paper 2 is described below in Figure 11.

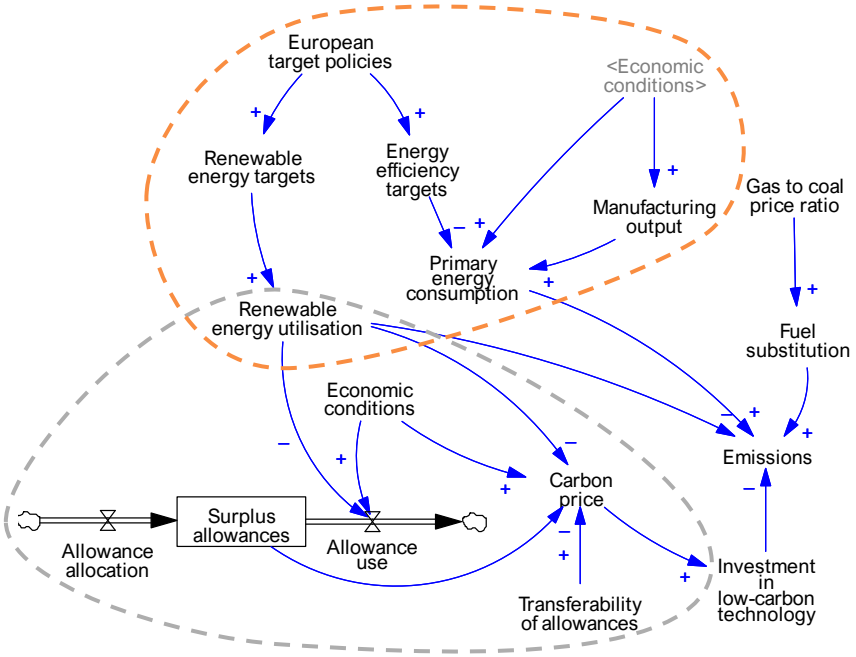

Figure 11 - Examination of the manually derived paper 2 causal map

The focus of this paper is to explore the contributing factors to the emissions changes during the initial stages of the EU ETS. From the abstract, the key findings of this paper can be summarised as emissions reductions stemming from the climate and energy package policies, the economic downturn, and price substitution effects. No impact from the carbon price was found, however, the economic downturn and renewable energy utilisation contributed to its low price thereby marginalising its influence (Gloaguen & Alberola, 2013).

The causal map for paper 2 presents a largely cohesive and comprehensive collection of factors with many inter-relations and no isolated structures. The map has very good coverage of the key results as presented in the abstract: there are separate factors indicating the individual components of the climate and energy package policies, in the orange area, the causal structure of ETS allowances in grey, as well as factors relating to economic conditions, and fuel substitution effects (synonymous with price substitution). Each has a clear causal connection denoting their emission reduction effect. Regarding the contributing factors for carbon price, the map also contains a clear description of how poor economic conditions and greater renewable energy utilisation reduced allowance demand thereby contributing to greater allowance surpluses, in turn depressing the carbon price and marginalising the carbon price's impact on emissions. One important omission, however, is a representation of the finding that no impact from the carbon price, in terms of emissions reduction, was found. Indeed, the link from carbon price → investment → emissions is present and may give a false impression that this connection was found to be important. The issue arises from the fact that this result is counterfactual – it expresses what *does not* happen. A causal map struggles to depict such insights.

Strengths and weaknesses of causal map representations

To convey a deeper understanding of the value of this type of conceptual model in policy analysis, the relative strengths and weaknesses of causal maps, as experienced in the analysis of the four emissions trading papers, are discussed:

Firstly, in all analysed cases, the causal maps provided comprehensive coverage of the points addressed in the abstract, usually with several more granular contributing factors included. This suggests that causal structures *are* sufficient to reconstruct the core findings of a paper and provide confidence in its sufficiency as a means to represent the results as a whole.

Secondly, this format provides a much more digestible representation of findings than an often 10+ page textual description. In the original format, causal connections are often buried in descriptive text and the interconnections between factors may not be apparent when they are spread between different paragraphs or sections. The direction of linkages between factors can also easily be overlooked. For example, in the passage "In order to protect industry from potential relocation risks, a list of sectors deemed at risk of carbon leakage was designed by the European Commission", it may be easy to miss the implicit positive connection carbon leakage → relocation risk. A causal map explicitly conveys these insights.

The factors identified in the source material can be aggregated, thereby avoiding any unnecessary repetition and uncovering connections between factors that may not have been immediately obvious given the use of different descriptive terms. For example, in paper 13, the causal connection between ETS regulation and firm revenues was explained on seven separate occasions throughout the paper, often using different terminology. In paper 2: the factor 'Renewable energy utilisation' is referred to (again using various terminology) both in connection with the impact of European climate and energy policies and in connection with allowance price factors. In the causal map, this factor's multi-functional role in the system is made clear by its connection to both structures.

The primary weakness of this application of causal maps is the lack of supporting context which underpins its representation:

- Whilst the papers often provide clear indications of the quantified impact of certain factors on one another, this is not visible in the map format. For example, paper 13 argues that the EU ETS induced emissions reductions of 10%, whilst the map only indicates a negative linkage.
- A causal map also cannot clearly display the disparate strength of linkages between different subsets of factors: for example, in paper 13, the paper argues that the heat and electricity sectors experienced a greater economic benefit from ETS regulation, compared to for example the manufacturing sector.
- Similarly, it can be difficult to ascertain the hierarchy of the linkages. Take, for example (in paper 13), the linkage ETS → emissions, and the separate linkage carbon price → emissions reduction investment incentive → emissions reduction adoption → emissions. Excerpts from the source material support both linkages, however, it is unclear the extent to which the linkage from carbon price to emissions is included in the higher-level ETS to emissions link. This issue is less of a criticism of the causal map format but rather an issue that arises when trying to explicitly display the various linkages present in a paper.
- Lastly, this representation lacks communication of the methodology underpinning the results, the methodology is important to contextualise the findings and to inform the reader of the limitations of the results. A causal is not designed to provide such insights.

A more process-related issue is the impact of how causal links are defined. The inclusion of a linkage is contingent on textual argumentation for a cause-effect relationship between factors. Questions arise on what degree of implicit relationships should be included. For example, consider the link (in paper 10): emissions → allowance price, it would perhaps seem obvious that there would also be a reciprocal relationship: allowance price → emissions, however, this point is not addressed within the text, so it was not included in the map. In such cases, the lack of implicit links may prevent a comprehensive description of the system.

Additionally, even when there is textual argumentation for a connection between factors, it can be difficult to determine what constitutes a strictly causal relation. For example, in the link: carbon leakage → relocation risk, the supporting text segment states "In order to protect industry from potential relocation risks … sectors deemed at risk of carbon leakage … qualify for free allowances". The expectation of relocation risk, amongst other factors, was used as one rationale for providing free allowances, yet in phases I and II of the EU ETS free allocations were provided regardless of relocation risk, and at the discretion of member states. Should there then be a link: relocation risk → free allowances? Only focusing on strictly causal links may give the impression that other relations are not important to the behaviour of the policy.

Finally, in a causal map representation, there is no communication of counterfactual conclusions, however, such insights can be useful. Excluding them in this representation, therefore, misses some potentially important information. For example, the results section states that "… we do not find a statistically significant effect for the medium-sized installations (2nd and 3rd quartile) … meaning that the emissions of ETS installations have not declined [relative to non-ETS installations]".

Discussion on causal map efficacy

Analysis has shown that causal maps can be used to graphically represent the causal relations present in policy analysis literature and that they can capture the conclusions from these papers. Its strength lies in its ability to explicitly convey the aggregated inter-relations between factors. Where it is less effective, however, is in communicating the more intricate supporting information that underpins these causal links, and in communicating counterfactual conclusions.

The issue of insufficient supporting information can largely be alleviated by supplementing the causal map with a means of referencing the contributing text segments that support a specific factor or linkage. In this way, the reader benefits from the high-level understanding of the connection and can seek more details if needed. This same conclusion was drawn by Kim & Andersen (2012) who subsequently argued for a data source reference table. The method developed in this project should therefore include a similar feature. The issue of counterfactual conclusions being missed is a more difficult shortcoming to address. Causal maps are not designed to capture such insights and so are ill-suited for this purpose. Finally, careful consideration must be given to what constitutes a 'causal relationship'. The definition used will heavily influence the type and number of connections present in a graph. Difficulty arises in the many fringe cases where it may not be entirely clear whether a causal link is argued, suggested, or implied. To ensure consistency and impartiality, analysts should endeavour to employ a constant definition.

Appendix B: ETS meta-reviews

Meta-review	Time period	Jurisdiction
(Green, 2021) Does carbon pricing reduce emissions? A review of ex-post analyses	<2021	Global
(Mascher, 2018) Striving for equivalency across the Alberta, British Columbia, Ontario and Québec carbon pricing systems: the Pan-Canadian carbon pricing benchmark	<2018	Alberta, British Columbia, Ontario and Québec
(Schmalensee & Stavins, 2017) Lessons learned from three decades of experience with cap and trade	1990-2020	US, EU-ETS
(Haites et al., 2018) Experience with carbon taxes and greenhouse gas emissions trading systems	<2015	Global
(R. Martin et al., 2012) An evidence review of the EU Emissions Trading System, focussing on effectiveness of the system in driving industrial abatement	<2012	EU-ETS
(Convery, 2020) Reflections—the emerging literature on emissions trading in Europe	<2009	EU-ETS
(B. Murray & Rivers, 2015) British Columbia's revenue-neutral carbon tax: A review of the latest "grand experiment" in environmental policy	2008-2015	British Columbia
(R. Martin et al., 2020) The impact of the European Union Emissions Trading Scheme on regulated firms: what is the evidence after ten years?	<2016	EU-ETS
(Laing et al., 2014) The effects and side-effects of the EU emissions trading scheme	<2014	EU-ETS
(Venmans, 2012) A literature-based multi-criteria evaluation of the EU ETS	2007-2012	EU-ETS

(Narassimhan et al., 2017) Carbon Pricing in Practice: A Review of the Evidence	<2017	Global
(Duan et al., 2014) Review of Carbon Emissions Trading Pilots in China	<2014	China
(Dong et al., 2016) From Pilot to the National Emissions Trading Scheme in China: International Practice and Domestic Experiences	<2016	China
(Ohlendorf et al., 2021) Distributional Impacts of Carbon Pricing: A Meta-Analysis	<2021	Global

Table 3 - ETS meta-reviews considered

Appendix C: Final ETS policy analysis source papers

Contributing review paper	Index	Individual papers
(Green, 2021)	1	(Anderson & Di Maria, 2011)
	2	(Gloaguen & Alberola, 2013)
	3	(Arimura & Abe, 2021)
	4	(Bayer & Aklin, 2020)
	5	(Bel & Joseph, 2015)
	6	(Cullenward, 2014)
	7	(Wagner et al., 2014)
	8	(Jaraite-Kažukauske & Di Maria, 2016)
	9	(Egenhofer et al., 2011)
	10	(Ellerman et al., 2016)
	11	(Kotnik et al., 2014)
	12	(Fell & Maniloff, 2018)
	13	(Dechezleprêtre et al., 2018)
	14	(Ellerman & Buchner, 2008)
	15	(G. Martin & Saikawa, 2017)
	16	(Ellerman & McGuinness, 2008)
	17	(B. C. Murray & Maniloff, 2015)
	18	(Petrick & Wagner, 2014)
	19	(Wakabayashi & Kimura, 2018)
(Schmalensee & stavins, 2017)	20	(Sijm et al., 2011)
	21	(Hibbard et al., 2015)
	22	(Wing & Kolodziej, 2008)
	23	(Ranson & Stavins, 2012)
	24	(Ellerman & Buchner, 2007)
	25	(Kruger et al., 2007)
	26	(Convery & Redmond, 2007)
	27	(Sartor et al., 2014)
	28	(Löfgren et al., 2015)

Table 4- Final ETS policy analysis source papers

Appendix D: Approaches to derive causal insights using NLP

Appendix C first provides some background on the nature of causal relations before going on to review the NLP literature on deriving causal insights and discussing the applicability of different methods for this purpose. This section then goes on to look more closely at one such method, relation extraction and outlines the specific approach chosen for this project's implementation.

What are causal relations?

Concepts of causality

Defining causality is a deceptively complex task. Despite intuitive notions of what constitutes causality, there is no unified theory or definition, scholars have presented numerous competing definitions, theories and counterexamples (Beebee et al., 2009; Copley & Wolff, 2014). Indeed, it is this complexity that contributes to the challenge of using NLP for causal relation extraction. If humans cannot agree on a definitive concept of causality, how can it be computationally operationalized? A comprehensive discussion on the field of causal relations is beyond the scope of this analysis, however, for this study we employ a high-level concept of causality, namely, manipulation. Manipulation posits that a relationship among some variables X and Y is causal if, were there a change in the value of X, the value of Y would also change (Drury et al., 2022; Psillos, 2007; Woodward, 2005). It is also worth noting at this stage, the concept of a counterfactual relation. A counterfactual causal relation implies a *lack* of causal relation between two variables.

Causality for NLP

When using NLP for causality-related purposes, some important terms are often used:

Explicit causality – This refers to causal relations in natural language text whereby variables are connected with explicit causal links, causative verbs, resultative phrases, conditions or causative adverbs and adjectives (Yang et al., 2021). This category is often easier to detect computationally.
Implicit causality – Unlike explicit causality, implicit causality involves more ambiguous connections. Readers must often use context, background knowledge and reasoning to determine the presence of causality in these cases. The extra level of comprehension required to detect implicit causality makes using NLP for implicit causality detection a much harder task.
Inter/Intra sentence causality – Inter-sentence causality refers to causal relations where both the cause-and-effect variables exist within the same sentence. Intra-sentence refers to when the variables are spread across multiple sentences. Intra-sentence causality is a significantly more challenging task given the increased comprehension required.
Embedded causality – This relates to segments of natural language wherein there exist multiple causal relations. Embedded causality occurs when a certain variable is included as both a cause and effect, in different causal relations.
In this section we have presented an introduction to the concept of causality, in the next section, we explore ways in which NLP has been used for deriving causal insights.

Overview of NLP for deriving causal insights

Having explored the academic literature relating to the extraction of causal relations, two distinct high-level approaches emerge: a top-down (co-occurrence) method which reduces a large body of text into core concepts and then finds connections between them and a bottom-up method (relation extraction) which first identifies connections and then aggregates them.

Co-occurrence based methods

This top-down category consists of methods that determine relations between concepts based on their co-occurrence in the source material. The most frequent application of co-occurrence-based methods regarding causal relations involves the combination of latent semantic analysis (LSA) and fuzzy association rule mining (FARM). This approach consists of three key stages which together produce a fuzzy cognitive map of concepts within a corpus and the causal weights between them. First, key concepts are extracted using an LSA topic modelling approach, second, FARM is used to determine the relations between the terms comprising the concepts, and finally, the concepts and their relational links can be constructed into a fuzzy cognitive map (FCM) (Han et al., 2019; J. Kim et al., 2016; Son et al., 2020).

So far this method has been used primarily in the field of scenario planning (Amer et al., 2013), which seeks to establish ‘‘a set of hypothetical events set in the future [which are] constructed to clarify a possible chain of causal events as well as their decision points’’ (Kahn & Wiener, 1967). Kim et al (2016) used this approach on a future-oriented corpus on the development of electric vehicles. The FCM was analysed both in a static sense to identify important scenario concepts, and in a dynamic sense to explore what-if scenario experiments. Similarly, Son et al., (2020) developed scenarios for the future of unmanned aerial vehicle technology by selecting high centrality concepts from the FCM and observing their expected impact on other development factors. While not exactly in line with the ambitions of *this* project, these studies indicate that a topic modelling and association rule approach can indeed identify concepts and relations. Regarding the application of this method for policy analysis, however, there is very little evidence. Han et al., (2019) provide the only example, in which they analyse a corpus of 828 documents from government and news website sources to construct an FCM that represents the system of state-owned capital layout related policies. The map was used to

deduct the evolution of concepts within the system, whilst avoiding the subjective bias of modellers. While the authors conclude that their method can provide "objective and neutral conclusions [for policy analysis]", it is important to recognise that their insights were derived from many assumptions and domain specific knowledge. What this example also shows is that this NLP method can effectively synthesise insights from a prohibitively large corpus. Analysing the 828 documents manually would be prohibitively time-consuming.

Whilst the limited number of examples in literature do advocate for the value of this approach in certain contexts, following consideration and testing, it has been deemed unsuitable for the purposes of this study because:

- Given that LSA seeks to reduce the dimensionality of provided information, it is ill-suited to extract granular concepts from within large volumes of text (Landauer et al., 1998). Only terms that frequently co-occur in similar pieces of text are likely to be grouped within a concept, a factor that may be presented as significant in the source material would be missed or incorrectly grouped with other factors if it does not meet this condition. Whilst LSA might reflect the high-level conclusions from a paper (i.e., that ETS has an impact on GHG emissions) if it cannot capture the lower level contributing relations, then the resulting map will only provide limited value for policy development.
- The justification for a relationship between concepts is based entirely on the association of terms in the source material, however, co-occurrence alone is not sufficient to assert a relationship (Fleuren & Alkema, 2015). While causality implies association, the reverse is not a given. For example, if a key conclusion from a paper was that an ETS had *no* impact on emissions, then the concepts encompassing ETS and emissions would still likely be closely associated and so have a connection in the FCM. This may then lead to the incorrect interpretation that ETS did in fact impact emissions.
- Finally, the degree of abstraction provided by LSA blurs the connection between output concepts and the source material. Concepts are built from isolated terms and there is no indication of which text segments they were sourced from. The same criticism can be made of the relations as the specific instances of co-occurrence are not presented. As a result, it can be difficult for an analyst to gain deeper insight and confidence in the FCM representation.

Relation extraction

Relation extraction is a bottom-up category of approaches that seek to understand the semantic relations between textual elements in a text segment (Bach & Badaskar, 2007). Causal relation extraction is the sub-field looking solely at *causal* relations between textual elements (Khoo & Na, 2006). As we will discuss in section 0, there are many methods to achieve this. At the most basic level, an algorithm will determine whether it thinks a sentence is causal or not which then requires manual coding to extract cause and effect factors. More complex projects examine larger text segments and aim to determine the cause-and-effect factors automatically. Once relations have been found, a causal map can be constructed using each individual link. Regardless of the specific type of approach, relation extraction provides two key benefits over the LSA + FARM method:

- Firstly, this method provides much greater coverage of factors by virtue of using a bottom-up approach. The inclusion of a factor is dependent only on the algorithm detecting causality, regardless of how many times that factor is mentioned in the document. In this way, even the most granular of factors can be included.
- Given that factors and relations are found directly from text segments, there is always a clear link to the source material. As a result, additional context can easily be gained by an analyst. This is useful when understanding requires consideration of context. It is also helpful for building confidence in the validity of the final causal map.

While by no means perfect, these benefits and the drawbacks of LSA + FARM mean that causal relation extraction should be considered a more suitable approach for the purposes of this study.

Methods to derive causal relations

Unlike the co-occurrence approach, the field of causal relation extraction is well developed with numerous studies presenting diverse algorithms and applications. Amongst literature reviews of causal relation extraction, categorical distinctions have been suggested according to the specific techniques presented in different works. Barik et al. (2016) define categories of 'methods using manual patterns', 'methods using semi-automated patterns', 'supervised learning methods' and finally 'statistical methods'. Asghar (2016) instead makes the distinction between statistical techniques – those methods leveraging implicit patterns and machine learning, and non-statistical techniques – those leveraging semantic and syntactic linguistic pattern matching, regardless of the degree of automation. What these categorisations both miss however is the new and burgeoning field of deep learning methods. The categorisations used in Yang et al., (2021) and Drury et al., (2022) both include a separate grouping for these methods. While Drury et al. locates deep learning under the broader category of machine learning, the tripartite grouping of knowledge-based, statistical ML and deep learning categories in Yang et al., (2021) allows easier comparisons between these techniques which is useful for our context. In the rest of section 2.3, we will describe methods in each of these categories and their applicability for the purposes of this project.

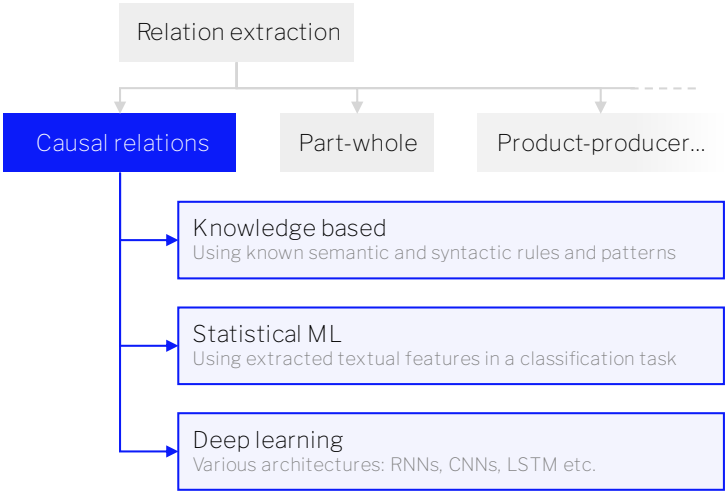

Figure 12 - Causal relation extraction techniques, adapted from (Yang et al., 2021)

Knowledge-based

These techniques rely on known semantic and/or syntactic features which are codified in rules or patterns. This can include for example identifying when specific causative verbs are followed by a noun, or if there are certain sequences of named entities, i.e. <mutation, relation, drug> triplets (Bui et al., 2010). Text segments that exhibit these features will be returned as output.

If a knowledge-based technique is to be used on consistent textual data with well-understood semantic and syntactic structure, then this approach can achieve good results, both in terms of precision and recall, all with relatively low computational cost (Beamer et al., 2008; Girju et al., 2009). As the complexity of the dataset increases, however, the task of encompassing the enormous range of causality-related features manually, quickly becomes unfeasible (Yang et al., 2021). This issue is exacerbated when a task requires detection of implicit causality, in which case a more ambiguous set of textual features must be considered. Given that in this project the source material uses complex natural language, that the policy analysis source material, in general, is diverse and complex, and that implicit causality features heavily, this limitation is severe.

Statistical ML

Unlike knowledge-based techniques, statistical ML identifies relevant features automatically from labelled data and then uses machine learning algorithms to perform the classification task based on these features. A common approach includes first finding candidate causal text segments based on sentence features and then pruning this list based on decision trees or Bayesian inference.

Statistical ML has been shown to achieve very good precision and recall results, often only marginally behind leading deep learning methods, and it can achieve this with significantly fewer computational resources (Airola et al., 2008; Pakray & Gelbukh, 2014; Zhao et al., 2016). The primary drawback of this technique is its limited portability. The extracted textual features and subsequent classification mean that the trained models often perform poorly when given data from dissimilar sources (Asghar, 2016). This issue is compounded by the fact that there are few comprehensive labelled causal training datasets (Yang et al., 2021). The ambition of this project is to present a method that can be applied to diverse policy analysis domains, and as such portability is desired.

Deep learning

Deep learning techniques utilise neural networks of various architectures, most commonly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) but also hybrid systems which combine multiple elements. More recently, transfer learning involving large pre-trained language models such as BERT (Devlin et al., 2018) has been shown to significantly improve performance (Beltagy et al., 2019). Once suitably trained, these methods return text segments deemed to be causal based on their learnt representation.

Against almost every benchmark dataset, deep learning algorithms have been shown to achieve the best results, often outperforming the F-score of the next best technique by 5-10% (Yang et al., 2021). Indeed, deep learning has been the technique of choice for the majority of recent work in the causal relation extraction field. Besides the superior performance, another key advantage that deep learning provides over the other approaches is its improved portability. This is a result of its more complex architecture which can deduce higher-level information from input textual data (Yang et al., 2021). Nevertheless, this approach does still suffer from portability issues when provided with poor or insufficient data, unlike statistical ML, however, deep learning can leverage transfer learning to help alleviate these problems to an extent (Kyriakakis et al., 2019; Li et al., 2021; Wu et al., 2019).

Applicability for this project

With consideration of the relative advantages and disadvantages of each causal relation extraction technique, a deep learning technique appears to be the most suitable primarily because of its high benchmark performance, and its relatively good portability. This justification is in line with the recommendations from Yang et al. (2021) who indicate deep learning as the most appropriate choice in the case of inconsistent data, and when good portability is required.

SCITE

To select the specific deep learning causal relation extraction algorithm to be employed in this project, a review of methods was conducted which centred around three criteria:

- The algorithm must be open source. Developing an implementation from scratch or documentation alone would not be feasible for this project, as such the search was limited to those projects with provided code of reasonable quality. This proved to be a significant constraint as many projects only included an academic article, others included sparsely documented or unmaintained code.
- Should be focused on *causal* relations. Many projects sought to identify alternative types of relations, alongside causal. In these cases, the methods were evaluated against their performance across all such relations. This often meant the purely causal performance was unclear.
- Attainment of reasonable benchmark performance. As stated earlier, there are a few benchmark datasets against which these algorithms can be tested. Although the focus of this project is not to achieve leading precision or recall, good results will provide better inputs for subsequent stages. We, therefore, used this criterion to inform our selection amongst those algorithms that satisfy the earlier criteria.

Additionally, a few features were deemed desirable in the selection of an algorithm: Firstly, methods that output both causality detection and cause-effect factor extraction. This feature would increase the automation of the process. Secondly, algorithms that could be used 'out of the box' were desirable. Leading methods almost all leverage exceedingly large computational resources. A requirement to retrain these methods would be costly and time-consuming.

Searching through Google Scholar, GitHub, Papers with code, and Hugging face, the method selected was:

SCITE: **S**elf-attentive BiLSTM-**C**RF w**I**th **T**ransferred **E**mbeddings

presented by Li et al., (2021). The project is open-source with well-documented code, it has an explicit focus on causal relations and achieves an F score of 0.85 against the SemEval 2010 task 8 dataset which is only ~0.05 below the current leading method. The algorithm extracts cause-effect factors alongside causality detection and can be repurposed for this project with relatively little adaptation required.

Appendix E: SCITE output, recall and precision metrics

After selection of the relevant sections and data cleaning, as described in step 2, 4542 sentences were included as input to the SCITE algorithm. Of these, 317 were deemed causal by SCITE. It took roughly 20 minutes for a typical paper of ~150 sentences to go from raw text to final outputs, with the majority of time being used to compile the contextualised word embeddings (a typical paper would take ~2 hours to review manually). Figure 13 shows a typical output. Each output includes the sentence number, sentence text and a list of tuples of suggested causal pairs. In some cases, a sentence can contain multiple causal pairs, as in Figure 14, in which case SCITE returns a list of multiple causal tuples.

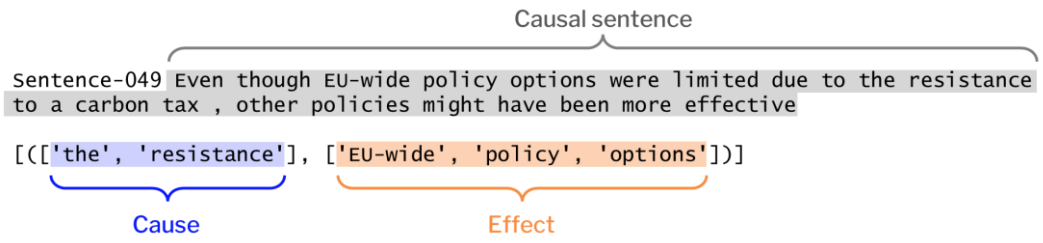

Figure 13 – Typical SCITE output, paper 4 sentence 49

```
Sentence-251 On the other hand , when assuming that the power companies in The
Netherlands receive 90 of their needed emission allowances for free , the
grandfathering effect far outweights the price effect , resulting in major total
windfall profits due to emissions trading based largely on free allocation

[(['the', 'grandfathering', 'effect'], ['major', 'total', 'windfall', 'profits']),
(['emissions', 'trading'], ['major', 'total', 'windfall', 'profits'])]
```

Figure 14 - Multiple extracted causal pairs, paper 20 sentence 251

What may already be apparent from Figure 13 is that these outputs are not perfect, in this case, the cause should really be 'the resistance *to carbon tax'*. To explore the variation in output accuracy, Figure 15 shows five types of typical output. First, we see a good result in which the sentence is causal, and the suggested causal pairs are accurate. Sentence 108 is causal however the suggested pairs require some interpretation, alone they do not convey the relation very well. Sentence 26 is a more extreme case where the sentence is causal and the suggested pairs at first seem reasonable, however, inspecting the sentence reveals that the second pair is incorrect. Instead of 'the economic downturn' causing 'higher emissions' it should be 'emissions *reduction'*. Sentence 64 is a common case whereby the text is causal, but not relevant to the policy analysis. Despite filtering out methodological sections of the paper, methodological/sensitivity analysis-related sentences are still present amongst other irrelevant sentences. Finally, sentence 142 is an example of the occasional output which is not causal at all.

```
------------------
Sentence-007 Price substitution effects induced by coal and gas prices also seem to have affected
emissions
[(['coal', 'and', 'gas', 'prices'], ['Price', 'substitution', 'effects'])]
```
Good result

```
------------------
Sentence-108 7 shows that an increase in the coal-gas price ratio from 04 to 1 would increase NGCC
capacity factors
[(['a', 'smaller', 'price', 'ratio'], ['a', 'smaller', 'increase'])]
```
Requires some interpretation

```
------------------
Sentence-026 This means that higher GDP growth rates lead to higher emissions and , vice versa , an
economic downturn comes in hand with an emission reduction which leads us to conclude that the
economic downturn among the EU-25 was a main driver of the reduction in emission rates
[(['higher', 'GDP', 'growth', 'rates'], ['higher', 'emissions']), (['the', 'economic', 'downturn'],
['higher', 'emissions'])]
```
Misleading cause-effects

```
------------------
Sentence-064 Furthermore , robust standard errors have been used to correct for any problems that
may cause heteroskedasticity
[(['any', 'problems'], ['heteroskedasticity'])]
```
Irrelevant

```
------------------
Sentence-142 What makes sense from the perspectives of developing in-state resources to meet state
electricity demand
[(['the', 'perspectives'], ['sense'])]
```
Not causal

Figure 15 – Types of SCITE output, from papers 2, 12, 5, 5, 21 respectively

The vast majority of outputs fell, to a greater or lesser degree, into the category of 'requiring some interpretation'. Good results and irrelevant results occurred with roughly equal frequency and misleading and non-causal sentences were less frequent. What this demonstrates is that the raw outputs alone are not of sufficient quality to be used directly in subsequent stages of analysis. Additionally, the algorithm outputs do not provide any explicit indication of the direction of linkage. This feature is very important when it comes to the final aggregation of relations into the causal map. A positive direction indicates that increasing the cause factor would result in an increase in the effect factor, conversely, a negative relationship indicates that increasing the cause factor would decrease the effect factor. For example, consider the causal relation from 'Industrial activity' → 'Energy demand', this would have a positive relationship given that increasing activity would most likely induce greater energy demand. On the other hand, the relation 'Energy efficiency' → 'Energy demand' would be negative, increasing efficiency would decrease energy demand.

Given the high level of inaccuracies, and the lack of relation direction, it was deemed necessary to manually review each output sentence to determine the true causal relations, causal pairs and direction. From this review, 154 sentences were verified as causal *and relevant*, which yielded 284 causal pairs. Table 55 shows the results of this review, where 'Total input sentences' is the number of sentences after filtering by relevant sections and cleaning, 'SCITE output sentences' is the raw output from SCITE, 'Verified sentences' is the raw outputs which are causal and relevant, and 'Causal pairs' is the number of pairs yielded from the relevant sentences.

Paper number	Total input sentences	SCITE output sentences	Verified sentences	Causal pairs
1	83	13	6	8
2	178	21	13	24
3	88	6	4	6
4	92	8	6	11

5	98	10	4	5
6	69	5	4	5
7	109	4	2	4
8	123	23	10	14
9	521	25	16	27
10	115	1	0	0
11	93	9	3	9
12	139	9	5	6
13	367	33	8	31
14	249	13	3	3
15	99	4	0	0
16	68	6	6	12
17	102	3	2	5
18	166	7	4	6
19	98	16	7	14
20	272	28	17	34
21	210	15	6	10
22	80	8	4	9
23	208	13	10	19
24	252	18	6	9
25	264	4	1	2
26	209	3	1	1
27	150	9	3	6
28	40	3	3	4
Total	**4542**	**317**	**154**	**284**

Table 5 - Overview of causality extraction results

What remains unclear from this review is an understanding of how many true causal sentences were missed by the SCITE output. This important value is formalised in the recall metric which is defined as the number of true positives over true positives and false negatives. In our case, we consider recall as the number of true causal sentences returned by SCITE (both relevant and irrelevant) divided by the total number of causal sentences from the input data. We include irrelevant causal sentences in this equation because SCITE is only meant to determine causality, it has no functionality to determine relevance. To calculate this value, a random subset of 4 papers (2, 5, 19 and 23) were manually reviewed, in their entirety, to extract all causal sentences. The results are shown in Equation 1.

$$Recall = \frac{True\ positives}{True\ positves + False\ negatives} = \frac{True\ causal\ sentences\ from\ output}{Total\ causal\ sentences\ from\ input}$$

	Paper 2	Paper 5	Paper 19	Paper 23	**Average**
$Recall =$	$\frac{17}{37} = 0.46$	$\frac{9}{30} = 0.30$	$\frac{9}{25} = 0.36$	$\frac{13}{32} = 0.41$	**0.38**

Equation 1 - Algorithm recall formula and results

The average recall across these papers is just 38% which is significantly lower than that achieved by SCITE when evaluated on its testing data. To determine some possible causes for this poor result, some common types of sentences missed by SCITE have been highlighted in Figure 16. We will discuss some possible reasons for these errors:

- *Sentence complexity* - As discussed in section 0, implicit causality is a more nuanced type of causality requiring consideration of context. This type of causality represents a step up in terms of complexity, so it is perhaps not surprising that these types of sentences are harder to detect. In example 1, the causal subject is implied, referred to only as “this”, additionally, the causality indicating features of the sentence are quite ambiguous, “motivated” is not commonly used in a causative capacity. As shown in example 2, sentences with embedded causality were often missed by SCITE. In this case, the sentence first indicates the

effects of linkage, then how these effects in turn impact control over further impacts. Complexity is again a likely cause of their exclusion. Another likely cause for the poor recall of these more 'complex' sentences, is the type of training data used for SCITE. The SemEval 2010 task 8 dataset documentation provides no elaboration on the nuances in causality relations, only stating that a cause-effect relation is when "An event or object yields an effect", and giving the simplistic example "those cancers were caused by radiation exposures" (Hendrickx et al., 2019, p. 1). Reviewing the dataset reveals that the overwhelming majority of sentences are similarly straightforward and explicit. Indeed this issue was noted by the authors of SCITE who, in their error analysis, stated that the low frequency of embedded causality sentences led to very poor recall for these sentences (Li et al., 2021, p. 214). Interestingly, the recall for embedded sentences recorded by the authors was in the range of ~20-30% and so much closer to the results experienced in this study.

- *Ambiguous language* - Issues of ambiguous language are also frequent in the excluded causal sentences. In cases like example 3, it is often not clear if the authors of the original text assert a causal relationship or not. Here it is suggested that linkage *may* cause certain effects, similarly, vague terms include 'suggest' or 'is reflected by'. In these cases, the ambiguity blurs the lines between causal and non-causal. Automated classification tasks are therefore difficult. The issue of data sparsity likely also contributes here, as the training data did not include these ambiguous instances.
- Finally, there are cases where it is unclear why SCITE did not detect causality. Even in cases of fairly simple, explicit causality, valid sentences were sometimes missed.

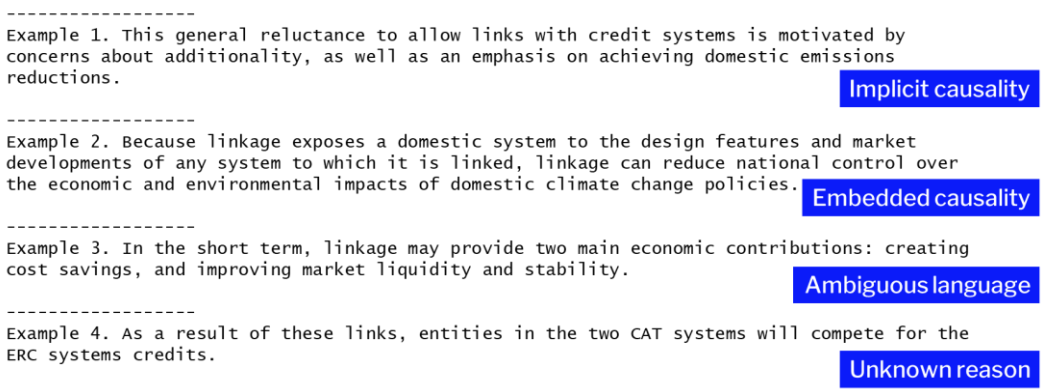

Figure 16 - Types of false-negative from SCITE

Another interesting metric is precision which gives an indication of the quality of outputs. It is defined as the total number of true positives divided by true positives and false negatives. In our case we again consider both relevant and irrelevant sentences, using those that are verified as causal for the numerator and the total number of output sentences for the denominator. Using the previously verified sentences, we calculate the precision over the entire dataset in Equation 2. 84% is a good result and indicates that the algorithm outputs can be trusted to be causal with reasonably high confidence. This result is very close to the 83% achieved by the authors (Li et al., 2021, p. 213).

$$Precision = \frac{True\ positives}{True\ positves + False\ positives} = \frac{True\ causal\ sentences\ from\ output}{Total\ sentences\ from\ output}$$

$$Precision = \frac{265}{317} = \mathbf{0.84}$$

Equation 2 - Algorithm precision formula and results

From the precision and recall values, it is possible to calculate the F-score. This is an often-used metric that conveys the balance between precision and recall. The F-score is defined as shown in Equation 3, the SCITE algorithm applied to our dataset gives an F-score of 0.52

$$F\ score = 2 * \frac{Precision * Recall}{Precision + Recall} = \mathbf{0.52}$$

Equation 3 - Algorithm F score formula and results

Appendix F: Final causal map reference table

Link ID	Relation IDs	Link ID	Relation IDs
001	23-167A	**080**	11-109E
002	23-015	**081**	28-025A, 28-025B
003	23-167C	**082**	27-022
004	23-167B	**083**	14-101
005	23-114, 23-067	**085**	9-126, 19-088
006	23-110A	**086**	2-075C
007	2-159	**087**	4-079B, 7-033B
008	23-110C	**088**	2-067B
009	23-049	**089**	4-028, 11-019E, 11-019F, 17-083C, 24-145
010	4-067B	**090**	4-079C
011	5-026A	**091**	2-075A, 5-091
012	11-019D	**093**	24-195
013	9-350	**094**	9-317
014	4-065, 6-002, 6-059, 6-067	**095**	24-224A
015	9-075D, 9-404D	**096**	23-130A
016	27-043A	**097**	23-134
017	6-017A	**098**	23-133A
018	22-019B	**099**	24-224B
020	12-126	**100**	23-130B
021	17-042B, 22-003B	**101**	23-133B
022	13-343A	**102**	23-133C
023	4-067A, 6-017B	**103**	20-263A
024	17-042A	**104**	20-265B
025	12-115, 12-133, 17-042B, 22-003B, 22-003A	**105**	24-087
026	12-133, 17-042B, 22-003B	**106**	24-087
027	11-017A	**107**	25-026A
028	13-066, 28-030	**108**	25-026B
029	8-065	**109**	13-343B
030	21-056A	**110**	20-052, 20-251A, 20-212A
031	8-005	**111**	13-036
033	4-067C, 7-097, 21-022	**112**	24-067A
034	1-068, 2-167, 3-064A, 4-047A, 9-101, 11-019B, 13-034, 18-003, 19-006, 19-076	**113**	9-059B, 20-074, 21-043, 22-019A, 22-056A, 24-067C
035	11-017B	**114**	26-199
036	21-033C	**115**	1-029, 9-271, 13-336D, 20-044C, 20-052, 20-251A, 20-212A, 20-256
037	21-033B	**116**	13-339B
038	2-028A	**117**	13-031A, 13-277, 20-239, 20-256
039	2-005A, 2-005C, 2-028C	**118**	13-031B
040	2-005A, 2-005C	**119**	13-031C
041	19-064C	**120**	13-031D
042	2-005A, 2-005B, 2-028B	**121**	13-031L, 13-336H
043	9-154	**122**	13-031K, 13-336G
044	19-008A, 19-040B, 19-064B	**123**	13-031J, 13-336F
045	2-005A, 2-171, 19-008B, 19-064C, 19-064D, 21-056B	**124**	1-029, 8-069B, 8-075B, 8-080, 8-076A, 9-271, 13-336D, 20-044C, 20-052, 20-251A, 20-251B, 20-212A, 20-212B, 20-256, 28-004
046	19-078A	**125**	9-075A, 9-256, 9-404A, 9-282B, 9-059A, 13-031G, 13-336C, 13-339D, 13-339E, 20-260, 20-205
047	19-064A	**126**	21-056C
048	9-497A	**127**	8-121
049	9-497B	**128**	20-044A
050	8-118A	**129**	20-044B
051	8-069A, 8-118B	**130**	20-219
052	8-068	**131**	20-186A
053	18-094	**132**	9-511
054	2-023, 2-129, 4-047B, 5-067, 5-026B, 5-079, 17-083A	**133**	9-510
055	2-005A, 2-029	**134**	9-059B, 13-339A, 20-074, 21-043, 22-056A, 24-067C

056	2-023, 2-046, 2-129, 4-047B, 5-067, 5-026B, 5-079, 17-083A	**135**	13-339C, 27-043B
057	20-133C, 20-139B	**137**	9-075A, 9-404A, 9-282A, 9-059A, 13-031G, 13-336C, 20-212B, 20-074, 21-043, 22-056A, 24-067C
058	1-035, 2-005A, 2-023, 2-129, 3-050, 3-064B, 3-078A, 4-047B, 5-067, 5-026B, 5-079, 12-085B, 17-083A, 17-083B, 19-008B, 19-064C, 19-064D, 19-078B, 22-056C	**138**	21-056C
059	1-018, 1-035, 2-005A, 2-023, 2-067A, 2-129, 3-050, 3-064B, 3-078A, 4-047B, 5-026B, 12-085B, 5-079, 17-083A, 17-083B, 19-008B, 19-064C, 19-064D, 19-078B, 22-056C	**139**	20-042B
060	2-005A, 2-067A, 2-075B	**140**	27-063B
061	2-031D, 11-019C	**141**	20-042A
062	12-085A	**142**	27-063A
063	2-007B, 2-031C, 14-211, 16-050, 16-039B	**143**	9-075B, 9-404B
064	9-163C	**144**	20-026A, 20-139D
065	22-024A	**145**	20-139A
066	2-031A, 16-004A, 16-039A, 16-063A	**146**	8-075A, 9-073
067	9-163B, 16-004B, 16-004C, 16-063B, 16-063C	**147**	13-031E, 13-336A, 20-074, 21-043, 22-056A, 24-067C
068	16-004C, 16-021B, 16-063C, 20-133B	**148**	9-075C, 9-404C
069	12-108, 16-004B, 16-021A, 16-063B	**149**	4-049
070	2-007A, 9-163B,	**150**	1-035, 3-050, 3-064B, 3-078A, 16-016, 12-085B, 17-083B, 22-056B
071	2-031B, 7-033C	**151**	8-077
072	2-075B, 14-063	**152**	13-031F, 13-336B
073	7-033A	**153**	20-026B, 20-026C
074	11-053	**154**	20-042C
075	1-013	**155**	20-044D
076	2-067B	**156**	20-044E
077	17-083C	**157**	20-139C
078	4-028	**158**	21-033A
079	11-019F	**159**	23-017A, 23-017B

Table 6 - Final causal map reference table

Appendix G: Network graph representation of the causal map

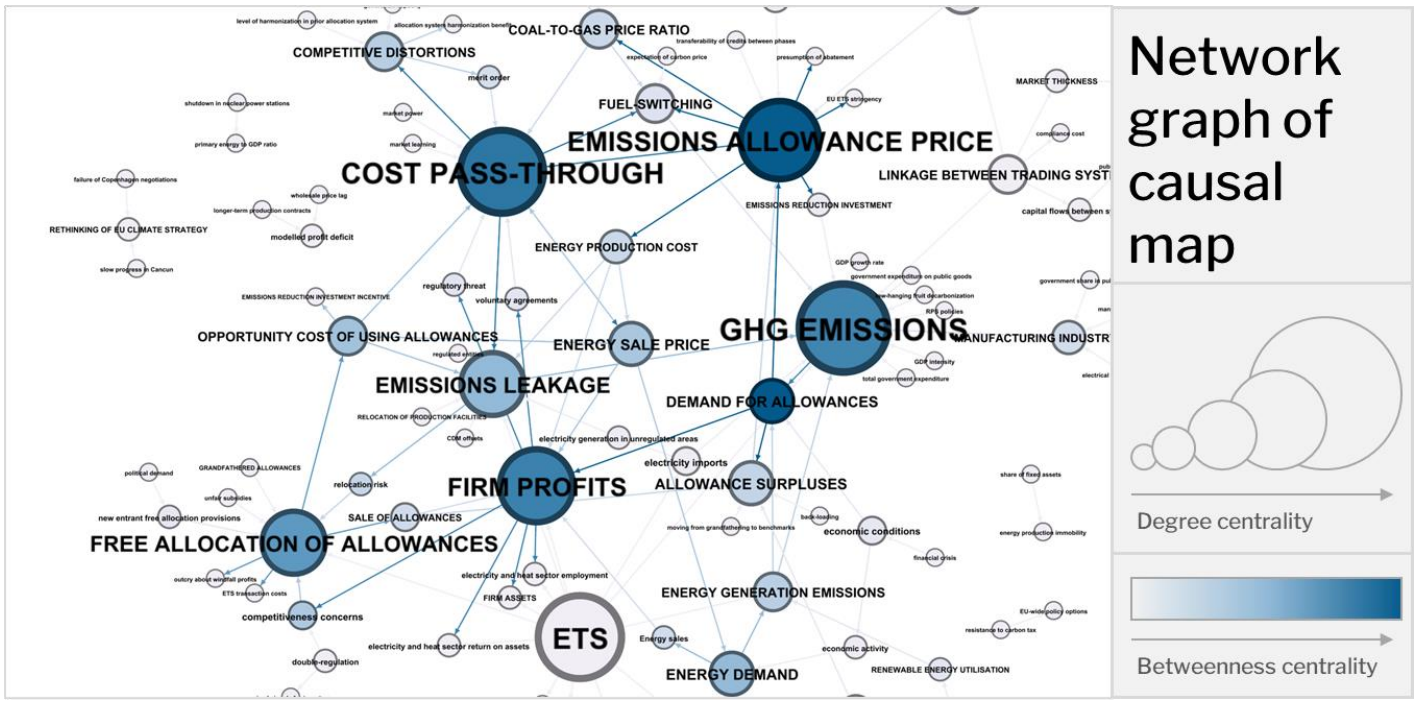

Figure 17 - Network graph representation of the causal map

7. Appendix H: EPA relevance

This topic is related to the Engineering and Policy (EPA) program as a result of its application to a grand challenge domain (global warming/energy transition policy domain) as well as its adoption of analytical and modelling methods. At its heart, this project seeks to analyse and bring clarity to a complex policy area (emissions trading schemes) by combining disparate information sources with the aim of contributing to more effective future policy development. To achieve this aim, several conceptual modelling and natural language processing methods have been employed (Using SCITE, semantic clustering and structuring the results into a causal map). In this sense the project satisfies the grand challenges, socio-political skill and analytics and modelling core competencies of the EPA program.