A semi-automated approach to policy-relevant evidence synthesis

combining natural language processing, causal mapping, and graph analytics for public policy

Hooper, Rory; Goyal, Nihit; Blok, Kornelis; Scholten, Lisa

DOI
10.1007/s11077-024-09548-3

Publication date
2024

Document Version
Final published version

Published in
Policy Sciences

Policy Sciences
https://doi.org/10.1007/s11077-024-09548-3

RESEARCH ARTICLE

A semi-automated approach to policy-relevant evidence synthesis: combining natural language processing, causal mapping, and graph analytics for public policy

Rory Hooper[1] · Nihit Goyal[1] · Kornelis Blok[1] · Lisa Scholten[1]

Accepted: 27 August 2024

Abstract
Although causal evidence synthesis is critical for the policy sciences—whether it be analysis *for* policy or analysis *of* policy—its repeatable, systematic, and transparent execution remains challenging due to the growing volume, variety, and velocity of policy-relevant evidence generation as well as the complex web of relationships within which policies are usually situated. To address these shortcomings, we develop a novel, semi-automated approach to synthesizing causal evidence from policy-relevant documents. Specifically, we propose the use of natural language processing (NLP) for the extraction of causal evidence and subsequent homogenization of the text; causal mapping for the collation, visualization, and summarization of complex interdependencies within the policy system; and graph analytics for further investigation of the structure and dynamics of the causal map. We illustrate this approach by applying it to a collection of 28 articles on the emissions trading scheme (ETS), a policy instrument of increasing importance for climate change mitigation. In all, we find 300 variables and 284 cause-effect pairs in our input dataset (consisting of 4524 sentences), which are reduced to 70 unique variables and 119 cause-effect pairs after homogenization. We create a causal map depicting these relationships and analyze it to demonstrate the perspectives and policy-relevant insights that can be obtained. We compare these with select manually conducted, previous meta-reviews of the policy instrument, and find them to be not only broadly consistent but also complementary. We conclude that, despite remaining limitations, this approach can help synthesize causal evidence for policy analysis, policy making, and policy research.

Keywords Evidence-informed policy · Emissions trading schemes (ETS) · Causal mapping · Machine learning (ML) · Natural language processing (NLP) · Policy analysis

✉ Nihit Goyal
nihit.goyal@tudelft.nl

[1] Faculty of Technology, Policy and Management, Delft University of Technology, TU Delft, The Netherlands

Published online: 20 September 2024

Introduction

Transparent, timely, repeatable, and contextual synthesis of causal evidence is important for analyzing and informing every stage of the policy cycle. Policy research relies on effective summarization, synthesis, and mobilization of knowledge *for* or *in* the policy process. Studies that aim to facilitate evidence-based or evidence-informed policy-making, policy evaluation, and policy learning require at least some degree of evidence synthesis (Sanderson, 2002). Relatedly, studies conducting ex-ante policy assessment benefit from a careful aggregation of existing research to map a complex system (Freebairn et al., 2016). Even reviews of policy research – such as the Policy Studies Yearbook or Theories of the Policy Process – involve critical appraisal of evidence (Norman, 2023; Weible & Sabatier, 2018). Methods for evidence synthesis have, therefore, received much attention from policy makers, analysts, and researchers alike.

However, several characteristics of the policy sciences make evidence synthesis challenging. First, the exponential growth in policy-relevant research increases resource and time requirements even as time-critical policy advice is more in demand (Bornmann & Mutz, 2015; Larsen & Von Ins, 2010; Nunez-Mir et al., 2016). Second, evidence is scattered across different policy areas, in academic and grey literature, wherein different terminologies are used to refer to similar phenomena (Goyal & Howlett, 2018; Saetren, 2005). Third, policies have (un)intended consequences on a variety of dimensions – such as programmatic, process, and political (Marsh & McConnell, 2010; Wakabayashi & Kimura, 2018) – all of which should be considered in evidence synthesis. Fourth, the process is complicated further due to different levels of policy analysis, spanning from cross-sectional, micro-level studies of individual policies to time-series, macro-level research on entire policy areas (Esty & Porter, 2005; Jiali, 1995; Thow et al., 2010; Warner & van Buuren, 2011).

The field of evidence synthesis has witnessed rapid development and now consists of several methods that facilitate systematic source identification, repeatable evidence appraisal, and transparent reporting. On the one hand, methods such as thematic synthesis (Thomas & Harden, 2008), framework synthesis (Carroll et al., 2011), scoping reviews (Peters et al., 2020), meta-narrative reviews (Wong et al., 2013a), and meta-ethnography (France et al., 2019) enable broad syntheses and critical appraisals of research, but do not seek to inform policy-making per se. On the other hand, methods such as systematic reviews (Pearson et al., 2015; Petticrew & Roberts, 2008), meta-analyses (Barza et al., 2009), and umbrella reviews (Fusar-Poli & Radua, 2018) help synthesize evidence on interventions in a critical, repeatable, and transparent manner. They can, however, be difficult to execute, limited in their ability to acknowledge diversity in evidence, simplistic in their treatment of complexity, and time-consuming (Haddaway et al., 2017). Although living systematic reviews (Millard et al., 2019) and rapid reviews permit execution in a time-sensitive situation (O'Leary et al., 2017), they do not address the other shortcomings.

A more configurative approach is necessary to capture the underlying system complexity (Anderson et al., 2013). Realist syntheses, which focus on mechanisms of how interventions work (Macura et al., 2019; Wong et al., 2013b), address complexity better (Greenhalgh et al., 2011). These can shed light on the underlying reasons behind success and facilitate theory-building or theory-testing. As they can also require significant resources, rapid realist reviews have been proposed (Saul et al., 2013). Meanwhile, meta-aggregation collates 'lines of action' in available evidence that can inform decision-making in an auditable, reliable, and transparent manner (Hannes & Lockwood, 2011). However, these methods, too, provide limited support in addressing diverse conceptual lenses of analyses in policy-relevant evidence.

This study presents a semi-automated approach to extract, aggregate, map, and analyze causal evidence from policy-relevant literature. To address the challenges outlined above, we combine natural language processing (NLP), causal mapping, and graph analytics. NLP refers to computational techniques for analyzing and representing natural language text (Liddy, 2001), and is increasingly used in the policy sciences (El-Taliawi et al., 2021; Goyal & Howlett, 2019, 2021). Meanwhile, causal mapping facilitates the organization and representation of causal evidence regarding interdependencies among parts of a system (Barbrook-Johnson & Penn, 2022), but has witnessed limited use in policy analysis. Finally, graph analytics – a field that is still underutilized for policy research – can help with systematic investigation of a causal map (Nguyen et al., 2013), a type of graph. While we use specific alternatives within each field for illustration, our approach itself is agnostic to the specific techniques chosen within NLP, causal mapping, and graph analytics.

This article is structured as follows. In Sect. "The potential of semi-automated causal evidence synthesis", we present an approach for semi-automated evidence synthesis. Subsequently, we illustrate the approach by synthesizing evidence on the emissions trading scheme (ETS) to create a causal map (Sect. "Synthesizing policy-relevant evidence: An illustration using the emissions trading scheme (ETS)"). In Sect. "Synthesizing policy-relevant evidence: An illustration using the emissions trading scheme (ETS)", we describe and analyze the causal map to demonstrate its suitability for generating policy-relevant insights. Thereafter, we discuss the implications of this research (Sect. "Discussion") and conclude the article (Sect. "Conclusion").

The potential of semi-automated causal evidence synthesis

Causal evidence synthesis is a multi-step process. The first step is the collation of documents that contain evidence pertaining to a research area, topic, question, or hypothesis. Existing machine learning techniques such as classification and clustering can help in this step (van de Schoot et al., 2021). We, however, focus on the remainder of the process.

Typically, the second step is the extraction and appraisal of evidence. While some protocols appraise causality based on research design – e.g., including only randomized control trials or meta-analyses of these (Burns et al., 2011) – such approaches exclude relevant evidence created through various qualitative, quantitative, and mixed-methods designs. An alternate approach is to extract causal claims from the relevant documents. Such claims might be present in the form of explicit causality (e.g. using causative adjectives, adverbs, or verbs), implicit causality (based on background knowledge and the line of reasoning), inter-sentence causality (cause and the effect are spread across multiple sentences), and embedded causality (a variable is included as a composite effect of one cause and the cause of another effect) (Yang et al., 2021). Due to the complexity of natural language and the growing volume of policy-relevant evidence, this is challenging even for a subject matter expert. While NLP is unlikely to substitute human intelligence, advances in the field can assist in reducing and redirecting human effort.

The approaches to causality detection can be broadly classified into (top-down) co-occurrence-based methods and (bottom-up) causal relation extraction methods. Co-occurrence-based methods reduce a large volume of text into core concepts and then identify connections among these concepts (Han et al., 2019; Kim et al., 2016; Son et al., 2020). Causal relation extraction methods identify the connections within the text and then aggregate them (Asghar, 2016; Bach & Badaskar, 2007; Khoo & Na, 2006). The latter is generally more appropriate for evidence synthesis as its bottom-up nature covers even infrequently occurring variables and retains the stated relationships among them (Barik et al., 2016). Causal relation extraction techniques can, in turn, be classified into knowledge-based, statistical machine learning, and deep learning techniques (Yang et al., 2021). Knowledge-based techniques rely on semantic and syntactic text characteristics, codified as patterns or rules, against which an algorithm can classify input text (e.g. a causative verb-noun pair) (Bui et al., 2010). These techniques perform well on simple text with explicit causality, but poorly on text with implicit causality (Beamer et al., 2008; Girju et al., 2009; Yang et al., 2021). Statistical machine learning identifies distinguishing characteristics based on annotated 'training' data using techniques such as Bayesian inference or decision trees. Statistical machine learning can handle implicit causality well, but it suffers from low portability, i.e., poor performance when the 'test' data is dissimilar from the training data (Asghar, 2016; Pakray & Gelbukh, 2014; Zhao et al., 2016). Finally, deep learning techniques utilize neural network architectures for causal relationship extraction; these perform well even on complex text (Yang et al., 2021). In addition, they can leverage 'transfer learning' to enhance portability of the algorithm to a new context (Beltagy et al., 2019; Devlin et al., 2018; Kyriakakis et al., 2019).

The next step is to make the evidence comparable through some form of homogenization or normalization. For example, in the case of a meta-analysis, a regression coefficient might be converted into a partial correlation coefficient to make the input comparable (Hansen et al., 2022). Whereas less relevant for qualitative evidence, a challenge for policy-relevant analysis is that the source material can span several (inter)disciplinary fields, with different vocabularies to refer to the same or similar phenomena (Goyal & Howlett, 2018; Saetren, 2005). For example, the use of 'governance arrangement' in governance studies and 'institution' in research on institutionalism might be synonymous with the use of 'policy instrument' in policy studies. While not a substitute for expert assessment, NLP techniques for measuring lexical or semantic similarity can be useful here. Lexical similarity relates to the similarity of a word set and is often measured using a dictionary-based approach (Cruanes et al., 2012). This approach is easy to implement but does not consider the context in which a term is used. In contrast, semantic similarity relates to the similarity in meaning between terms and is measured such that the distance between terms (or other units, such as sentences, paragraphs, documents) represents the likeness of their meaning (Liu et al., 2007). For complex text, semantic similarity is likely to be a more robust metric (Inan, 2020).

Thereafter, the evidence needs to be aggregated. For quantitative evidence synthesis, a weighted mean effect size for the cause-effect relationship might be computed (Hansen et al., 2022). However, as policies often have consequences across many dimensions – program, process, and politics (McConnell, 2010) – that vary over time (Compton & 't Hart, 2019; Goyal, 2021), it is essential to consider multiple effects for any cause. Further, relevant evidence might be available from diverse contexts varying in levels, geographies, scales, and time periods, making it important to highlight the relevant effects of contextual variables and internal mechanisms on the causal chain. In addition, the research designs of the sources could range from argumentative work to small-n case studies; medium or large-n observational, quasi-experimental, and experimental research; simulation modelling; and co-design or participatory knowledge. All of this lends itself to a more qualitative synthesis, for which in-depth information is usually summarized as conceptual diagrams, tables, or a narrative. Collating and visually presenting evidence as done in systems science can add value here.

System mapping facilitates collation of existing evidence to answer a specific question, identify knowledge clusters, or articulate a knowledge gap (James et al., 2016). Similarly, systematic evidence mapping helps summarize, query, and visualize evidence, for example, using a database and/or a graphical representation (Peters et al., 2021; Wolffe et al., 2019). Causal mapping is likely to be especially useful for causal evidence synthesis. It integrates causal chain analysis with the mapping of complex interrelationships, thereby facilitating collation of existing evidence or expert knowledge about (dynamic) causal interdependencies between parts of a system (Ackermann & Alexander, 2016). This allows to answer specific questions about effects of interest, identify knowledge clusters, or articulate a knowledge gap (James et al., 2016). Causal mapping is well-suited for combining analyses with different scopes and expanding the representation when new information is introduced (Eden et al., 1992).

As a causal map is a type of directed knowledge graph, it can be investigated using graph analytics. Here, by graph analytics we refer to techniques in causal chain analysis, graph theory, and network analysis. Illustratively, topographic analysis creates insight based on the structural layout of the causal map. An assessment of the strength of the linkages, for example, indicates the number of studies providing evidence on specific cause-effect relationships (Montibeller & Belton, 2006). This can shed light not only on interventions that are supported by much evidence, but also on parts of the system that may warrant further investigation. Similarly, effect trees–a collection of downstream variables affected by a given factor, and cause trees–a collection of upstream variables that influence a factor–can help analyze the cause and consequences, respectively, of an intervention (Eden et al., 1992). Centrality analysis can provide insight into the likely importance of a variable–either perceived or real–based on whether, for example, it has several linkages to other variables (i.e., high degree centrality) or lies on the shortest path between several other variables (i.e., high betweenness centrality). Some of the other possibilities for investigating a causal map are causal inference analysis and causal loop analysis. Causal inference refers to the process of determining the causal impact of one factor on another (Axelrod, 1976), either the partial effect (the causal impact of one variable on another along a specific path), or the total effect (the overall impact of one variable on another taking all paths into consideration). Causal loop analysis involves the identification and description of feedback mechanisms within a causal map, i.e., positive or negative feedback within (a part of) the system, resulting in reinforcing or balancing behavior, respectively. Apart from aiding causal evidence synthesis, causal maps can also inform subsequent policy analysis, for example, in the form of Bayesian belief analysis or simulation modelling (Pullin et al., 2016).

An overview of the semi-automated approach is shown in Fig. 1. Beginning with a selection of relevant text, causal relations are extracted, similar factors are grouped, and an aggregate causal map is built. The map is then analyzed to derive policy-relevant insights. It is important to note that a variety of different NLP algorithms can be used to achieve

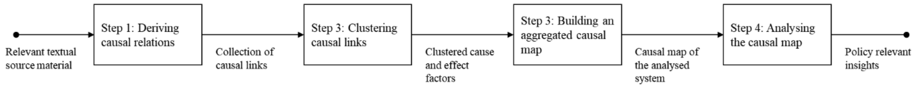

Relevant textual source material
Step 1: Deriving causal relations
Collection of causal links
Step 3: Clustering causal links
Clustered cause and effect factors
Step 3: Building an aggregated causal map
Causal map of the analysed system
Step 4: Analysing the causal map
Policy relevant insights

Fig. 1 Input-output diagram of the four-step process for causal evidence synthesis

the desired outputs of each step, and our approach itself is algorithm-agnostic. Algorithm selection should carefully consider the relevant metrics for a given case to ensure sufficient performance. In this way the approach can be rooted in and benefit from ongoing advancements in NLP.

Synthesizing policy-relevant evidence: an illustration using the emissions trading scheme (ETS)

We illustrate the process of synthesizing policy-relevant evidence and creating a causal map through the semi-automated approach by using the case of the ETS.

Emissions trading schemes

ETS are policy instruments that aim to impose a cost on the emission of greenhouse gases (GHGs). This policy instrument involves setting a 'cap' on certain types of emissions; regulating bodies then 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 allowances but must surrender a quantity equivalent 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 obligation. In this way, ETS provide a mechanism to reduce emissions, to specified levels, whilst encouraging the most cost-efficient abatement (ICAP, 2022a).

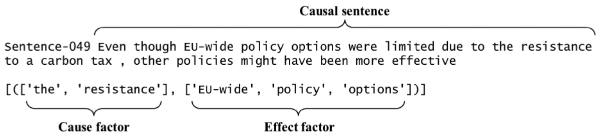
Evidence-synthesis from ETS policy analysis literature is encumbered by many of the limitations discussed earlier. First, there is an enormous literature base. For example, Google Scholar returns over 23,100 results for a search of "ETS" and "policy analysis". Even though a majority of these are likely to be excluded from a synthesis relying on empirical evidence, the resulting dataset could potentially consist of 50–100 sources. Second, the terminology employed for discussing ETS performance varies across these studies. ETS policy studies often only examine a single performance dimension despite compelling analysis from programmatic, process, and political perspectives. And, finally, analysis is conducted at various levels of scope from case studies to reviews seeking to summarize the wholesale abatement impact of carbon pricing. Therefore, this instrument is an appropriate choice for demonstrating our approach.

While several possibilities exist for compiling relevant source material, for this illustration we identified meta-reviews of ETS and selected constituent articles. This allowed us to compare the insights created by our approach against those in the meta-reviews. From the 14 meta-reviews on ETS identified in our search, we selected two: Green (2021), due to its explicit ex-post and quantitative direction encompassing studies from all major ETS jurisdictions, and Schmalensee and Stavins (2017) due to its specific discussion on political considerations and system design factors, which reflect more process and political perspectives. Collating the papers within both reviews and removing those deemed irrelevant yielded a final total of 28 source papers (see Appendix: Table A1).

Extracting cause-effect relationships

The first stage of our approach involves examining each sentence of the source material to determine whether it exhibits causality, and–if so–extracting the causal factor(s), the effect factor(s), and the direction of the causal relation(s). Take, for example, the causal sentence "The higher emissions allowance price caused a decrease in coal power generation". Here, 'emissions allowance price' is the cause factor, 'coal power generation' is the effect factor, the direction of the relationship is negative (i.e., an increase in the former results in a reduction in the latter). To semi-automate this process, we employed the 'Self-attentive BiLSTM-CRF wIth Transferred Embeddings' or SCITE algorithm, developed by Li et al. (2021). This algorithm was chosen, amongst various alternatives, based on three criteria. Firstly, it uses a deep learning, relation-extraction approach which we prefer to knowledge-based or statistical approaches for complex, inconsistent data, in-line with recommendations from Yang et al. (2021). Secondly, it is an open-source project, ensuring transparency and accessibility. Finally, it performed highly against relevant performance benchmarks for recall and precision. Recall, i.e., the percentage of true-positive to true-positive and false-negative outputs, is an important metric because it indicates the percentage of causal relations that were correctly identified. A higher recall value results in more causal relations being included in the model. Precision indicates the percentage of true-positives to true-positives and false-positives and, hence, the percentage of causality-extraction outputs that are indeed causal. Higher precision reduces the number of incorrect outputs carried forward to later stages. While we use SCITE for the cause-effect extraction stage of our approach, other suitable algorithms exist and will continue to emerge (see e.g., Akkasi & Moens, 2021; Ancin-Murguzur & Hausner, 2021; Liu et al., 2024).

To derive causal relations using SCITE, the 28 documents were processed in three stages. First, we selected only the abstract, results, discussion, and conclusion sections of the articles. The literature review and methods were excluded because these often-returned irrelevant methodological links, and relations suggested in other literature instead of findings of that article. Next, the text was automatically segmented into sentences, cleaned (removing non-alphanumeric characters, separation of punctuation etc.) and further segmented into words. Subsequently, 'layered embeddings' – necessary for the algorithm to learn a representation of the input data (Levy & Goldberg, 2014) – were generated for each sentence (Akbik et al., 2018). The sentences and corresponding embeddings were processed through a pre-trained SCITE model resulting in a list of sentences deemed causal, and their suggested cause-effect relationships. An example causal relation returned by SCITE is shown in Fig. 2.



```
Sentence-049 Even though EU-wide policy options were limited due to the resistance
to a carbon tax , other policies might have been more effective

[(['the', 'resistance'], ['EU-wide', 'policy', 'options'])]
```

Fig. 2 An illustrative output of the SCITE algorithm

In all, our data consisted of 4542 sentences that were cleaned and provided as input to SCITE. Of these, 317 were deemed as causal by the algorithm. The raw output, however, did not indicate the direction of causality (i.e., positive or negative) and included a high level of inaccuracy. This necessitated a manual review of each SCITE output to determine the true causal relations, causal pairs, and direction. This stage was also necessary to remove irrelevant causal sentences. Ultimately, 154 sentences were verified as causal and relevant. These sentences contained 284 causal pairs. Our validation of this step revealed that the pre-trained model of SCITE achieved approximately 84% precision and 38% recall for our data. Despite the current manual review stages required, the use of the algorithm reduced the time taken to extract causal relations from each paper from 3 h to approximately 20 min.

Clustering causal links

Next, we concatenated the individual causal pairs, their contributing sentence(s), cause-effect factors, and relationship direction. To ensure backward traceability, we provided a unique cause-effect id (in the form [article number]-[sentence number][causal pair letter]) to each causal pair, which highlights the prevalence of similar causal pairs. For example, when talking about the use of coal power, one causal pair may include the factor ‘coal power generation’ whereas another ‘coal utilization’. Including both factors would quickly cloud the causal map with repetitive links. To address this issue, semantic clustering can help in grouping factors that have the same meaning. This would reduce the number of (duplicated) factors and increase the (legitimate) interconnections, resulting in a more cohesive causal map.

Often this task is achieved in a two-stage process whereby ‘embeddings’ (a way of representing a word and/or sentence as a high-dimensional vector) are generated for each phrase and subsequently clustered using a suitable algorithm (Jacksi & Salih, 2020). By grouping phrase embeddings, this approach intends to capture some semantic connection between the phrases. More recently, large language models (LLMs) have demonstrated their capability for capturing semantic meaning (Hansen & Hebart, 2022; Le Mens et al., 2023) which could also prove effective for semantic clustering. Although we employed the above-mentioned embedding and clustering method for this project due to the lack of case studies that evaluate generative LLMs at the time of writing, future applications would do well to consider emerging methods for this task. We chose SBERT as the embedding algorithm due to its accessible open-source format and state-of-the-art performance against semantic similarity benchmarks at the time (Reimers & Gurevych, 2019).

These embeddings were then grouped using a clustering algorithm. Again, various algorithms could be used (e.g., k-means or agglomerative hierarchical clustering), each with its relative strengths (Rokach & Maimon, 2005). In this case, DBSCAN (Ester et al., 1996) was chosen due to its ability to handle uneven cluster sizes and outliers (Rokach & Maimon, 2005). We expect these in the data because many synonymous terms may exist for a certain factor (e.g., ‘emissions’, ‘GHG emissions’, and ‘carbon emissions’) requiring large clusters, whilst many unique terms may also be present which should remain un-clustered. DBSCAN was applied using a cosine distance metric (given its ability to handle high-dimensional vectors, such as embeddings) and a minimum cluster size of 2 (given that at least two semantically similar factors could warrant clustering). One overarching factor term is then manually chosen for each cluster and used to overwrite the factors present in that cluster where they occur in causal pairs (see Table 1). Note that the use of SBERT and DBSCAN algorithms here demonstrate the clustering stage of our approach, they are proposed as appropriate but not necessarily optimal algorithms for this task.

Applying the above approach, the initial collation of causal pairs resulted in the identification of 300 cause-effect factors in our dataset. Next, the cosine distance parameter of DBSCAN was iteratively tuned to achieve qualitatively determined ‘reasonable’ clusters, i.e. where phrases of the same semantic meaning (in the context of ETS) were grouped and those sufficiently unique phrases remained un-clustered. Using a final distance value of 0.3, 230 of the cause/effect factors were grouped into 49 clusters and the remaining 70 factors were deemed unique. From the clustered factors, we could prune duplicate cause-effect pairs leading to 119 unique cause-effect pairs.

Building an aggregated causal map

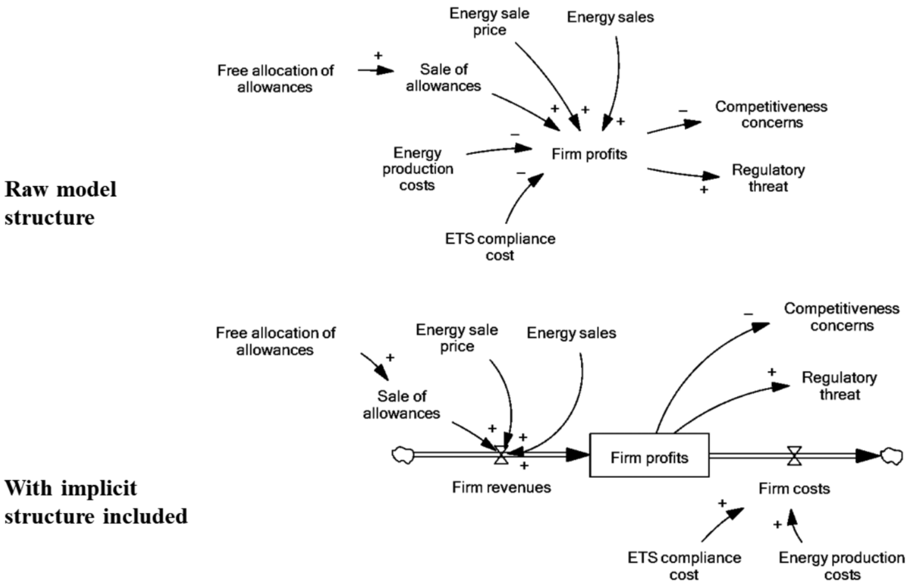
The next step is to generate the aggregate causal map. Using the clustered cause-effect pairs, individual factors can be added to the map as nodes and their causal links included as the edges between them. The causal map grows in complexity as more factors appear and connections between structures emerge.

While based on the causal links identified in earlier stages, this map generation stage inevitably involves coder interpretation. Despite clustering factors, additional grouping may be warranted. Implicit structures are also likely to become more evident, the inclusion of which can allow for richer synthesis. For example, many relations discussed the impact of the free allocation of emission allowances and subsequent sales on the profits of power-generating firms. Initially, this behavior is captured by various links to and from ‘firm profits’. The same behavior may instead be captured by a stock-flow structure, whereby revenues are aggregated as inflow to profit and costs as outflow, better reflecting the underlying dynamics. Figure 3 demonstrates how including the implicit stock-flow structure allows clearer demarcation of factors influencing revenues, costs, and profits.

Another issue that can arise in this process is that of ‘intermediate variables’, whereby one causal chain depicts a connection from A → C and another from A → B → C. In such cases, it is often unclear whether the link A → C implies the existence of intermediate factor

Table 1 An illustration of clustered causal pairs. The factors that have been clustered are shown in capital case

Corpus-ID	Raw cause effect pairs	Clustered cause effect pairs
2–028 A	[european level commitments, +, 20-20-2010 targets]	[European level commitments, +, 2020 EU CLIMATE ENERGY PACKAGE]
2-028B	[20-20-2010 targets, +, renewable energy directive]	[2020 EU CLIMATE ENERGY PACKAGE, +, RENEWABLE ENERGY DIRECTIVE]
2–028 C	[20-20-2010 targets, +, energy efficiency directive]	[2020 EU CLIMATE ENERGY PACKAGE, +, energy efficiency directive]
2–029	[renewable energy directive, +, renewable energy utilisation]	[RENEWABLE ENERGY DIRECTIVE, +, RENEWABLE ENERGY UTILISATION]
2–031 A	[carbon price, +, fuel-switching]	[EMISSIONS ALLOWANCE PRICE, +, FUEL-SWITCHING]


Raw model structure

With implicit structure included

Fig. 3 An example illustrating the inclusion of implicit structures in the causal diagram

B, it is unaware of factor B, it considers B irrelevant, or it suggests 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 interpretation. In situations where an intermediate variable is suggested but not explicitly mentioned by a causal link, the analyst can refer to the contributing text segment to gain more context. In other cases, support cannot be found for an intermediate link, and an additional path is added to the causal map alongside the intermediate path, i.e., including A → B → C alongside A → C.

What should be apparent is that model generation is an iterative process. As the map evolves, implicit structures emerge and intermediate variables appear, frequent reorganization is warranted. There is no objectively complete causal map; rather, the causal map should be refined until all causal links have been incorporated, and the overall structure can be used to garner relevant insights. The link to the contributing text segments is also necessary to provide additional context when examining the causal map. By labelling the arrows and providing a reference table of supporting causal IDs, the analyst can (help the readers) examine each causal link in more detail, if desired.

Our iterative process produced the final causal map shown in Fig. 4.

The synthesized evidence: probing the causal map on the emissions trading scheme

In this section, we briefly describe the causal map and demonstrate how it can be further analyzed to create policy-relevant insight.

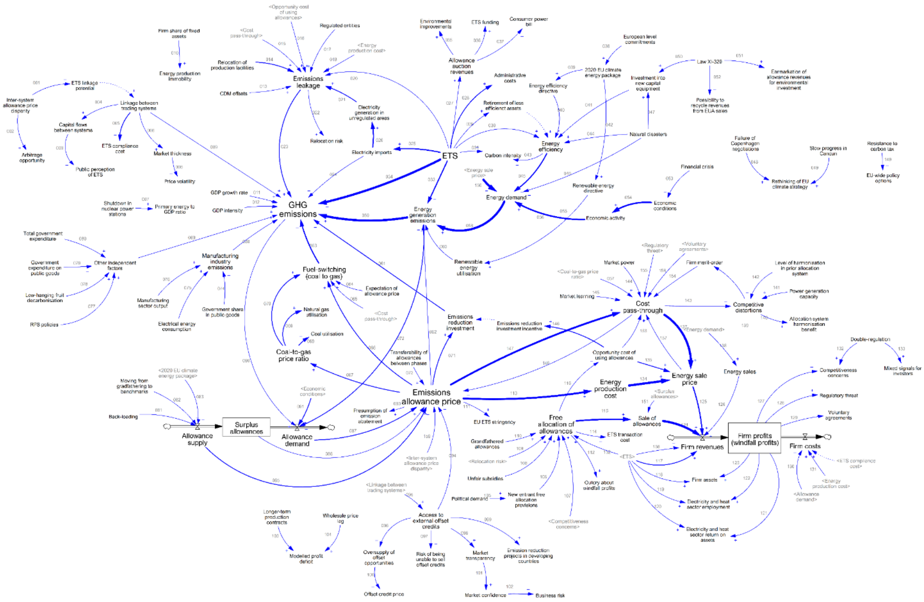

ETS
GHG emissions
Emissions leakage
Energy generation emissions
Energy demand
Energy efficiency
Emissions allowance price
Cost pass-through
Energy sale price
Energy production cost
Free allocation of allowances
Firm profits (windfall profits)
Firm revenues
Firm costs
Surplus allowances
Allowance supply
Allowance demand
Fuel-switching (coal to gas)
Coal-to-gas price ratio
<Opportunity cost of using allowances>
Firm share of fixed assets
Energy production immobility
<Cost pass-through>
Regulated entities
<Energy production cost>
Relocation of production facilities
CDM offsets
Electricity generation in unregulated areas
Relocation risk
Electricity imports
ETS linkage potential
Inter-system allowance price disparity
Linkage between trading systems
Arbitrage opportunity
Capital flows between systems
Public perception of ETS
ETS compliance cost
Market thickness
Price volatility
GDP growth rate
GDP intensity
Shutdown in nuclear power stations
Primary energy to GDP ratio
Total government expenditure
Government expenditure on public goods
Other independent factors
Low-hanging fruit decarbonisation
RPS policies
Manufacturing industry emissions
Manufacturing sector output
Electrical energy consumption
Government share in public goods
Environmental improvements
ETS funding
Consumer power bill
Allowance auction revenues
Administrative costs
Retirement of less efficient assets
Carbon intensity
<Energy sale price>
European level commitments
2020 EU climate energy package
Energy efficiency directive
Investment into new capital equipment
Law XI-329
Earmarkation of allowance revenues for environmental investment
Possibility to recycle revenues from EUA sales
Natural disasters
Renewable energy directive
Failure of Copenhagen negotiations
Slow progress in Cancun
Rethinking of EU climate strategy
Resistance to carbon tax
EU-wide policy options
Financial crisis
Economic conditions
Economic activity
Renewable energy utilisation
Expectation of allowance price
Natural gas utilisation
Coal utilisation
Emissions reduction investment
Emissions reduction investment incentive
Transferability of allowances between phases
<2020 EU climate energy package>
Moving from gradfathering to benchmarks
Back-loading
<Economic conditions>
Presumption of emission abatement
EU ETS stringency
Grandfathered allowances
<Relocation risk>
Unfair subsidies
<Inter-system allowance price disparity>
<Linkage between trading systems>
Political demand
New entrant free allocation provisions
<Competitiveness concerns>
Outcry about windfall profits
ETS transaction cost
Sale of allowances
<Surplus allowances>
<ETS>
Access to external offset credits
Oversupply of offset opportunities
Offset credit price
Risk of being unable to sell offset credits
Market transparency
Market confidence
Business risk
Emission reduction projects in developing countries
Longer-term production contracts
Wholesale price lag
Modelled profit deficit
<Regulatory threat>
<Voluntary agreements>
Market power
Firm merit-order
<Coal-to-gas price ratio>
Market learning
Opportunity cost of using allowances
<Energy demand>
Energy sales
Competitive distortions
Level of harmonisation in prior allocation system
Power generation capacity
Allocation system harmonisation benefit
Double-regulation
Competitiveness concerns
Mixed signals for investors
Regulatory threat
Voluntary agreements
Firm assets
Electricity and heat sector employment
Electricity and heat sector return on assets
<ETS compliance cost>
<Energy production cost>
<Allowance demand>

Fig. 4 An aggregated causal map showing the relationships influencing the ETS. A 'stock' variable is represented as a rectangular node while a 'flow' variable is represented as a text label. The thickness of the edge connecting two nodes indicates the number of publications providing evidence regarding the relationship. Nodes in angle brackets"<>" indicate factors that are already captured elsewhere in the map and mentioned again for ease of visualisation

Description of the aggregated ETS causal map

Before analyzing the causal map (Fig. 4), it is worth describing some core structural elements.

To begin with, the relation between ETS and GHG emissions. We observe that weight of the link is high and the relationship negative ('-'), indicating that several studies provide evidence that ETS is associated with a reduction in GHG emissions. We also observe two mediators that link from ETS to GHG emissions. First, ETS contributes to a reduction in 'emissions from energy generation', resulting in a reduction in GHG emissions. Second, ETS increases 'electricity imports', which also result in a reduction in GHG emissions within a particular jurisdiction. However, the causal map presents the complexity of the system and highlights that different factors could determine the overall influence of the ETS on GHG emissions. Some of the key factors include emissions allowance price, emissions leakage, and free allocation of allowances.

Emissions allowance price inducing GHG emission reductions

This structure concerns the variables linking 'emissions allowance price' and 'GHG emissions'. There is a clear influence of allowance price on emissions via fuel switching, elaborated by the coal-to-gas price ratio factor. This is consistent with the idea that allowance price will impact the cost of coal more than that of gas (as coal is more carbon-intensive), thereby inducing fuel switching. The link from allowance price to emissions via 'cost pass-through'/'energy production cost', 'energy sale price', and 'energy demand' captures the idea that price increases are passed on to consumers, who then reduce their energy demand. Additionally, there is a path through emissions reduction investment, as higher allowance prices encourage polluters to enact measures to reduce emissions. Through the sign of linkage, the map also conveys the expected impact that changes in any of these factors would have. For example, a lower allowance price would: (i) reduce the coal-to-gas price ratio and curtail fuel switching; (ii) limit the energy sale price (due to lower production cost), thereby having lesser influence on energy demand reduction; and (iii) lessen the emissions reduction investment.

Emissions leakage diminishing ETS efficacy

This structure concerns 'emissions leakage' and the different ways in which leakage impacts the system. One can observe the direct positive link from 'ETS' regulation to 'emissions leakage' to 'GHG emissions', reflecting the idea that ETS regulation incentivizes polluting activity to relocate outside regulated jurisdictions, thereby avoiding ETS-induced emissions reductions. Examining the structural location of emissions leakage also reveals its negative role in important causal chains. For example, consider again the path from 'allowance price' through 'cost pass-through'/'energy production cost', 'energy demand' to 'GHG emissions'.' This causal chain captures the idea that a higher allowance price reduces emissions. However, as both cost pass-through and energy production cost have a positive relationship with emissions leakage (which in turn has a negative relationship with GHG emissions), the effect of allowance price on emissions is mitigated, to some extent, by emissions leakage.

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 reflects the behavior, observed particularly in the earlier phases of the EU ETS, whereby firms profited from the sale of freely distributed emissions allowances. The basic mechanism is clearly present with a path from 'free allocation' to 'sale of allowances', 'firm revenues' and, thereby, '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.

Analysis of the aggregated ETS causal map

An analysis of the causal map can be conducted using techniques such as topographic analysis, causal inference analysis, and causal loop analysis. Here, we illustrate the type of insights such analyses could generate.

Topographic analysis

The most supported connections on the map are not surprising (Table 2). Links #034 ('ETS' and 'GHG emissions'), #058 ('energy demand' and 'energy generation emissions') and

Table 2 Causal relations with most unique supporting articles

Relation-ID	Causal relation	Number of unique supporting articles	Supporting corpus-IDs
034	ETS → GHG emissions (-)	9	1–068, 2-167, 3–064 A, 4–047 A, 9-101, 11-019B, 13–034, 18 − 003, 19 − 006, 19–076
058	Energy demand → Energy generation emissions (+)	9	1–035, 2–005 A, 2–023, 2-129, 3–050, 3-064B, 3–078 A, 4-047B, 5–067, 5-026B, 5–079, 12-085B, 17–083 A, 17-083B, 19-008B, 19–064 C, 19-064D, 19-078B, 22–056 C
059	Energy generation emissions → GHG emissions (+)	9	1–018, 1–035, 2–005 A, 2–023, 2–067 A, 2-129, 3–050, 3-064B, 3–078 A, 4-047B, 5-026B, 12-085B, 5–079, 17–083 A, 17-083B, 19-008B, 19–064 C, 19-064D, 19-078B, 22–056 C
124	Sale of allowances → Firm revenues (+)	6	1–029, 8-069B, 8-075B, 8–080, 8–076 A, 9-271, 13-336D, 20–044 C, 20–052, 20–251 A, 20-251B, 20–212 A, 20-212B, 20–256, 28 − 004
134	Energy production costs → Energy sale price (+)	6	9-059B, 13–339 A, 20–074, 21–043, 22–056 A, 24–067 C

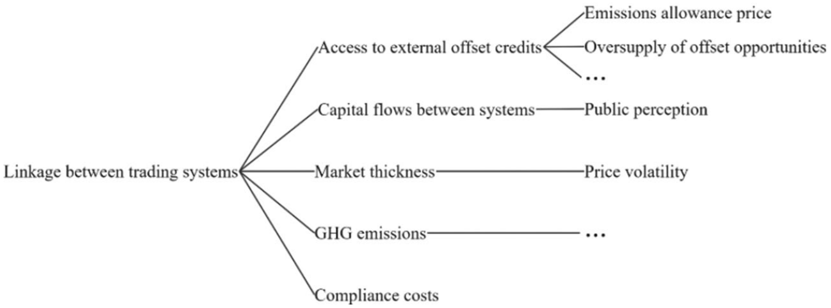

Linkage between trading systems
Access to external offset credits
Emissions allowance price
Oversupply of offset opportunities
...
Capital flows between systems
Public perception
Market thickness
Price volatility
GHG emissions
...
Compliance costs

Fig. 5 Effect-tree, the collection of downstream variables affected by a given factor, for 'Linkage between trading systems'

#059 ('energy generation emissions' and 'GHG emissions') are each supported by nine articles and relate to obvious links between central factors in the system. This high degree of shared knowledge implies that these connections are relevant in explaining the ETS system. Conversely, #071 ('emissions allowance price' and 'emissions reduction investment') and #061 ('emissions reduction investment' and 'GHG emissions') have only two articles supporting them, despite their centrality to the system. The apparent lack of knowledge regarding these relationships suggests that further research would be warranted.

Consider the effect tree of ETS system linkage shown in Fig. 5. If an analyst were 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 because 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 recommend measures to mitigate the negative downstream effects, sever the negative effect branches, or introduce reinforcing branches that promote the desired behavior.

Table 3 Shows relevant network metrics derived from the causal map. The high in-degree (14) of GHG emissions reflects this factor's position at the end of many causal paths. Conversely, ETS with the highest out-degree (14) is a source of causal paths. Allowance

demand, emissions allowance price, cost pass-through and firm profits have the highest betweenness, reflecting their position as key levers within the system. Demand for allowances has much higher betweenness than in- or out-degree; hence, while it is included in many causal paths that determine system behavior, its influence is limited to its impact on emissions allowance price and firm profits. Eccentricity reflects which factors are the farthest away from others, potentially making their causal influence on outcomes harder to trace. High eccentricity scores for Law-XI-829, european-level commitments and renewable energy directive, for example, indicate that it may be difficult to attribute changes in system outcomes to these, given the many intermediary mechanisms on the path to certain outcomes. Yet, this shows how different policy instruments interact with the ETS and potentially influence its effectiveness.

Depending on the questions of interest, other network analysis metrics and methods could be used.

Causal inference analysis

Consider the path from 'emissions allowance price' to 'coal-to-gas price ratio', 'fuel switching (coal to gas)', and 'GHG emissions'. The partial effect is negative, consistent with the idea that a greater allowance price will lead to a price disparity between coal and gas fuel sources which in turn leads to fuel switching from coal to gas, thereby reducing emissions. Article 2 supported the link between price ratio and fuel switching but did not include the connection between price-ratio and emissions allowance price. Article 16 did mention this relation but the link between price ration and fuel-switching was omitted. It is only by combining the causality of these two articles that the entire causal chain from allowance price to emissions becomes apparent. If a policy maker sought to encourage fuel-switching, an understanding of the aggregated path shows that increasing the allowance price would likely contribute to this end.

Looking instead at the total effect, there is an obvious path from 'cost pass-through' to 'GHG emissions' via 'energy sale price', 'energy demand', and '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 consumption and associated emissions. Such an effect is well-studied and understood, with aspects of this path covered by 15 articles. However, taking instead the path from 'cost pass-through' to 'emissions leakage' to 'GHG

Table 3 Network metrics of the causal maps, showing nodes with highest in-degree, out-degree, betweenness centrality, closeness centrality, and eccentricity

In-degree		Out-degree		Betweenness		Eccentricity	
GHG emissions	14	ETS	14	Allowance demand	0.069	Law XI-329	9
Cost pass-through	9	Emissions allowance price	7	Emissions allowance price	0.068	European level commitments	8
Emissions leakage	8	Firm profits	6	Cost pass-through	0.056	Generation capacity	8
Emissions allowance price	6	Cost pass-through	5	Firm profits	0.051	Renewable energy directive	8
Firm profits	6	Access to external credits	5	GHG emissions	0.050	Expectation of carbon price	8

emissions' instead yields a positive partial effect, consistent with the idea that greater pass-through costs contribute to greater leakage and associated net emissions. The relationship between pass-through and leakage is suggested only in Article 27. In this case, the total effect between these two factors is, thus, undetermined. If targeting emissions reductions, a policy maker might examine the degree of cost pass-through. By only considering the energy sale price pathway, they might conclude that encouraging cost pass-through would be fruitful. Examining the total effect of cost pass-through on emissions would, instead, reveal that emissions leakage can mitigate the reduction effect somewhat.

Causal loop analysis

Consider the balancing loop between emissions allowance price and GHG emissions (Fig. 6). As a reduction in GHG emissions lessens allowance demand, there is a potential dampening effect on the emissions allowance price if allowance supply is not cut back to keep pace with these reductions. This would, in turn, impact subsequent GHG emissions. That is, if future allowance supply caps are not set sufficiently low, then successful abatement efforts may reduce the allowance price and induce a balancing effect on GHG emissions. Indeed, a similar issue was experienced in the first phase of the EU ETS, whereby an allowance oversupply contributed to an allowance price collapse (Schmalensee & Stavins, 2017). An investigation of this causal loop indicates that, alongside stringent allowance caps, a price collar could work to mitigate this issue.

This brief illustration shows that our approach results in different types of policy-relevant insights. First, the causal map provides an overview of the system and the different factors that influence policy effectiveness within it. Second, the topographic analysis establishes which factors and relationships appear to be influential and whether the existing evidence or research is aligned with their centrality to the system. Third, the causal inference analysis and the causal loop analysis point to 'feedback' or non-linearity, and help identify key pathways that result in balancing or reinforcing dynamics within the system. Fourth, the analysis highlights the interaction among policy instruments, with other governance arrangements on energy efficiency and renewable energy also influencing the effectiveness of the ETS to a greater or lesser extent.

Comparisons with the meta-reviews

Next, we examine whether the causal map captures the insights of the meta-reviews from which the source material was selected. Our approach performs well in the identification of granular features contributing to high-level behavior. Green (2021), for example, notes

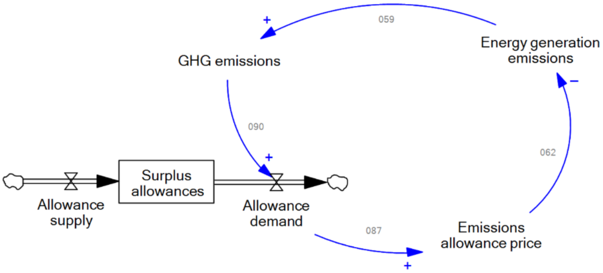


Fig. 6 A 'balancing' causal loop between allowance price and emissions: higher allowance price reduces emissions, reducing allowance demand and thereby reducing allowance price

the effect of the ETS on fuel switching as well as the (modest) contribution of fuel switching, energy efficiency, and reduced fuel consumption on GHG emissions, all of which are captured in the causal map and subsequent analysis. Further, Green discusses the mitigating role of consistently low carbon price on the impact of ETS on GHG emissions, along with the issues associated with offset credits. By tracing the effect of reducing allowance price, its negative impact on GHG emissions can clearly be seen. Looking at the 'access to external offset credits' factor, issues relating to market confidence and financial risks that can negatively impact allowance price are observed. However, a key insight from Green was that the quantitative impact of emission trading is, so far, limited. This counterfactual finding (relating to the *lack* of relationship) is not picked up by causal-relation extraction and hence the results of our approach. Similarly, geography- or sector-specific insights are difficult to discern from the causal map (such as the efficacy of an ETS in the energy sector versus manufacturing or the discussion regarding specific geographies such as in Green, 2021).

In contrast, Schmalensee and Stavins (2017) discuss the performance of various ETS systems qualitatively. They explain the potentially large revenues that can be generated through allowance auctions, highlight the importance of free allowance allocations in gaining political support for ETS policy, and note that these allocations are motivated by concerns of adverse impact of ETS on competitiveness. With the causal analysis, all these factors are apparent. Schmalensee and Stavins (2017) further 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 an examination of the various causal paths from allowance price to GHG emissions, it can be interpreted that fluctuating price would, in turn, contribute to variation in emission abatement. Similar to Green (2021), the insights from this review that are not apparent in our causal analysis relate to heterogeneities in performance of different ETS. For example, Schmalensee and Stavins (2017) note that carbon leakage is particularly concerning for subnational systems.

This comparison, firstly, provides confidence that our approach can faithfully derive many insights identified in a traditional literature review. In addition, it also highlights the strengths of our approach: a systems perspective, a focus on underlying mechanisms influencing policy effectiveness, incorporation of policy 'mixes' described in the source material, and a repeatable appraisal of the gaps in existing evidence. At the same time, in its present form, our approach does not detect the lack of a causal relationship nor lend itself to quantification of the magnitude of a causal relationship. Further, heterogeneities in the relationship—for example, based on geography, level of government, or sector—are also difficult to discern from a single causal map.

Discussion

We have presented a novel, semi-automated, NLP approach to extract causal statements from policy analysis literature, aggregate them into a causal map of policy behavior, and derive policy-relevant insights. The results show that the method was able to address four shortcomings of state-of-the-art approaches for evidence synthesis for evidence-informed policy making as illustrated for ETS.

Firstly, it provides a configurative approach to reviewing and synthesizing available documented causal evidence. This allows analysts to better capture complex mechanisms underlying observable policy effects and to understand their effectiveness within the policy system. Secondly, the semi-automated approach enables an analyst to do so with significantly less effort than traditional review and evidence synthesis methods, at high levels of accuracy. Using NLP is relatively easy and less time-consuming than manual text review, reducing the time to derive causal relations by almost 90%, from ~3 h to ~20 min per article. As a result, a greater number of sources can be considered with the same resources, increasing the evidence base and reducing potential biases arising from a more limited dataset. The high precision achieved by the relation extraction algorithm also suggests that a high-quality analysis can be obtained. Thirdly, the clustering algorithm helps with integrating policy evidence that is scattered across policy issues or domains. The use of NLP reduced the time taken to cluster semantically similar concepts by approximately 50% (from 2 h to 1 h for about 300 factors). Lastly, the proposed approach can harmonize disconnected information from various levels of scope, performance perspectives, and taxonomies while also incorporating upstream and downstream effects into one comprehensive causal model. In combination, this equips analysts and policy makers with a more systemic understanding of a policy domain beyond the direct cause-effect inferences that are typically obtained from traditional means of evidence synthesis.

As demonstrated for ETS policy, the derived causal map captures most of the insights obtained from manual evidence syntheses as presented in the review articles by Schmalensee and Stavins (2017) and Green (2021). The comparison has highlighted some strengths and weaknesses of this causal map. The strengths include that the generated causal map can faithfully represent the granular features contributing to ETS system behavior and that most policy behavior as highlighted in the reviews is captured. The map performs well in distilling the factors and dynamics present in the disparate source material that are not easily obtained nor communicated otherwise. This can supplement traditional methods in presenting an explicit, parsimonious, aggregated systems perspective. Quantified insights and insights by sector while not represented explicitly in this causal map, however, could be uncovered by referencing the contributing text segments.

Despite these promising results, several limitations remain that should be addressed in future work to ensure high quality analysis and to realize the potential of this approach for evidence-informed policy. A major limitation concerns the low recall (~38%) achieved by the causal extractions algorithm we used. A consequence is that relevant causal links contributing to system behavior were probably missed. The low recall is likely explained by the data sources used (and the complexity of cause-effects therein, relative to the training data) rather than an inherent issue of the algorithm, which has achieved recall rates of up to 86% on benchmark data sets (Li et al., 2021). Addressing this shortcoming, however, is complicated by the sparsity of training data (Asghar, 2016). Here, the creation of benchmark datasets specific to public policy could be invaluable (see, for example, Sewerin et al., 2023 for such an effort in the case of policy design). Further, the rapidly developing capabilities of deep learning techniques for causal relation extraction (Yang et al., 2021) could also provide avenues for improvement.

More broadly, the significance of algorithm selection and case-specific evaluation cannot be overstated when utilizing NLP. Researchers or practitioners who utilize it should carefully consider the relevant metrics for their application and test various algorithms against these metrics to ensure sufficient performance for their specific case. Further, they should consider fine-tuning off-the-shelf algorithms with context or domain specific data, as errors resulting from their use could result in ineffective or even harmful policy implications. Here, too, the creation of models and tools that are specific to public policy (e.g., creating a 'PolicyBERT') could be helpful (see, for example, Webersinke et al., 2021).

Another potential limitation relates to the possible exclusion of important insights in the source material in the resulting causal map. Two underlying issues here pertain to the lack of consideration for counterfactuals and the current inability of the causal relation extraction algorithm to distinguish between hypothesized and empirically substantiated cause-effects. Recent developments using adversarial training for causality extraction models might be able to address this (Feder et al., 2021). Also, a related issue is the lack of a well-specified context within which the extracted cause-effect relations are deemed valid. It is possible to uncover additional contextual information by analyzing the underlying textual information present in the reference table before creating the causal map. However, this is inconvenient, especially when new information is collected iteratively. Again, developments in NLP might overcome these limitations soon: for example, newer word embedding techniques could facilitate mapping of contextual variables in a dataset (Pelevina et al., 2017; Selva Birunda & Kanniga Devi, 2021).

Yet another limitation relates to the scalability of this approach. Our illustrative example consists of only 28 source articles. A solution for *large*-scale evidence synthesis of hundreds or thousands of articles would require further development and automation of some of the stages of our approach. The relation derivation stage should already be suitable for this task provided that the precision and recall performance are deemed acceptable for the application. The clustering stage, however, would need to reliably aggregate thousands of causal relations into a comprehensible number of causal linkages without abstracting away necessary context in absence of manual intervention. Similarly, a manual approach would no longer be feasible for the causal-map generation. Here, advances in automating the workflow from causal relationship extraction via relations table building to creating a visual causal map (e.g. building on Ancin-Murguzur & Hausner, 2021) could reduce the need for iterative updating of the reference table and map and facilitate interactive exploration by analysts and policy makers. Additionally, experimentation with more effective means of communicating a systems perspective would be useful as causal maps can become difficult to comprehend when dealing with extremely large, complex systems.

Further to these important areas for improvement, future research may seek to use our approach to compare and contrast the causal maps that can be extracted from different sources and fields of evidence generation. This can help to expose conflicting information about causal paths from different sources of evidence. A directed network map representation allows to gauge contributions of various factors within and across causal chains, including potential for weighting of links or factors e.g., by frequency of mentioning, source type, soundness of the empirical basis, or authoritativeness of the source. The directed map representation, furthermore, provides additional analytic capacity. This includes topographic analyses using graph theory-based concepts to identify the most central factors as system levers or barriers to effective policy. Alternatively, causal reinforcing or balancing factors can be identified and explored, which is not commonly done in ETS policy analysis.

In a broad domain such as climate mitigation policy, it is challenging to keep an overview of the many policies and their effects, both individually and in concert. Here, a causal map presentation allows to either zoom in to evaluate a specific policy, or to zoom out and identify feedback with other policies that enhance or limit policy effectiveness. It may also guide empirical validation of the key cause-effects identified. Beyond that, their expansion to system maps (i.e., that include possible policy actions, policy effects, and contextual variables that would influence the cause-effect pathway, see e.g., Enserink et al., 2022) would provide an ideal foundation for model-supported exploration of policy behavior and policy system change over time (e.g., using system dynamics or agent-based modelling). Altogether, these would result in a more comprehensive and complexity-proof evidence-base for policy research, policy analysis, and policy making.

Conclusion

Evidence-based or evidence-informed policy relies on the ability to effectively summarize, synthesize, and mobilize knowledge for the policy process, yet characteristics of policy sciences as a field make evidence synthesis challenging in practice. The exponential growth in policy research significantly increases the resources necessary for evidence synthesis, evidence is often scattered across different policy areas and employing distinct terminology; policy impacts can be measured in a variety of dimensions, such as programmatic, process, and political, and at different levels from micro-level studies of individual policies to macro-level research on entire policy areas - all of which need to be taken into account when synthesizing evidence.

To address these shortcomings, this article has introduced a novel analysis method that can semi-automatically derive and aggregate causal relations from policy analysis literature into a causal map of policy effects. Applying this method to a collection of 28 ETS literature sources produced a causal map consisting of 159 unique causal links. Evaluation of this result has demonstrated that the approach allows analysts to better capture complex mechanisms and interactions underlying observable policy effects with significantly less effort than traditional review and evidence synthesis methods. It also supports synthesis of policy evidence scattered across policy issues or areas and can harmonize disconnected information from various levels of scope, performance perspectives, and taxonomies. Comparison of insights obtained from the causal map against those from a manual review of the same source material has demonstrated that most of the insights can be captured by this approach, whilst providing a more configurative perspective of the features contributing to policy behavior. Finally, in providing a causal map representation of a policy area, new tools for policy evaluation, such as topographic, causal inference, and causal loop analysis, become available to analysts. While promising in many respects, some notable limitations include the poor recall (~38%) achieved in this application of the method. This may contribute to structural gaps in the map and a lack of contextualization for cause-effect relations which can inhibit understanding of the nuance behind certain behavior.

Regardless of algorithmic performance or the quality of insights obtained for ETS, the method and implementation presented in this study only represents a first step in combining NLP, causal mapping, and graph analytics for policy-relevant evidence synthesis. The proposed method, and future iterations, appears to contribute a promising new tool for policy analysts across domains, helping to provide a more comprehensive understanding of the factors and relations affecting policy and ultimately improving the evidence base on which to inform policy development.

Appendix 1

Table A1 The source articles on the ETS used for the synthesis

Contributing review paper	Index	Individual papers
(Green, 2021)	1	(B. 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	(Martin & Saikawa, 2017)
	16	(Ellerman & McGuinness, 2008)
	17	(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)

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

Ackermann, F., & Alexander, J. (2016). Researching complex projects: Using causal mapping to take a systems perspective. *International Journal of Project Management*, *34*(6), 891–901. https://doi.org/10.1016/j.ijproman.2016.04.001

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649.

Ancin-Murguzur, F. J., & Hausner, V. H. (2021). causalizeR: A text mining algorithm to identify causal relationships in scientific literature. *PeerJ*, *9*, e11850. https://doi.org/10.7717/peerj.11850

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

Anderson, L. M., Oliver, S. R., Michie, S., Rehfuess, E., Noyes, J., & Shemilt, I. (2013). Investigating complexity in systematic reviews of interventions by using a spectrum of methods. *Journal of Clinical Epidemiology*, *66*(11), 1223–1229. https://doi.org/10.1016/j.jclinepi.2013.06.014

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.

Barbrook-Johnson, P., & Penn, A. S. (2022). *Systems Mapping: How to build and use causal models of systems*. Springer Nature.

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

Barza, M., Trikalinos, T. A., & Lau, J. (2009). Statistical considerations in Meta-analysis. *Infectious Disease Clinics of North America*, *23*(2), 195–210. https://doi.org/10.1016/j.idc.2009.01.003

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.

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

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.

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

Burns, P. B., Rohrich, R. J., & Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and Reconstructive Surgery*, *128*(1), 305–310. https://doi.org/10.1097/PRS.0b013e318219c171

Carroll, C., Booth, A., & Cooper, K. (2011). A worked example of ‘best fit’ framework synthesis: A systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology*, *11*. https://doi.org/10.1186/1471-2288-11-29

Compton, M., & ’t Hart, P. (2019). *Great policy successes*. First). Oxford University Press.

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

Cruanes, J., Roma-Ferri, M. T., & Lloret, E. (2012). *Measuring lexical similarity methods for textual mapping in nursing diagnoses in Spanish and SNOMED-CT*. *180*, 255–259. https://doi.org/10.3233/978-1-61499-101-4-255

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*.

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 1756736). Social Science Research Network. https://papers.ssrn.com/abstract=1756736

El-Taliawi, O. G., Goyal, N., & Howlett, M. (2021). Holding out the promise of Lasswell's dream: Big data analytics in public policy research and teaching. *Review of Policy Research*, *38*(6), 640–660.

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., & 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

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

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon.

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. 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

Feder, A., Oved, N., Shalit, U., & Reichart, R. (2021). CausaLM: Causal Model Explanation through Counterfactual Language models. *Computational Linguistics*, *47*(2), 333–386. https://doi.org/10.1162/coli_a_00404

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

France, E. F., Cunningham, M., Ring, N., Uny, I., Duncan, E. A. S., Jepson, R. G., Maxwell, M., Roberts, R. J., Turley, R. L., Booth, A., Britten, N., Flemming, K., Gallagher, I., Garside, R., Hannes, K., Lewin, S., Noblit, G. W., Pope, C., Thomas, J., & Noyes, J. (2019). Improving reporting of meta-ethnography: The eMERGe reporting guidance. *Psycho-Oncology*, *28*(3), 447–458. https://doi.org/10.1002/pon.4915

Freebairn, L., Atkinson, J., Kelly, P., McDonnell, G., & Rychetnik, L. (2016). Simulation modelling as a tool for knowledge mobilisation in health policy settings: A case study protocol. *Health Research Policy and Systems*, *14*(1). https://doi.org/10.1186/s12961-016-0143-y

Fusar-Poli, P., & Radua, J. (2018). Ten simple rules for conducting umbrella reviews. *Evidence-Based Mental Health*, *21*(3), 95–100. https://doi.org/10.1136/ebmental-2018-300014

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*.

Goyal, N. (2021). Explaining Policy Success using the multiple streams Framework: Political success despite programmatic failure of the Solar Energy Policy in Gujarat, India. *Politics & Policy*, *49*(5), 1021–1060. https://doi.org/10.1111/polp.12426

Goyal, N., & Howlett, M. (2018). Lessons learned and not learned: Bibliometric analysis of Policy Learning. In C. A. Dunlop, C. M. Radaelli, & P. Trein (Eds.), *Learning in Public Policy: Analysis, modes and outcomes* (pp. 27–49). Springer International Publishing.

Goyal, N., & Howlett, M. (2019). Combining internal and external evaluations within a multilevel evaluation framework: Computational text analysis of lessons from the Asian Development Bank. *Evaluation*, *25*(3), 366–380. https://doi.org/10.1177/1356389019827035

Goyal, N., & Howlett, M. (2021). Measuring the Mix of policy responses to COVID-19: Comparative policy analysis using topic modelling. *Journal of Comparative Policy Analysis: Research and Practice*, *23*(2), 250–261.

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

Greenhalgh, T., Wong, G., Westhorp, G., & Pawson, R. (2011). Protocol - realist and meta-narrative evidence synthesis: Evolving standards (RAMESES). *BMC Medical Research Methodology*, *11*(1), 115. https://doi.org/10.1186/1471-2288-11-115

Haddaway, N. R., Land, M., & Macura, B. (2017). A little learning is a dangerous thing: A call for better understanding of the term 'systematic review'. *Environment International*, *99*, 356–360. https://doi.org/10.1016/j.envint.2016.12.020

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

Hannes, K., & Lockwood, C. (2011). Pragmatism as the philosophical foundation for the Joanna Briggs meta-aggregative approach to qualitative evidence synthesis. *Journal of Advanced Nursing*, *67*(7), 1632–1642. https://doi.org/10.1111/j.1365-2648.2011.05636.x

Hansen, H., & Hebart, M. N. (2022). *Semantic features of object concepts generated with GPT-3* (arXiv:2202.03753). arXiv. https://doi.org/10.48550/arXiv.2202.03753

Hansen, C., Steinmetz, H., & Block, J. (2022). How to conduct a meta-analysis in eight steps: A practical guide. *Management Review Quarterly*, *72*(1), 1–19. https://doi.org/10.1007/s11301-021-00247-4

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.

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

ICAP (2022b). Emissions Trading Worldwide: Status Report 2022. Berlin: International Carbon Action Partnership.

Inan, E. (2020). SimiT: A text similarity method using lexicon and dependency representations. *New Generation Computing*, *38*(3), 509–530. https://doi.org/10.1007/s00354-020-00099-8

Jacksi, K., & Salih, N. (2020). State of the art document clustering algorithms based on semantic similarity. *Jurnal Informatika*, *14*(2), 58. https://doi.org/10.26555/jifo.v14i2.a17513

James, K. L., Randall, N. P., & Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, *5*(1). https://doi.org/10.1186/s13750-016-0059-6

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

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

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. https://doi.org/10.1016/j.eswa.2016.03.043. Scopus.

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., Saudabayev, A., & Ginés i Ametllé, J. (2019). Transfer learning for causal sentence detection. *Proceedings of the 18th BioNLP Workshop and Shared Task*, *292-297*. https://doi.org/10.18653/v1/W19-5031

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.

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, *120*(49), e2309350120. https://doi.org/10.1073/pnas.2309350120

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, *2: Short Papers*, 302–308.

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. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Liu, X., Zhou, Y., & Zheng, R. (2007). Measuring Semantic Similarity in Wordnet. 2007 International Conference on Machine Learning and Cybernetics, 6, 3431-3435.

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*.

Macura, B., Suškevičs, M., Garside, R., Hannes, K., Rees, R., & Rodela, R. (2019). Systematic reviews of qualitative evidence for environmental policy and management: An overview of different methodological options. *Environmental Evidence*, *8*(1). https://doi.org/10.1186/s13750-019-0168-0

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). https://doi.org/10.1038/s41558-017-0001-0

McConnell, A. (2010). Policy Success, Policy failure and Grey Areas In-Between. *Journal of Public Policy*, *30*(3), 345–362. https://doi.org/10.1017/S0143814X10000152

Millard, T., Synnot, A., Elliott, J., Green, S., McDonald, S., & Turner, T. (2019). Feasibility and acceptability of living systematic reviews: Results from a mixed-methods evaluation. *Systematic Reviews*, *8*(1). https://doi.org/10.1186/s13643-019-1248-5

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

Nguyen, D., & Keshav Pingali. (2013). Andrew Lenharth, &. *A lightweight infrastructure for graph analytics*. Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, Farminton, Pennsylvania.

Norman, E. R. (2023). Policy studies yearbook annual review 2022–23: Note from the editor and call for papers. Politics & Policy, 51(3), 348–354. https://doi.org/10.1111/polp.12545

Nunez-Mir, G. C., Iannone, I. I. I., Pijanowski, B. V., Kong, B. C., 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

O'Leary, D. F., Casey, M., O'Connor, L., Stokes, D., Fealy, G. M., O'Brien, D., Smith, R., McNamara, M. S., & Egan, C. (2017). Using rapid reviews: An example from a study conducted to inform policy-making. *Journal of Advanced Nursing*, *73*(3), 742–752. https://doi.org/10.1111/jan.13231

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

Pearson, A., White, H., Bath-Hextall, F., Salmond, S., Apostolo, J., & Kirkpatrick, P. (2015). A mixed-methods approach to systematic reviews. *International Journal of Evidence-Based Healthcare*, *13*(3), 121–131. https://doi.org/10.1097/XEB.0000000000000052

Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2017). *Making Sense of Word Embeddings* (arXiv:1708.03390). arXiv. https://doi.org/10.48550/arXiv.1708.03390

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, *18*(10), 2119–2126. https://doi.org/10.11124/JBIES-20-00167

Peters, M. D. J., Marnie, C., Colquhoun, H., Garritty, C. M., Hempel, S., Horsley, T., Langlois, E. V., Lillie, E., O'Brien, K. K., Tunçalp, Ö., Wilson, M. G., Zarin, W., & Tricco, A. C. (2021). Scoping reviews: Reinforcing and advancing the methodology and application. *Systematic Reviews*, *10*(1). https://doi.org/10.1186/s13643-021-01821-3

Petrick, S., & Wagner, U. J. (2014). *The Impact of Carbon Trading on Industry: Evidence from German Manufacturing Firms* (SSRN Scholarly Paper 2389800). Social Science Research Network. https://doi.org/10.2139/ssrn.2389800

Petticrew, M., & Roberts, H. (2008). Systematic Reviews in the Social Sciences: A practical guide. John Wiley & Sons.

Pullin, A., Frampton, G., Jongman, R., Kohl, C., Livoreil, B., Lux, A., Pataki, G., Petrokofsky, G., Podhora, A., Saarikoski, H., Santamaria, L., Schindler, S., Sousa-Pinto, I., Vandewalle, M., & Wittmer, H. (2016). Selecting appropriate methods of knowledge synthesis to inform biodiversity policy. *Biodiversity and Conservation*, *25*(7), 1285–1300. https://doi.org/10.1007/s10531-016-1131-9

Ranson, M., & Stavins, R. N. (2013). Post-Durban climate policy architecture based on linkage of cap-and-trade systems. Chicago Journal of International Law, 13, 403–438.

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

Rokach, L., & Maimon, O. (2005). Clustering Methods. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 321–352). Springer US. https://doi.org/10.1007/0-387-25465-X_15

Saetren, H. (2005). Facts and myths about research on public policy implementation: Out-of-fashion, allegedly dead, but still very much alive and relevant. *Policy Studies Journal*, *33*(4), 559–582. https://doi.org/10.1111/j.1541-0072.2005.00133.x

Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. *Public Administration*, *80*(1), 1–22.

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

Saul, J. E., Willis, C. D., Bitz, J., & Best, A. (2013). A time-responsive tool for informing policy making: Rapid realist review. *Implementation Science*, *8*(1). https://doi.org/10.1186/1748-5908-8-103

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

Selva Birunda, S., & Kanniga Devi, R. (2021). A Review on Word Embedding Techniques for Text Classification. In J. S. Raj, A. M. Iliyasu, R. Bestak, & Z. A. Baig (Eds.), *Innovative Data Communication Technologies and Application* (pp. 267–281). Springer. https://doi.org/10.1007/978-981-15-9651-3_23

Sewerin, S., Kaack, L. H., Küttel, J., Sigurdsson, F., Martikainen, O., Esshaki, A., & Hafner, F. (2023). Towards understanding policy design through text-as-data approaches: The policy design annotations (POLIANNA) dataset. *Scientific Data*, *10*(1), 896.

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

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *Bmc Medical Research Methodology*, *8*, 45. https://doi.org/10.1186/1471-2288-8-45

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. https://doi.org/10.1016/j.foodpol.2010.06.005

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2). https://doi.org/10.1038/s42256-020-00287-7

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. https://doi.org/10.1177/0020852311419387

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBERT: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Weible, C.M., & Sabatier, P.A. (2018). Theories of the Policy Process (4th ed.). Routledge. https://doi.org/10.4324/9780429494284

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

Wolffe, T. A. M., Whaley, P., Halsall, C., Rooney, A. A., & Walker, V. R. (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environment International*, *130*, 104871. https://doi.org/10.1016/j.envint.2019.05.065

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013a). RAMESES publication standards: Meta-narrative reviews. *BMC Medicine*, *11*, 20. https://doi.org/10.1186/1741-7015-11-20

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013b). RAMESES publication standards: Realist syntheses. *BMC Medicine*, *11*(1). https://doi.org/10.1186/1741-7015-11-21

World Bank (2023). *State and Trends of Carbon Pricing 2023*. https://doi.org/10.1596/39796

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